Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2575548 A1 2006/07/27

(21) 2 575 548

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) Date de dépôt PCT/PCT Filing Date: 2005/07/29
(87) Date publication PCT/PCT Publication Date: 2006/07/27
(85) Entrée phase nationale/National Entry: 2007/01/29
(86) N° demande PCT/PCT Application No.: US 2005/027239
(87) N° publication PCT/PCT Publication No.: 2006/078318
(30) Priorités/Priorities: 2004/07/29 (US60/592,805); 2004/09/13 (US60/609,833); 2004/10/08 (US60/616,833); 2004/12/07 (US60/633,418); 2004/12/30 (US60/640,069); 2005/03/11 (US60/660,321); 2005/04/22 (US60/673,754); 2005/06/21 (US60/693,001); 2005/07/01 (US60/695,453); 2005/07/11 (US60/697,643)

(51) Cl.Int./Int.Cl. *A61K 39/02* (2006.01)

(71) Demandeur/Applicant:
CHIRON CORPORATION, US

(72) Inventeurs/Inventors:
TELFORD, JOHN L., IT;
GRANDI, GUIDO, IT;
LAUER, PETER, US;
MORA, MARIROSA, IT;
ROS, IMMACULADA MARGARIT Y., IT;
MAIONE, DOMENICO, IT;
BENSI, GUILIANO, IT;
...

(74) Agent: BORDEN LADNER GERVAIS LLP

(54) Titre : COMPOSITIONS IMMUNOGENES POUR BACTERIES A GRAM POSITIF TELLES QUE STREPTOCOCCUS AGALACTIAE

(54) Title: IMMUNOGENIC COMPOSITIONS FOR GRAM POSITIVE BACTERIA SUCH AS STREPTOCOCCUS AGALACTIAE

(57) Abrégé/Abstract:
The invention relates to the identification of a new adhesin islands within the genomes of several Group A and Group B Streptococcus serotypes and isolates. The adhesin islands are thought to encode surface proteins which are important in the bacteria's virulence. Thus, the adhesin island proteins of the invention may be used in immunogenic compositions for prophylactic or therapeutic immunization against GAS or GBS infection. For example, the invention may include an immunogenic composition comprising one or more of the discovered adhesin island proteins.

CA 2575548 A1 2006/07/27

(21) 2 575 548

(13) A1

(72) **Inventeurs(suite)/Inventors(continued)**: RINAUDO, DANIELA, IT; MASIGNANI, VEGA, IT; BAROCCHI, MICHELLE, IT; RAPPULOI, RINO, IT

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 July 2006 (27.07.2006)

PCT

(10) International Publication Number
WO 2006/078318 A2

(51) International Patent Classification:
A61K 39/02 (2006.01)

(21) International Application Number:
PCT/US2005/027239

(22) International Filing Date: 29 July 2005 (29.07.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/592,805	29 July 2004 (29.07.2004)	US
60/609,833	13 September 2004 (13.09.2004)	US
60/616,833	8 October 2004 (08.10.2004)	US
60/633,418	7 December 2004 (07.12.2004)	US
60/640,069	30 December 2004 (30.12.2004)	US
60/660,321	11 March 2005 (11.03.2005)	US
60/673,754	22 April 2005 (22.04.2005)	US
60/693,001	21 June 2005 (21.06.2005)	US
60/695,453	1 July 2005 (01.07.2005)	US
60/697,643	11 July 2005 (11.07.2005)	US

(71) Applicant *(for all designated States except US)*: **CHIRON CORPORATION** [US/US]; 4560 Horton Street, Emeryville, CA 9460-2916 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **TELFORD, John, L.** [IT/IT]; Chiron Vaccines, 1, Via Fiorentina Siena, I-Siena (IT). **GRANDI, Guido** [IT/IT]; CHIRON VACCINES, 1, Via Fiorentia, I-53100 Siena (IT). **LAUER, Peter** [US/US]; 1438 Milvia Street, Berkeley, CA 94709-1917 (US). **MORA, Marirosa** [IT/IT]; CHIRON VACCINES, 1, Via Fiorentina, I-53100 Siena (IT). **ROS, Immaculada, Margarit, Y.** [IT/IT]; CHIRON VACCINES, 1, Via Fiorentina, I-53100 Siena (IT). **MAIONE, Domenico** [IT/IT]; CHIRON VACCINES, 1, Via Fiorentina, I-53100 Siena (IT). **BENSI, Guiliano** [IT/IT]; CHIRON VACCINES, 1, Via Fiorentina, I-53100 Siena (IT). **RINAUDO, Daniela** [IT/IT]; CHIRON VACCINES, 1, Via Fiorentina, I-53100 Siena (IT). **MASIGNANI, Vega** [IT/IT]; CHIRON VACCINES, 1, Via Fiorentina, I-53100 Siena (IT). **BAROCCHI, Michelle** [IT/IT]; CHIRON VACCINES, 1, Via Fiorentina, I-53100 Siena (IT). **RAPPULOI, Rino** [IT/IT]; CHIRON VACINES, 1, Via Fiorentina, I-53100 Siena (IT).

(74) Agent: **HALE, Rebecca, M.**; CHIRON CORPORATION, 4560 Horton Street, Emeryville, CA 94608-2916 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: IMMUNOGENIC COMPOSITIONS FOR GRAM POSITIVE BACTERIA SUCH AS STREPTOCOCCUS AGALACTIAE

(57) Abstract: The invention relates to the identification of a new adhesin islands within the genomes of several Group A and Group B Streptococcus serotypes and isolates. The adhesin islands are thought to encode surface proteins which are important in the bacteria's virulence. Thus, the adhesin island proteins of the invention may be used in immunogenic compositions for prophylactic or therapeutic immunization against GAS or GBS infection. For example, the invention may include an immunogenic composition comprising one or more of the discovered adhesin island proteins.

WO 2006/078318 A2

DEMANDE OU BREVET VOLUMINEUX

LA PRÉSENTE PARTIE DE CETTE DEMANDE OU CE BREVET COMPREND PLUS D'UN TOME.

CECI EST LE TOME 1 DE 2
CONTENANT LES PAGES 1 À 202

NOTE : Pour les tomes additionels, veuillez contacter le Bureau canadien des brevets

JUMBO APPLICATIONS/PATENTS

THIS SECTION OF THE APPLICATION/PATENT CONTAINS MORE THAN ONE VOLUME

THIS IS VOLUME 1 OF 2
CONTAINING PAGES 1 TO 202

NOTE: For additional volumes, please contact the Canadian Patent Office

NOM DU FICHIER / FILE NAME :

NOTE POUR LE TOME / VOLUME NOTE:

IMMUNOGENIC COMPOSITIONS FOR GRAM POSITIVE BACTERIA SUCH AS *STREPTOCOCCUS AGALACTIAE*

FIELD OF THE INVENTION

The invention relates to the identification of adhesin islands within the genome *Streptococcus agalactiae* ("GBS") and the use of adhesin island amino acid sequences encoded by these adhesin islands in compositions for the treatment or prevention of GBS infection. Similar sequences have been identified in other Gram positive bacteria. The invention further includes immunogenic compositions comprising adhesin island amino acid sequences of Gram positive bacteria for the treatment or prevention of infection of Gram positive bacteria. Preferred immunogenic compositions of the invention include an adhesin island surface protein which may be formulated or purified in an oligomeric or pilus form.

BACKGROUND OF THE INVENTION

GBS has emerged in the last 20 years as the major cause of neonatal sepsis and meningitis that affects 0.5 – 3 per 1000 live births, and an important cause of morbidity among older age groups affecting 5 – 8 per 100,000 of the population. Current disease management strategies rely on intrapartum antibiotics and neonatal monitoring which have reduced neonatal case mortality from >50% in the 1970's to less than 10% in the 1990's. Nevertheless, there is still considerable morbidity and mortality and the management is expensive. 15 – 35% of pregnant women are asymptomatic carriers and at high risk of transmitting the disease to their babies. Risk of neonatal infection is associated with low serotype specific maternal antibodies and high titers are believed to be protective. In addition, invasive GBS disease is increasingly recognized in elderly adults with underlying disease such as diabetes and cancer.

The "B" in "GBS" refers to the Lancefield classification, which is based on the antigenicity of a carbohydrate which is soluble in dilute acid and called the C carbohydrate. Lancefield identified 13 types of C carbohydrate, designated A to O, that could be serologically differentiated. The organisms that most commonly infect humans are found in groups A, B, D, and G. Within group B, strains can be divided into at least 9 serotypes (Ia, Ib, Ia/c, II, III, IV, V, VI, VII and VIII) based on the structure of their polysaccharide capsule. In the past, serotypes Ia, Ib, II, and III were equally prevalent in normal vaginal carriage and early onset sepsis in newborns. Type V GBS has emerged as an important cause of GBS infection in the USA, however, and strains of types VI and VIII have become prevalent among Japanese women.

The genome sequence of a serotype V strain 2603 V/R has been published (See Tettelin *et al.* (2002) *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.182380799) and various polypeptides for use a vaccine antigens have been identified (WO 02/34771). The vaccines currently in clinical trials, however, are based primarily on polysaccharide antigens. These suffer from serotype-specificity and poor immunogenicity, and so there is a need for effective vaccines against *S.agalactiae* infection.

PCT/US2005/027239

S. agalactiae is classified as a gram positive bacterium, a collection of about 21 genera of bacteria that colonize humans, have a generally spherical shape, a positive Gram stain reaction and lack endospores. Gram positive bacteria are frequent human pathogens and include *Staphylococcus* (such as *S. aureus*), *Streptococcus* (such as *S. pyogenes* (GBS), *S. pyogenes* (GAS), *S. pneumonaie*, *S. mutans*), *Enterococcus* (such as *E. faecalis* and *E. faecium*), *Clostridium* (such as *C. difficile*), *Listeria* (such as *L. monocytogenes*) and *Corynebacterium* (such as *C. diphtheria*).

It is an object of the invention to provide further and improved compositions for providing immunity against disease and/or infection of Gram positive bacteria. The compositions are based on the identification of adhesin islands within Streptococcal genomes and the use of amino acid sequences encoded by these islands in therapeutic or prophylactic compositions. The invention further includes compositions comprising immunogenic adhesin island proteins within other Gram positive bacteria in therapeutic or prophylactic compositions.

SUMMARY OF THE INVENTION

Applicants have identified a new adhesin island, "GBS Adhesin Island 1", "AI-1" or "GBS AI-1", within the genomes of several Group B Streptococcus serotypes and isolates. This adhesin island is thought to encode surface proteins which are important in the bacteria's virulence. In addition, Applicants have discovered that surface proteins within GBS Adhesin Islands form a previously unseen pilus structure on the surface of GBS bacteria. Amino acid sequences encoded by such GBS Adhesin Islands may be used in immunogenic compositions for the treatment or prevention of GBS infection.

A preferred immunogenic composition of the invention comprises an AI-1 surface protein, such as GBS 80, which may be formulated or purified in an oligomeric (pilus) form. In a preferred embodiment, the oligomeric form is a hyperoligomer. Electron micrographs depicting some of the first visualizations of this pilus structure in a wild type GBS strain are shown in Figures 16, 17, 49, and 50. In addition, Applicants have transformed a GBS strain with a plasmid comprising the AI surface protein GBS 80 which resulted in increased production of that AI surface protein. The electron micrographs of this mutant GBS strain in Figures 13 – 15 reveal long, hyper-oligomeric structures comprising GBS 80 which appear to cover portions of the surface of the bacteria and stretch far out into the supernatant. These hyper-oligomeric pilus structures comprising a GBS AI surface protein may be purified or otherwise formulated for use in immunogenic compositions.

GBS AI-1 comprises a series of approximately five open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("AI-1 proteins"). Specifically, AI-1 includes polynucleotide sequences encoding for two or more of GBS 80, GBS 104, GBS 52, SAG0647 and SAG0648. One or more of the AI-1 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the AI-1 open reading frames may be replaced by a sequence having sequence homology (sequence identity) to the replaced ORF.

PCT/US2005/027239

AI-1 typically resides on an approximately 16.1 kb transposon-like element frequently inserted into the open reading frame for *trmA*. One or more of the AI-1 surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. The AI surface proteins of the invention may affect the ability of the GBS bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of GBS to translocate through an epithelial cell layer. Preferably, one or more AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The sortase proteins are thought to be involved in the secretion and anchoring of the LPXTG containing surface proteins. AI-1 may encode at least one surface protein. Alternatively, AI-1 may encode at least two surface proteins and at least one sortase. Preferably, AI-1 encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif or other sortase substrate motif.

The GBS AI-1 protein of the composition may be selected from the group consisting of GBS 80, GBS 104, GBS 52, SAG0647 and SAG0648. GBS AI-1 surface proteins GBS 80 and GBS 104 are preferred for use in the immunogenic compositions of the invention.

In addition to the open reading frames encoding the AI-1 proteins, AI-1 may also include a divergently transcribed transcriptional regulator such as *araC* (*i.e.*, the transcriptional regulator is located near or adjacent to the AI protein open reading frames, but it transcribed in the opposite direction). It is believed that *araC* may regulate the expression of the GBS AI operon. (*See* Korbel et al., Nature Biotechnology (2004) 22(7): 911 – 917 for a discussion of divergently transcribed regulators in *E. coli*).

A second adhesin island, "Adhesin Island-2", "AI-2" or "GBS AI-2", has also been identified in numerous GBS serotypes. Amino acid sequences encoded by the open reading frames of AI-2 may also be used in immunogenic compositions for the treatment or prevention of GBS infection.

GBS AI-2 comprises a series of approximately five open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, AI-2 includes open reading frames encoding for two or more of GBS 67, GBS 59, GBS 150, SAG1405, SAG1406, 01520, 01521, 01522, 01523, 01523, 01524 and 01525. The GBS AI-2 sequences may be divided into two subgroups. In one embodiment, AI-2 includes open reading frames encoding for two or more of GBS 67, GBS 59, GBS 150, SAG1405, and SAG1406. This collection of open reading frames may be generally referred to as GBS AI-2 subgroup 1. Alternatively, AI-2 may include open reading frames encoding for two or more of 01520, 01521, 01522, 01523, 01523, 01524 and 01525. This collection of open reading frames may be generally referred to as GBS AI-2 subgroup 2.

One or more of the AI-2 open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the AI-2 open reading frames may be replaced by a sequence having sequence homology (sequence identity) to the replaced ORF.

PCT/US05/27239

One or more of the AI-2 surface proteins typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. These sortase proteins are thought to be involved in the secretion and anchoring of the LPXTG containing surface proteins. AI-2 may encode for at least one surface protein. Alternatively, AI-2 may encode for at least two surface proteins and at least one sortase. Preferably, AI-2 encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif.

The AI-2 protein of the composition may be selected from the group consisting of GBS 67, GBS 59, GBS 150, SAG1405, SAG1406, 01520, 01521, 01522, 01523, 01523, 01524 and 01525. AI-2 surface proteins GBS 67, GBS 59, and 01524 are preferred AI-2 proteins for use in the immunogenic compositions of the invention. GBS 67 or GBS 59 is particularly preferred.

GBS AI-2 may also include a divergently transcribed transcriptional regulator such as a RofA like protein (for example *rogB*). As in AI-1, *rogB* is thought to regulate the expression of the AI-2 operon.

The GBS AI proteins of the invention may be used in immunogenic compositions for prophylactic or therapeutic immunization against GBS infection. For example, the invention may include an immunogenic composition comprising one or more GBS AI-1 proteins and one or more GBS AI-2 proteins.

The immunogenic compositions may also be selected to provide protection against an increased range of GBS serotypes and strain isolates. For example, the immunogenic composition may comprise a first and second GBS AI protein, wherein a full length polynucleotide sequence encoding for the first GBS AI protein is not present in a genome comprising a full length polynucleotide sequence encoding for the second GBS AI protein. In addition, each antigen selected for use in the immunogenic compositions will preferably be present in the genomes of multiple GBS serotypes and strain isolates. Preferably, each antigen is presnt in the genomes of at least two (*i.e.*, 3, 4, 5, 6, 7, 8, 9, 10, or more) GBS strain isolates. More preferably, each antigen is present in the genomes of at least two (*i.e.*, at least 3, 4, 5 or more) GBS serotypes.

Within GBS AI-1, Applicants have found that Group B Streptococcus surface exposure of GBS 104 is dependent on the concurrent expression of GBS 80. It is thought that GBS 80 is involved in the transport or localization of GBS 104 to the surface of the bacteria. The two proteins may be oligomerized or otherwise chemically or physically associated. It is possible that this association involves a conformational change in GBS 104 that facilitates its transition to the surface of the GBS bacteria. In addition, one or more AI sortases may also be involved in this surface localization and chemical or physical association. Similar relationships are thought to exist within GBS AI-2. The compositions of the invention may therefore include at least two AI proteins, wherein the two AI proteins are physically or chemically associated. Preferably, the two AI proteins form an oligomer. Preferably, one or more of the AI proteins are in a hyper-oligomeric form. In one embodiment, the associated AI proteins may be purified or isolated from a GBS bacteria or recombinant host cell.

PCT/US2005/027239

It is also an object of the invention to provide further and improved compositions for providing prophylactic or therapeutic protection against disease and/or infection of Gram positive bacteria. The compositions are based on the identification of adhesin islands within Streptococcal genomes and the use of amino acid sequences encoded by these islands in therapeutic or prophylactic compositions. The invention further includes compositions comprising immunogenic adhesin island proteins within other Gram positive bacteria in therapeutic or prophylactic compositions. Preferred Gram positive adhesin island proteins for use in the invention may be derived from *Staphylococcus* (such as *S. aureus*), *Streptococcus* (such as *S. agalactiae* (GBS), *S. pyogenes* (GAS), *S. pneumonaie*, *S. mutans*), *Enterococcus* (such as *E. faecalis* and *E. faecium*), *Clostridium* (such as *C. difficile*), *Listeria* (such as *L. monocytogenes*) and *Corynebacterium* (such as *C. diphtheria*). Preferably, the Gram positive adhesin island surface proteins are in oligomeric or hyperologimeric form.

For example, Applicants have identified adhesin islands within the genomes of several Group A Streptococcus serotypes and isolates. These adhesion islands are thought to encode surface proteins which are important in the bacteria's virulence, and Applicants have obtained the first electron micrographs revealing the presence of these adhesin island proteins in hyperoligomeric pilus structures on the surface of Group A Streptococcus.

Group A Streptococcus is a human specific pathogen which causes a wide variety of diseases ranging from pharyngitis and impetigo through life threatening invasive disease and necrotizing fasciitis. In addition, post-streptococcal autoimmune responses are still a major cause of cardiac pathology in children.

Group A Streptococcal infection of its human host can generally occur in three phases. The first phase involves attachment and/or invasion of the bacteria into host tissue and multiplication of the bacteria within the extracellular spaces. Generally this attachment phase begins in the throat or the skin. The deeper the tissue level infected, the more severe the damage that can be caused. In the second stage of infection, the bacteria secretes a soluble toxin that diffuses into the surrounding tissue or even systemically through the vasculature. This toxin binds to susceptible host cell receptors and triggers innappropropriate immune responses by these host cells, resulting in pathology. Because the toxin can diffuse throughout the host, the necrosis directly caused by the GAS toxins may be physically located in sites distant from the bacterial infection. The final phase of GAS infection can occur long after the original bacteria have been cleared from the host system. At this stage, the host's previous immune response to the GAS bacteria due to cross reactivity between epitopes of a GAS surface protein, M, and host tissues, such as the heart. A general review of GAS infection can be found in Principles of Bacterial Pathogeneis, Groisman ed., Chapter 15 (2001).

In order to prevent the pathogenic effects associated with the later stages of GAS infection, an effective vaccine against GAS will preferably facilitate host elimination of the bacteria during the initial attachment and invasion stage.

Isolates of Group A Streptococcus are historically classified according to the M surface protein described aboye. The M protein is surface exposed trypsin-sensitive protein generally

comprising two polypeptide chains complexed in an alpha helical formation. The carboxyl terminus is anchored in the cytoplasmic membrane and is highly conserved among all group A streptococci. The amino terminus, which extend through the cell wall to the cell surface, is responsible for the antigenic variability observed among the 80 or more serotypes of M proteins.

A second layer of classification is based on a variable, trypsin-resistant surface antigen, commonly referred to as the T-antigen. Decades of epidemiology based on M and T serological typing have been central to studies on the biological diversity and disease causing potential of Group A Streptococci. While the M-protein component and its inherent variability have been extensively characterized, even after five decades of study, there is still very little known about the structure and variability of T-antigens. Antisera to define T types is commercially available from several sources, including Sevapharma (http://www.sevapharma.cz/en).

The gene coding for one form of T-antigen, T-type 6, from an M6 strain of GAS (D741) has been cloned and characterized and maps to an approximately 11 kb highly variable pathogenicity island. Schneewind et al., J Bacteriol. (1990) 172(6):3310 – 3317. This island is known as the Fibronectin-binding, Collagen-binding T-antigen (FCT) region because it contains, in addition to the T6 coding gene (*tee6*), members of a family of genes coding for Extra Cellular Matrix (ECM) binding proteins. Bessen et al., Infection & Immunity (2002) 70(3):1159-1167. Several of the protein products of this gene family have been shown to directly bind either fibronectin and/or collagen. *See* Hanski et al., Infection & Immunity (1992) 60(12):5119-5125; Talay et al., Infection & Immunity (1992(60(9):3837-3844; Jaffe et al. (1996) 21(2):373-384; Rocha et al., Adv Exp Med Biol. (1997) 418:737-739; Kreikemeyer et al., J Biol Chem (2004) 279(16):15850-15859; Podbielski et al., Mol. Microbiol. (1999) 31(4):1051-64; and Kreikemeyer et al., Int. J. Med Microbiol (2004) 294(2-3):177-88. In some cases direct evidence for a role of these proteins in adhesion and invasion has been obtained.

Applicants raised antiserum against a recombinant product of the *tee6* gene and used it to explore the expression of T6 in M6 strain 2724. In immunoblot of mutanolysin extracts of this strain, the antiserum recognized, in addition to a band corresponding to the predicted molecular mass of the product, very high molecular weight ladders ranging in mobility from about 100 kDa to beyond the resolution of the 3-8% gradient gels used.

This pattern of high molecular weight products is similar to that observed in immunoblots of the protein components of the pili identified in *Streptococcus agalactiae* (described above) and previously in *Corynebacterium diphtheriae*. Electron microscropy of strain M6_2724 with antisera specific for the product of *tee6* revealed abundant surface staining and long pilus like structures extending up to 700 nanometers from the bacterial surface, revealing that the T6 protein, one of the antigens recognized in the original Lancefiled serotyping system, is located within a GAS Adhesin Island (GAS AI-1) and forms long covalently linked pilus structures.

Applicants have identified at least four different Group A Streptococcus Adhesin Islands. While these GAS AI sequences can be identified in numerous M types, Applicants have surprisingly

discovered a correlation between the four main pilus subunits from the four different GAS AI types and specific T classifications. While other trypsin-resistant surface exposed proteins are likely also implicated in the T classification designations, the discovery of the role of the GAS adhesin islands (and the associated hyper-oligomeric pilus like structures) in T classification and GAS serotype variance has important implications for prevention and treatment of GAS infections. Applicants have identified protein components within each of the GAS adhesin islands which are associated with the pilus formation. These proteins are believed to be involved in the bacteria's initial adherence mechanisms. Immunological recognition of these proteins may allow the host immune response to slow or prevent the bacteria's transition into the more pathogenic later stages of infection.

In addition, Applicants have discovered that the GBS pili structures appear to be implicated in the formation of biofilms (populations of bacteria growing on a surface, often enclosed in an exopolysaccharide matrix). Biofilms are generally associated with bacterial resistance, as antibiotic treatments and host immune response are frequently unable to erradicate all of the bacteria components of the biofilm. Direction of a host immune response against surface proteins exposed during the first steps of bacterial attachment (i.e., before complete biofilm formation) is preferable.

The invention therefore provides for improved immunogenic compositions against GAS infection which may target GAS bacteria during their initial attachment efforts to the host epithelial cells and may provide protection against a wide range of GAS serotypes. The immunogenic compositions of the invention include GAS AI surface proteins which may be formulated in an oligomeric, or hyperoligomeric (pilus) form. The immunogenic compositions of the invention may include one or more GAS AI surface proteins. The invention also includes combinations of GAS AI surface proteins. Combinations of GAS AI surface proteins may be selected from the same adhesin island or they may be selected from different GAS adhesin islands.

Amino acid sequence encoded by such GAS Adhesin Islands may be used in immunogenic compositions for the treatment or prevention of GAS infection. Preferred immunogenic compositions of the invention comprise a GAS AI surface protein which has been formulated or purified in an oligomeric (pilus) form. In a preferred embodiment, the oligomeric form is a hyperoligomer.

GAS Adhesin Islands generally include a series of open reading frames within a GAS genome that encode for a collection of surface proteins and sortases. A GAS Adhesin Island may encode for an amino acid sequence comprising at least one surface protein. The Adhesin Island, therefore, may encode at least one surface protein. Alternatively, a GAS Adhesin Island may encode for at least two surface proteins and at least one sortase. Preferably, a GAS Adhesin Island encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. One or more GAS AI surface proteins may participate in the formation of a pilus structure on the surface of the Gram positive bacteria.

PCT Published Page

GAS Adhesin Islands of the invention preferably include a divergently transcribed transcriptional regulator. The transcriptional regulator may regulate the expression of the GAS AI operon. Examples of transcriptional regulators found in GAS AI sequences include *RofA* and *Nra*.

The GAS AI surface proteins may bind or otherwise adhere to fibrinogen, fibronectin, or collagen. One or more of the GAS AI surface proteins may comprise a fimbrial structural subunit.

One or more of the GAS AI surface proteins may include an LPXTG motif or other sortase substrate motif. The LPXTG motif may be followed by a hydrophobic region and a charged C terminus, which are thought to retard the protein in the cell membrane to facilitate recognition by the membrane-localized sortase. See Barnett, et al., J. Bacteriology (2004) 186 (17): 5865-5875.

GAS AI sequences may be generally categorized as Type 1, Type 2, Type 3, or Type 4, depending on the number and type of sortase sequences within the island and the percentage identity of other proteins (with the exception of RofA and cpa) within the island. Schematics of the GAS adhesin islands are set forth in FIGURE 51A and FIGURE 162. "GAS Adhesin Island-1 or "GAS AI-1" comprises a series of approximately five open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("GAS AI-1 proteins"). GAS AI-1 preferably comprises surface proteins, a srtB sortase and a rofA divergently transcribed transcriptional regulator. GAS AI-1 surface proteins may include a fibronectin binding protein, a collagen adhesion protein and a fimbrial structural subunit. The fimbrial structural subunit (also known as *tee6*) is thought to form the shaft portion of the pilus like structure, while the collagen adhesion protein (Cpa) is thought to act as an accessory protein facilitating the formation of the pilus structure, exposed on the surface of the bacterial capsule.

Specifically, GAS AI-1 includes polynucleotide sequences encoding for two or more of M6_Spy0157, M6_Spy0158, M6_Spy0159, M6_Spy0160, M6_Spy0161. The GAS AI-1 may also include polynucleotide sequences encoding for any one of CDC SS 410_fimbrial, ISS3650_fimbrial, DSM2071_fimbrial

A preferred immunogenic composition of the invention comprises a GAS AI-1 surface protein which may be formulated or purified in an oligomeric (pilus) form. In a preferred embodiment, the oligomeric form is a hyperoligomer. The immunogenic composition of the invention may alternatively comprise an isolated GAS AI-1 surface protein in oligomeric (pilus) form. The oligomer or hyperoligomeric pilus structures comprising GAS AI-1 surface proteins may be purified or otherwise formulated for use in immunogenic compositions.

One or more of the GAS AI-1 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the GAS AI-1 open reading frames may be replaced by a sequence having sequence homology (sequence identity) to the replaced ORF.

One or more of the GAS AI-1 surface proteins typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. These sortase proteins are thought to be involved in the secretion and anchoring of the LPXTG containing surface proteins. GAS AI-1 may

encode for at least one surface protein. Alternatively, GAS AI-1 may encode for at least two surface proteins and at least one sortase. Preferably, GAS AI-1 encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif.

GAS AI-1 preferably includes a srtB sortase. GAS srtB sortases may preferably anchor surface proteins with an LPSTG motif (SEQ ID NO: 166), particularly where the motif is followed by a serine.

The GAS AI-1 protein of the composition may be selected from the group consisting of M6_Spy0157, M6_Spy0158, M6_Spy0159, M6_Spy0160 M6_Spy0161, CDC SS 410_fimbrial, ISS3650_fimbrial, and DSM2071_fimbrial. GAS AI-1 surface proteins M6_Spy0157 (a fibronectin binding protein), M6_Spy0159 (a collagen adhesion protein, Cpa), M6_Spy0160 (a fimbrial structural subunit, *tee6*), CDC SS 410_fimbrial (a fimbrial structural subunit), ISS3650_fimbrial (a fimbrial structural subunit), and DSM2071_fimbrial (a fimbrial structural subunit) are preferred GAS AI-1 proteins for use in the immunogenic compositions of the invention. The fimbrial structural subunit tee6 and the collagen adhesion protein Cpa are preferred GAS AI –1 surface proteins. Preferably, each of these GAS AI-1 surface proteins includes an LPXTG sortase substrate motif, such as LPXTG (SEQ ID NO: 122) or LPXSG (SEQ ID NO: 134) (conservative replacement of threonine with serine).

In addition to the open reading frames encoding the GAS AI-1 proteins, GAS AI-1 may also include a divergently transcribed transcriptional regulator such as *rofA* (*i.e.*, the transcriptional regulator is located near or adjacent to the GAS AI protein open reading frames, but it transcribed in the opposite direction).

The GAS AI-1 surface proteins may be used alone, in combination with other GAS AI-1 surface proteins or in combination with other GAS AI surface proteins. Preferably, the immunogenic compositions of the invention include the GAS AI-1 fimbrial structural subunit (*tee6*) and the GAS AI-1 collagen binding protein. Still more preferably, the immunogenic compositions of the invention include the GAS AI-1 fimbrial structural subunit (*tee6*).

A second GAS adhesion island, "GAS Adhesin Island-2" or "GAS AI-2," has also been identified in GAS serotypes. Amino acid sequences encoded by the open reading frames of GAS AI-2 may also be used in immunogenic compositions for the treatment or prevention of GAS infection.

A preferred immunogenic composition of the invention comprises a GAS AI-2 surface protein which may be formulated or purified in an oligomeric (pilus) form. In a preferred embodiment, the oligomeric form is a hyperoligomer. A preferred immunogenic composition of the invention alternatively comprises an isolated GAS AI-2 surface protein in oligomeric (pilus) form. The oligomer or hyperoligomeric pilus structures comprising GAS AI-2 surface proteins may be purified or otherwise formulated for use in immunogenic compositions.

GAS AI-2 comprises a series of approximately eight open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("GAS AI-2 proteins").

GAS AI-2 preferably comprises surface proteins, a srtB sortase, a srtC1 sortase and a rofA divergently transcribed transcriptional regulator.

Specifically, GAS AI-2 includes polynucleotide sequences encoding for two or more of GAS15, Spy0127, GAS16, GAS17, GAS18, Spy0131, Spy0133, and GAS20.

One or more of the GAS AI-2 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the GAS AI-2 open reading frames may be replaced by a sequence having sequence homology (sequence identity) to the replaced ORF.

One or more of the GAS AI-2 surface proteins typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. These sortase proteins are thought to be involved in the secretion and anchoring of the LPXTG containing surface proteins. GAS AI-2 may encode for at least one surface protein. Alternatively, GAS AI-2 may encode for at least two surface proteins and at least one sortase. Preferably, GAS AI-2 encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif.

GAS AI-2 preferably includes a srtB sortase and a srtC1 sortase. As discussed above, GAS srtB sortases may preferably anchor surface proteins with an LPSTG motif (SEQ ID NO: 166), particularly where the motif is followed by a serine. GAS srtC1 sortase may preferentially anchor surface proteins with a V(P/V)PTG (SEQ ID NO:167) motif. GAS srtC1 may be differentially regulated by *rofA*.

The GAS AI-2 protein of the composition may be selected from the group consisting of GAS15, Spy0127, GAS16, GAS17, GAS18, Spy0131, Spy0133, and GAS20. GAS AI-2 surface proteins GAS15 (Cpa), GAS16 (thought to be a fimbrial protein, M1_128), GAS18 (M1_Spy0130), and GAS20 are preferred for use in the immunogenic compositions of the invention. GAS 16 is thought to form the shaft portion of the pilus like structure, while GAS 15 (the collagen adhesion protein Cpa) and GAS 18 are thought to act as accessory proteins facilitating the formation of the pilus structure, exposed on the surface of the bacterial capsule. Preferably, each of these GAS AI-2 surface proteins includes an LPXTG sortase substrate motif, such as LPXTG (SEQ ID NO: 122), VVXTG (SEQ ID NO: 135), or EVXTG (SEQ ID NO: 136).

In addition to the open reading frames encoding the GAS AI-2 proteins, GAS AI-2 may also include a divergently transcribed transcriptional regulator such as *rofA* (*i.e.*, the transcriptional regulator is located near or adjacent to the GAS AI protein open reading frames, but it transcribed in the opposite direction).The GAS AI-2 surface proteins may be used alone, in combination with other GAS AI-2 surface proteins or in combination with other GAS AI surface proteins. Preferably, the immunogenic compositions of the invention include the GAS AI-2 fimbrial protein (GAS 16), the GAS AI-2 collagen binding protein (GAS 15) and GAS 18 (M1_Spy0130). More preferably, the immunogenic compositions of the invention include the GAS AI-2 fimbrial protein (GAS 16).

A third GAS adhesion island, "GAS Adhesin Island-3" or "GAS AI-3," has also been identified in numerous GAS serotypes. Amino acid sequences encoded by the open reading frames of

GAS AI-3 may also be used in immunogenic compositions for the treatment or prevention of GAS infection.

A preferred immunogenic composition of the invention comprises a GAS AI-3 surface protein which may be formulated or purified in an oligomeric (pilus) form. In a preferred embodiment, the oligomeric form is a hyperoligomer. A preferred immunogenic composition of the invention alternatively comprises an isolated GAS AI-3 surface protein in oligomeric (pilus) form. The oligomer or hyperoligomeric pilus structures comprising GAS AI-3 surface proteins may be purified or otherwise formulated for use in immunogenic compositions.GAS AI-3 comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("GAS AI-3 proteins"). GAS AI-3 preferably comprises surface proteins, a srtC2 sortase, and a Negative transcriptional regulator (Nra) divergently transcribed transcriptional regulator. GAS AI-3 surface proteins may include a collagen binding protein, a fimbrial protein, and a F2 like fibronectin-binding protein. GAS AI-3 surface proteins may also include a hypothetical surface protein. The fimbrial protein is thought to form the shaft portion of the pilus like structure, while the collagen adhesion protein (Cpa) and the hypothetical surface protein are thought to act as accessory proteins facilitating the formation of the pilus structure, exposed on the surface of the bacterial capsule. Preferred AI-3 surface proteins include the fimbrial proein, the collagen binding protein and the hypothetical protein. Preferably, each of these GAS AI-3 surface proteins include an LPXTG sortase substrate motif, such as LPXTG (SEQ ID NO: 122), VPXTG (SEQ ID NO: 137), QVXTG (SEQ ID NO: 138) or LPXAG (SEQ ID NO: 139).

Specifically, GAS AI-3 includes polynucleotide sequences encoding for two or more of SpyM3_0098, SpyM3_0099, SpyM3_0100, SpyM3_0101, SpyM3_0102, SpyM3_0103, SpyM3_0104, Sps0100, Sps0101, Sps0102, Sps0103, Sps0104, Sps0105, Sps0106, orf78, orf79, orf80, orf81, orf82, orf83, orf84, spyM18_0126, spyM18_0127, spyM18_0128, spyM18_0129, spyM18_0130, spyM18_0131, spyM18_0132, SpyoM01000156, SpyoM01000155, SpyoM01000154, SpyoM01000153, SpyoM01000152, SpyoM01000151, SpyoM01000150, SpyoM01000149, ISS3040_fimbrial, ISS3776_fimbrial, and ISS4959_fimbrial. In one embodiment, GAS AI-3 may include open reading frames encoding for two or more of SpyM3_0098, SpyM3_0099, SpyM3_0100, SpyM3_0101, SpyM3_0102, SpyM3_0103, and SpyM3_0104. Alternatively, GAS AI-3 may include open reading frames encoding for two or more of Sps0100, Sps0101, Sps0102, Sps0103, Sps0104, Sps0105, and Sps0106. Alternatively, GAS AI-3 may include open reading frames encoding for two or more of orf78, orf79, orf80, orf81, orf82, orf83, and orf84. Alternatively, GAS AI-3 may include open reading frames encoding for two or more of spyM18_0126, spyM18_0127, spyM18_0128, spyM18_0129, spyM18_0130, spyM18_0131, and spyM18_0132. Alternatively, GAS AI-3 may include open reading frames encoding for two or more of SpyoM01000156, SpyoM01000155, SpyoM01000154, SpyoM01000153, SpyoM01000152, SpyoM01000151, SpyoM01000150, and SpyoM01000149. Alternatively, GAS AI-1 may also include polynucleotide sequences encoding for any one of ISS3040_fimbrial, ISS3776_fimbrial, and ISS4959_fimbrial.

One or more of the GAS AI-3 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the GAS AI-3 open reading frames may be replaced by a sequence having sequence homology (sequence identity) to the replaced ORF.

One or more of the GAS AI-3 surface proteins typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. These sortase proteins are thought to be involved in the secretion and anchoring of the LPXTG containing surface proteins. GAS AI-3 may encode for at least one surface protein. Alternatively, GAS AI-3 may encode for at least two surface proteins and at least one sortase. Preferably, GAS AI-3 encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif.

GAS AI-3 preferably includes a srtC2 type sortase. GAS srtC2 type sortases may preferably anchor surface proteins with a QVPTG (SEQ ID NO: 140) motif, particularly when the motif is followed by a hydrophobic region and a charged C terminus tail. GAS SrtC2 may be differentially regulated by Nra.

The GAS AI-3 protein of the composition may be selected from the group consisting of SpyM3_0098, SpyM3_0099, SpyM3_0100, SpyM3_0101, SpyM3_0102, SpyM3_0103, SpyM3_0104, Sps0100, Sps0101, Sps0102, Sps0103, Sps0104, Sps0105, Sps0106, orf78, orf79, orf80, orf81, orf82, orf83, orf84, spyM18_0126, spyM18_0127, spyM18_0128, spyM18_0129, spyM18_0130, spyM18_0131, spyM18_0132, SpyoM01000156, SpyoM01000155, SpyoM01000154, SpyoM01000153, SpyoM01000152, SpyoM01000151, SpyoM01000150, SpyoM01000149, ISS3040_fimbrial, ISS3776_fimbrial, and ISS4959_fimbrial. GAS AI-3 surface proteins SpyM3_0098, SpyM3_0100, SpyM3_0102, SpyM3_0104, SPs0100, SPs0102, SPs0104, SPs0106, orf78, orf80, orf82, orf84, spyM18_0126, spyM18_0128, spyM18_0130, spyM18_0132, SpyoM01000155, SpyoM01000153, SpyoM01000151, SpyoM01000149, ISS3040_fimbrial, ISS3776_fimbrial, and ISS4959_fimbrial are preferred GAS AI-3 proteins for use in the immunogenic compositions of the invention.

In addition to the open reading frames encoding the GAS AI-3 proteins, GAS AI-3 may also include a transcriptional regulator such as *Nra*.

GAS AI-3 may also include a LepA putative signal peptidase I protein.

The GAS AI-3 surface proteins may be used alone, in combination with other GAS AI-3 surface proteins or in combination with other GAS AI surface proteins. Preferably, the immunogenic compositions of the invention include the GAS AI-3 fimbrial protein, the GAS AI-3 collagen binding protein, the GAS AI-3 surface protein (such as SpyM3_0102, M3_Sps0104, M5_orf82, or spyM18_0130), and fibronectin binding protein PrtF2. More preferably, the immunogenic compositions of the invention include the GAS AI-3 fimbrial protein, the GAS AI-3 collagen binding protein, and the GAS AI-3 surface protein. Still more preferably, the immunogenic compositions of the invention include the GAS AI-3 fimbrial protein.

Representative examples of the GAS AI-3 fimbrial protein include SpyM3_0100, M3_Sps0102, M5_orf80, spyM18_128, SpyoM01000153, ISS3040_fimbrial, ISS3776_fimbrial, ISS4959_fimbrial.

Representative examples of the GAS AI-3 collagen binding protein include SpyM3_0098, M3_Sps0100, M5_orf 78, spyM18_0126, and SpyoM01000155.

Representative examples of the GAS AI-3 fibronectin binding protein PrtF2 include SpyM3_0104, M3_Sps0106, M5_orf84 and spyM18_0132, and SpyoM01000149.

A fourth GAS adhesion island, "GAS Adhesin Island-4" or "GAS AI-4," has also been identified in GAS serotypes. Amino acid sequences encoded by the open reading frames of GAS AI-4 may also be used in immunogenic compositions for the treatment or prevention of GAS infection.

A preferred immunogenic composition of the invention comprises a GAS AI-4 surface protein which may be formulated or purified in an oligomeric (pilus) form. In a preferred embodiment, the oligomeric form is a hyperoligomer. A preferred immunogenic composition of the invention alternatively comprises an isolated GAS AI-4 surface protein in oligomeric (pilus) form. The oligomer or hyperoligomeric pilus structures comprising GAS AI-3 surface proteins may be purified or otherwise formulated for use in immunogenic compositions. The oligomeric or hyperoligomeric pilus structures comprising GAS AI-4 surface proteins may be purified or otherwise formulated for use in immunogenic compositions.

GAS AI-4 comprises a series of approximately eight open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("GAS AI-4 proteins"). This GAS adhesin island 4 ("GAS AI-4") comprises surface proteins, a srtC2 sortase, and a RofA regulatory protein. GAS AI-4 surface proteins within may include a fimbrial protein, F1 and F2 like fibronectin-binding proteins, and a capsular polysaccharide adhesion protein (cpa). GAS AI-4 surface proteins may also include a hypothetical surface protein in an open reading frame (orf).

The fimbral protein (EftLSL) is thought to form the shaft portion of the pilus like structure, while the collagen adhesion protein (Cpa) and the hypothetical protein are thought to act as accessory proteins facilitating the formation of the pilus structure, exposed on the surface of the bacterial capsule. Preferably, each of these GAS AI-4 surface proteins include an LPXTG sortase substrate motif, such as LPXTG (SEQ ID NO: 122), VPXTG (SEQ ID NO: 137), QVXTG (SEQ ID NO: 138) or LPXAG (SEQ ID NO: 139).

Specifically, GAS AI-4 includes polynucleotide sequences encoding for two or more of 19224134, 19224135, 19224136, 19224137, 19224138, 19224139, 19224140, and 19224141. A GAS AI-4 polynucleotide may also include polynucleotide sequences encoding for any one of 20010296_fimbrial, 20020069_fimbrial, CDC SS 635_fimbrial, ISS4883_fimbrial, ISS4538_fimbrial. One or more of the GAS AI-4 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the GAS AI-4 open reading frames may be replaced by a sequence having sequence homology (sequence identity) to the replaced ORF.

One or more of the GAS AI-4 surface proteins typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. These sortase proteins are thought to be involved in the secretion and anchoring of the LPXTG containing surface proteins. GAS AI-4 may encode for at least one surface protein. Alternatively, GAS AI-4 may encode for at least two surface proteins and at least one sortase. Preferably, GAS AI-4 encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif.

GAS AI-4 includes a SrtC2 type sortase. GAS SrtC2 type sortases may preferably anchor surface proteins with a QVPTG (SEQ ID NO: 140) motif, particularly when the motif is followed by a hydrophobic region and a charged C terminus tail.

The GAS AI-4 protein of the composition may be selected from the group consisting of 19224134, 19224135, 19224136, 19224137, 19224138, 19224139, 19224140, 19224141, 20010296_fimbrial, 20020069_fimbrial, CDC SS 635_fimbrial, ISS4883_fimbrial, and ISS4538_fimbrial. GAS AI-4 surface proteins 19224134, 19224135, 19224137, 19224139, 19224141, 20010296_fimbrial, 20020069_fimbrial, CDC SS 635_fimbrial, ISS4883_fimbrial, ISS4538_fimbrial are preferred proteins for use in the immunogenic compositions of the invention.

In addition to the open reading frames encoding the GAS AI-4 proteins, GAS AI-4 may also include a divergently transcribed transcriptional regulator such as *RofA* (*i.e.*, the transcriptional regulator is located near or adjacent to the AI protein open reading frames, but it transcribed in the opposite direction.

GAS AI-4 may also include a LepA putative signal peptidase I protein and a MsmRL protein. The GAS AI-4 surface proteins may be used alone, in combination with other GAS AI-4 surface proteins or in combination with other GAS AI surface proteins. Preferably, the immunogenic compositions of the invention include the GAS AI-4 fimbrial protein (EftLSL or 20010296_fimbrial, 20020069_fimbrial, CDC SS 635_fimbrial, ISS4883_fimbrial, or ISS4538_fimbrial), the GAS AI-4 collagen binding protein, the GAS AI-4 surface protein (such as M12 isolate A735 orf 2), and fibronectin binding protein PrtF1 and PrtF2. More preferably, the immunogenic compositions of the invention include the GAS AI-4 fimbrial protein, the GAS AI-4 collagen binding protein, and the GAS AI-4 surface protein. Still more preferably, the immunogenic compositions of the invention include the GAS AI-4 fimbrial protein.

The GAS AI proteins of the invention may be used in immunogenic compositions for prophylactic or therapeutic immunization against GAS infection. For example, the invention may include an immunogenic composition comprising one or more GAS AI-1 proteins and one or more of any of GAS AI-2, GAS AI-3, or GAS AI-4 proteins. For example, the invention includes an immunogenic composition comprising at least two GAS AI proteins where each protein is selected from a different GAS adhesin island. The two GAS AI proteins may be selected from one of the following GAS AI combinations: GAS AI-1 and GAS AI-2; GAS AI-1 and GAS AI-3; GAS AI-1 and GAS AI-4; GAS AI-2 and GAS AI-3; GAS AI-2 and GAS AI-4; and GAS AI 3 and GAS AI-4. Preferably the combination includes fimbrial proteins from one or more GAS adhesin islands.

The immunogenic compositions may also be selected to provide protection against an increased range of GAS serotypes and strain isolates. For example, the immunogenic composition may comprise a first and second GAS AI protein, wherein a full length polynucleotide sequence encoding for the first GAS AI protein is not present in a genome comprising a full length polynucleotide sequence encoding for the second GAS AI protein. In addition, each antigen selected for use in the immunogenic compositions will preferably be present in the genomes of multiple GAS serotypes and strain isolates. Preferably, each antigen is present in the genomes of at least two (*i.e.*, 3, 4, 5, 6, 7, 8, 9, 10, or more) GAS strain isolates. More preferably, each antigen is present in the genomes of at least two (*i.e.*, at least 3, 4, 5, or more) GAS serotypes.

Applicants have also identified adhesin islands within the genome of *Streptococcus pneumoniae*. These adhesion islands are thought to encode surface proteins which are important in the bacteria's virulence. Amino acid sequence encoded by such *S. pneumoniae* Adhesin Islands may be used in immunogenic compositions for the treatment or prevention of *S. pneumoniae* infection. Preferred immunogenic compositions of the invention comprise a *S. pneumoniae* AI surface protein which has been formulated or purified in an oligomeric (pilus) form. In a preferred embodiment, the oligomeric form is a hyperoligomer. A preferred immunogenic composition of the invention alternatively comprises an isolated *S. pneumoniae* surface protein in oligomeric (pilus) form. The oligomer or hyperoligomeric pilus structures comprising *S. pneumoniae* surface proteins may be purified or otherwise formulated for use in immunogenic compositions.

The *S. pneumoniae* Adhesin Islands generally include a series of open reading frames within a S. pneumoniae genome that encode for a collection of surface proteins and sortases. A S. pneumoniae Adhesin Island may encode for an amino acid sequence comprising at least one surface protein. Alternatively, the S. pneumoniae Adhesin Island may encode for at least two surface proteins and at least one sortase. Preferably, a S. pneumoniae Adhesin Island encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPTXG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. One or more S. pneumoniae AI surface proteins may participate in the formation of a pilus structure on the surface of the S. pneumoniae bacteria.

The S. pneumoniae Adhesin Islands of the invention preferably include a divergently transcribed transcriptional regulator. The transcriptional regulator may regulate the expression of the S. pneumonaie AI operon. An example of a transcriptional regulator found in S. pneumoniae AI sequences is *rlrA*.

A schematic of the organization of a S. pneumoniae AI locus is provided in Figure 137. The locus comprises open reading frames encoding a transcriptional regulator (rlrA), cell wall surface proteins (rrgA, rrgB, rrgC) and sortases (srt B, srtC, srtD).

S. pneumoniae AI sequences may be generally divided into two groups of homology, S. pneuamoniae AI-a and AI-b. S. pneumoniae strains that comprise AI-a include 14 CSR 10, 19A

Hungary 6, 23 F Poland 15, 670, 6B Finland 12, and 6B Spain 2. S. pneumoniae AI strains that comprise AI-b include 19F Taiwan 14, 9V Spain 3, 23F Taiwan 15 and TIGR 4.

S. pneumoniae AI from TIGR4 comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("*S. pneumoniae* AI proteins"). Specifically, *S. pneumoniae* AI from TIGR4 includes polynucleotide sequences encoding for two or more of SP0462, SP0463, SP0464, SP0465, SP0466, SP0467, and SP0468.

One or more of the *S. pneumoniae* AI from TIGR4 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* AI from TIGR4 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

S. pneumoniae strain 670 AI comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("*S. pneumoniae* AI proteins"). Specifically, *S. pneumoniae* strain 670 AI includes polynucleotide sequences encoding for two or more of orf1_670, orf3_670, orf4_670, orf5_670, orf6_670, orf7_670, and orf8_670.

One or more of the *S. pneumoniae* strain 670 AI polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* strain 670 AI open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

S. pneumoniae AI from 14 CSR10 comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("*S. pneumoniae* AI proteins"). Specifically, *S. pneumoniae* AI from 14 CSR10 includes polynucleotide sequences encoding for two or more of ORF2_14CSR, ORF3_14CSR, ORF4_14CSR, ORF5_14CSR, ORF6_14CSR, ORF7_14CSR, and ORF8_14CSR.

One or more of the *S. pneumoniae* AI from 14 CSR10 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* AI from 14 CSR10 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

S. pneumoniae AI from 19A Hungary 6 comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("*S. pneumoniae* AI proteins"). Specifically, *S. pneumoniae* AI from 19A Hungary 6 includes polynucleotide sequences encoding for two or more of ORF2_19AH, ORF3_19AH, ORF4_19AH, ORF5_19AH, ORF6_19AH, ORF7_19AH, and ORF8_19AH.

One or more of the *S. pneumoniae* AI from 19A Hungary 6 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* AI from 19A Hungary 6 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

PC *S. pneumoniae* AI from 19F Taiwan 14 comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("*S. pneumoniae* AI proteins"). Specifically, *S. pneumoniae* AI from 19F Taiwan 14 includes polynucleotide sequences encoding for two or more of ORF2_19FTW, ORF3_19FTW, ORF4_19FTW, ORF5_19FTW, ORF6_19FTW, ORF7_19FTW, and ORF8_19FTW.

One or more of the *S. pneumoniae* AI from 19F Taiwan 14 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* AI from 19F Taiwan 14 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

S. pneumoniae AI from 23F Poland 16 comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("*S. pneumoniae* AI proteins"). Specifically, *S. pneumoniae* AI from 23F Poland 16 includes polynucleotide sequences encoding for two or more of ORF2_23FP, ORF3_23FP, ORF4_23FP, ORF5_23FP, ORF6_23FP, ORF7_23FP, and ORF8_23FP.

One or more of the *S. pneumoniae* AI from 23F Poland 16 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* AI from 23F Poland 16 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

S. pneumoniae AI from 23F Taiwan 15 comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("*S. pneumoniae* AI proteins"). Specifically, *S. pneumoniae* AI from 23F Taiwan 15 includes polynucleotide sequences encoding for two or more of ORF2_23FTW, ORF3_23FTW, ORF4_23FTW, ORF5_23FTW, ORF6_23FTW, ORF7_23FTW, and ORF8_23FTW.

One or more of the *S. pneumoniae* AI from 23F Taiwan 15 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* AI from 23F Taiwan 15 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

S. pneumoniae AI from 6B Finland 12 comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("*S. pneumoniae* AI proteins"). Specifically, *S. pneumoniae* AI from 6B Finland 12 includes polynucleotide sequences encoding for two or more of ORF2_6BF, ORF3_6BF, ORF4_6BF, ORF5_6BF, ORF6_6BF, ORF7_6BF, and ORF8_6BF.

One or more of the *S. pneumoniae* AI from 6B Finland 12 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* AI from 6B Finland 12 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

S. pneumoniae AI from 6B Spain 2 comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases

("*S. pneumoniae* AI proteins"). Specifically, *S. pneumoniae* AI from 6B Spain 2 includes polynucleotide sequences encoding for two or more of ORF2_6BSP, ORF3_6BSP, ORF4_6BSP, ORF5_6BSP, ORF6_6BSP, ORF7_6BSP, and ORF8_6BSP.

One or more of the *S. pneumoniae* AI from 6B Spain 2 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* AI from 6B Spain 2 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

S. pneumoniae AI from 9V Spain 3 comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("*S. pneumoniae* AI proteins"). Specifically, *S. pneumoniae* AI from 9V Spain 3 includes polynucleotide sequences encoding for two or more of ORF2_9VSP, ORF3_9VSP, ORF4_9VSP, ORF5_9VSP, ORF6_9VSP, ORF7_9VSP, and ORF8_9VSP.

One or more of the *S. pneumoniae* AI from 9V Spain 3 polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* AI from 9V Spain 3 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the *S. pneumoniae* AI surface proteins typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. These sortase proteins are thought to be involved in the secretion and anchoring of the LPXTG containing surface proteins. *S. pneumoniae* AI may encode for at least one surface protein. The Adhesin Island, may encode at least one surface protein. Alternatively, *S. pneumoniae* AI may encode for at least two surface proteins and at least one sortase. Preferably, *S. pneumoniae* AI encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif.

The *S. pneumoniae* AI protein of the composition may be selected from the group consisting of SP0462, SP0463, SP0464, SP0465, SP0466, SP0467, SP0468, orf1_670, orf3_670, orf4_670, orf5_670, orf6_670, orf7_670, orf8_670, ORF2_14CSR, ORF3_14CSR, ORF4_14CSR, ORF5_14CSR, ORF6_14CSR, ORF7_14CSR, ORF8_14CSR, ORF2_19AH, ORF3_19AH, ORF4_19AH, ORF5_19AH, ORF6_19AH, ORF7_19AH, ORF8_19AH, ORF2_19FTW, ORF3_19FTW, ORF4_19FTW, ORF5_19FTW, ORF6_19FTW, ORF7_19FTW, ORF8_19FTW, ORF2_23FP, ORF3_23FP, ORF4_23FP, ORF5_23FP, ORF6_23FP, ORF7_23FP, ORF8_23FP, ORF2_23FTW, ORF3_23FTW, ORF4_23FTW, ORF5_23FTW, ORF6_23FTW, ORF7_23FTW, ORF8_23FTW, ORF2_6BF, ORF3_6BF, ORF4_6BF, ORF5_6BF, ORF6_6BF, ORF7_6BF, ORF8_6BF, ORF2_6BSP, ORF3_6BSP, ORF4_6BSP, ORF5_6BSP, ORF6_6BSP, ORF7_6BSP, ORF8_6BSP, ORF2_9VSP, ORF3_9VSP, ORF4_9VSP, ORF5_9VSP, ORF6_9VSP, ORF7_9VSP and, ORF8_9VSP.

S. pneumoniae AI surface proteins are preferred proteins for use in the immunogenic compositions of the invention. In one embodiment, the compositions of the invention comprise combinations of two or more S pneumoniae AI surface proteins. Preferably such combinations are

selected from two or more of the group consisting of SP0462, SP0463, SP0464, orf3_670, orf4_670, orf5_670, ORF3_14CSR, ORF4_14CSR, ORF5_14CSR, ORF3_19AH, ORF4_19AH, ORF5_19AH, ORF3_19FTW, ORF4_19FTW, ORF5_19FTW, ORF3_23FP, ORF4_23FP, ORF5_23FP, ORF3_23FTW, ORF4_23FTW, ORF5_23FTW, ORF3_6BF, ORF4_6BF, ORF5_6BF, ORF3_6BSP, ORF4_6BSP, ORF5_6BSP, ORF3_9VSP, ORF4_9VSP, and ORF5_9VSP.

In addition to the open reading frames encoding the *S. pneumoniae* AI proteins, *S. pneumoniae* AI may also include a transcriptional regulator.

The *S. pneumoniae* AI proteins of the invention may be used in immunogenic compositions for prophylactic or therapeutic immunization against *S. pneumoniae* infection. For example, the invention may include an immunogenic composition comprising one or more *S. pneumoniae* from TIGR4 AI proteins and one or more *S. pneumoniae* strain 670 proteins. The immunogenic composition may comprise one or more AI proteins from any one or more of *S. pneumoniae* strains TIGR4, 19A Hungary 6, 6B Finland 12, 6B Spain 2, 9V Spain 3, 14 CSR 10, 19F Taiwan 14, 23F Taiwan 15, 23F Poland 16, and 670.

The immunogenic compositions may also be selected to provide protection against an increased range of *S. pneumoniae* serotypes and strain isolates. For example, the immunogenic composition may comprise a first and second *S. pneumoniae* AI protein, wherein a full length polynucleotide sequence encoding for the first *S. pneumoniae* AI protein is not present in a genome comprising a full length polynucleotide sequence encoding for the second *S. pneumoniae* AI protein. In addition, each antigen selected for use in the immunogenic compositions will preferably be present in the genomes of multiple *S. pneumoniae* serotypes and strain isolates. Preferably, each antigen is present in the genomes of at least two (*i.e.*, 3, 4, 5, 6, 7, 8, 9, 10, or more) *S. pneumoniae* strain isolates. More preferably, each antigen is present in the genomes of at least two (*i.e.*, at least 3, 4, 5, or more) *S. pneumoniae* serotypes.

The immunogenic compositions may also be selected to provide protection against an increased range of serotypes and strain isolates of a Gram positive bacteria. For example, the immunogenic composition may comprise a first and second Gram positive bacteria AI protein, wherein a full length polynucleotide sequence encoding for the first Gram positive bacteria AI protein is not present in a genome comprising a full length polynucleotide sequence encoding for the second Gram positive bacteria AI protein. In addition, each antigen selected for use in the immunogenic compositions will preferably be present in the genomes of multiple serotypes and strain isolates of the Gram positive bacteria. Preferably, each antigen is present in the genomes of at least two (*i.e.*, 3, 4, 5, 6, 7, 8, 9, 10, or more) Gram positive bacteria strain isolates. More preferably, each antigen is present in the genomes of at least two (*i.e.*, at least 3, 4, 5, or more) Gram positive bacteria serotypes.One or both of the first and second AI proteins may preferably be in oligomeric or hyperoligomeric form.

Adhesin island surface proteins from two or more Gram positive bacterial genus or species may be combined to provide an immunogenic composition for prophylactic or therapeutic treatment

of disease or infection of two more Gram positive bacterial genus or species. Optionally, the adhesin island surface proteins may be associated together in an oligomeric or hyperoligomeric structure.

In one embodiment, the invention comprises adhesin island surface proteins from two or more *Streptococcus* species. For example, the invention includes a composition comprising a GBS AI surface protein and a GAS adhesin island surface protein. As another example, the invention includes a composition comprising a GAS adhesin island surface protein and a *S. pneumoniae* adhesin island surface protein. One or both of the GAS AI surface protein and the *S. pneumoniae* AI surface protein may be in oligomeric or hyperoligomeric form. As a further example, the invention includes a composition comprising a GBS adhesin island surface protein and a *S. pneumoniae* adhesin island surface protein.

In one embodiment, the invention comprises an adhesin island surface protein from two or more Gram positive bacterial genus. For example, the invention includes a composition comprising a *Streptococcus* adhesin island protein and a *Corynebacterium* adhesin island protein. One or more of the Gram positive bacteria AI surface proteins may be in an oligomeric or hyperoligomeric form.

In addition, the AI polynucleotides and amino acid sequences of the invention may also be used in diagnostics to identify the presence or absence of GBS (or a Gram positive bacteria) in a biological sample. They may be used to generate antibodies which can be used to identify the presence of absence of an AI protein in a biological sample or in a prophylactic or therapeutic treatment for GBS (or a Gram positive bacterial) infection. Further, the AI polynucleotides and amino acid sequences of the invention may also be used to identify small molecule compounds which inhibit or decrease the virulence associated activity of the AI.

BRIEF DESCRIPTION OF THE FIGURES

FIGURE 1 presents a schematic depiction of Adhesin Island 1 ("AI-1") comprising open reading frames for GBS 80, GBS 52, SAG0647, SAG0648 and GBS 104.

FIGURE 2 illustrates the identification of AI-1 sequences in several GBS serotypes and strain isolates (GBS serotype V, strain isolate 2603; GBS serotype III, strain isolate nem316; GBS serotype II, strain isolate 18RS21; GBS serotype V, strain isolate CJB111; GBS serotype III, strain isolate COH1 and GBS serotype 1a, strain isolate A909). (An AI-1 was not identified in GBS serotype 1b, strain isolate H36B or GBS serotype 1a, strain isolate 515).

FIGURE 3 presents a schematic depiction of the correlation between AI-1 and the Adhesin Island 2 ("AI-2") within the GBS serotype V, strain isolate 2603 genome. (This AI-2 comprises open reading frames for GBS 67, GBS 59, SAG1406, SAG1405 and GBS 150).

FIGURE 4 illustrates the identification of AI-2 comprising open reading frames encoding for GBS 67, GBS 59, SAG1406, SAG1404 and GBS 150 (or sequences having sequence homology thereto) in several GBS serotypes and strain isolates (GBS serotype V, strain isolate 2603; GBS serotype III, strain isolate NEM316; GBS serotype 1b, strain isolate H36B; GBS serotype V, strain isolate CJB111; GBS serotype II, strain isolate 18RS21; and GBS serotype 1a, strain isolate 515). Figure 4 also illustrates the identification of AI-2 comprising open reading frames encoding for 01520

(a sortase), 01521, 01522 (a sortase), 01523 (spb1), 01524 and 01525 (or sequences having sequence homology thereto).

FIGURE 5 presents data showing that GBS 80 binds to fibronectin and fibrinogen in ELISA.

FIGURE 6 illustrates that all genes in AI-1 are co-transcribed as an operon.

FIGURE 7 presents schematic depictions of in-frame deletion mutations within AI-1.

FIGURE 8 presents FACS data showing that GBS 80 is required for surface localization of GBS 104.

FIGURE 9 presents FACS data showing that sortases SAG0647 and SAG0648 play a semi-redundant role in surface exposure of GBS 80 and GBS 104.

FIGURE 10 presents Western Blots of the in-frame deletion mutants probed with anti-GBS80 and anti-GBS 104 antisera.

FIGURE 11: Electron micrograph of surface exposed pili structures in *Streptococcus agalactiae* containing GBS 80.

FIGURE 12: PHD predicted secondary structure of GBS 067.

FIGURE 13, 14 and 15: Electron micrograph of surface exposed pili structures of strain isolate COH1 of *Streptococcus agalactiae* containing a plasmid insert encoding GBS 80.

FIGURE 16 and 17: Electron micrograph of surface exposed pili structure of wild type strain isolate COH1 of *Streptococcus agalactiae.*

FIGURE 18: Alignment of polynucleotide sequences of AI-1 from serotype V, strain isolates 2603 and CJB111; serotype II, strain isolate 18RS21; serotype III, strain isolates COH1 and NEM316; and serotype 1a, strain isolate A909.

FIGURE 19: Alignment of polynucleotide sequences of AI-2 from serotype V, strain isolates 2603 and CJB111; serotype II, strain isolate 18RS21; serotype 1b, strain isolate H36B; and serotype 1a, strain isolate 515.

FIGURE 20: Alignment of polynucleotide sequences of AI-2 from serotype V, strain isolate 2603 and serotype III, strain isolate NEM316.

FIGURE 21: Alignment of polynucleotide sequences of AI-2 from serotype III, strain isolate COH1 and serotype Ia, strain isolate A909.

FIGURE 22: Alignment of amino acid sequences of AI-1 surface protein GBS 80 from serotype V, strain isolates 2603 and CJB111; serotype 1a, strain isolate A909; serotype III, strain isolates COH1 and NEM316.

FIGURE 23: Alignment of amino acid sequences of AI-1 surface protein GBS 104 from serotype V, strain isolates 2603 and CJB111; serotype III, strain isolates COH1 and NEM316; and serotype II, strain isolate 18RS21.

FIGURE 24: Alignment of amino acid sequences of AI-2 surface protein GBS 067 from serotype V, strain isolates 2603 and CJB111; serotype 1a, strain isolate 515; serotype II, strain isolate 18RS21; serotype Ib, strain isolate H36B; and serotype III, strain isolate NEM316.

FIGURE 25: Illustrates that GBS closely associates with tight junctions and cross the monolayer of ME180 cervical epithelial cells by a paracellular route.

FIGURE 26: Illustrates GBS infection of ME180 cells.

FIGURE 27: Illustrates that GBS 80 recombinant protein does not bind to epithelial cells.

FIGURE 28: Illustrates that deletion of GBS 80 does not effect the capacity of GBS strain 2603 V/R to adhere and invade ME180 cervical epithelial cells.

FIGURE 29: Illustrates binding of recombinant GBS 104 protein to epithelial cells.

FIGURE 30: Illustrates that deletion of GBS 104 in the GBS strain COH1, reduces the capacity of GBS to adhere to ME180 cervical epithelial cells.

FIGURE 31: Illustrates that GBS 80 knockout mutant strain partially loses the ability to translocate through an epithelial cell monolayer.

FIGURE 32: Illustrates that deletion of GBS 104, but not GBS 80, reduces the capacity of GBS to invade J774 macrophage-like cell line.

FIGURE 33: Illustrates that GBS 104 knockout mutant strain translocates through an epithelial monolayer less efficiently than the isogenic wild type.

FIGURE 34: Negative stained electron micrographs of GBS serotype III, strain isolate COH1, containing a plasmid insert to over-express GBS 80.

FIGURE 35: Electron micrographs of surface exposed pili structures on GBS serotype III, strain isolate COH1, containing a plasmid insert to over-express GBS 80, stained with anti-GBS 80 antibodies (visualized with 10 nm gold particles).

FIGURE 36: Electron micrographs of surface exposed pili structures on GBS serotype III, strain isolate COH1, containing a plasmid insert to over-express GBS 80, stained with anti-GBS 80 antibodies (visualized with 10 nm gold particles).

FIGURE 37: Electron micrographs of surface exposed pili structures on GBS serotype III, strain isolate COH1, containing a plasmid insert to over-express GBS 80, stained with anti-GBS 80 antibodies (visualized with 20 n m gold particles).

FIGURE 38: Electron micrographs of surface exposed pili structures on GBS serotype III, strain isolate COH1, containing a plasmid insert to over-express GBS 80, stained with anti-GBS 104 antibodies or preimmune sera (visualized with 10 nm gold particles).

FIGURE 39: Electron micrographs of surface exposed pili structures on GBS serotype III, strain isolate COH1, containing a plasmid insert to over-express GBS 80, stained with anti-GBS 80 antibodies (visualized with 20 nm gold particles) and anti-GBS 104 antibodies (visualized with 10 nm gold particles).

FIGURE 40: Electron micrographs of surface exposed pili structures on GBS serotype III, strain isolate COH1, containing a plasmid insert to over-express GBS 80, stained with anti-GBS 80 antibodies (visualized with 20 nm gold particles) and anti-GBS 104 antibodies (visualized with 10 nm gold particles).

FIGURE 41: Illustrates that GBS 80 is necessary for polymer formation and GBS104 and sortase SAG0648 are necessary for efficient assembly of pili.

FIGURE 42: Illustrates that GBS 67 is part of a second pilus and that GBS 80 is polymerized in strain 515.

FIGURE 43: Illustrates that two macro-molecules are visible in Coh1, one of which is the GBS 80 pilin.

FIGURE 44: Illustrates pilin assembly.

FIGURE 45: Illustrates that GBS 52 is a minor component of the GBS pilus.

FIGURE 46: Illustrates that the pilus is found in the supernatant of a bacterial culture.

FIGURE 47: Illustrates that the pilus is found in the supernatant of bacterial cultures in all phases.

FIGURE 48: Illustrates that in Coh1, only the GBS 80 protein and one sortase (sag0647 or sag0648) is required for polymerization.

FIGURE 49: IEM image of GBS 80 staining of a GBS serotype VIII strain JM9030013 that express pili.

FIGURE 50: IEM image of GBS 104 staining of a GBS serotype VIII strain JM9030013 that express pili.

FIGURE 51A: Schematic depiction of open reading frames comprising a GAS AI-2 serotype M1 isolate, GAS AI-3 serotype M3, M5, M18, and M49 isolates, a GAS AI-4 serotype M12 isolate, and an GAS AI-1 serotype M6 isolate.

FIGURE 51B: Amino acid alignment of SrtC1-type sortase of a GAS AI-2 serotype M1 isolate, SrtC2-type sortases of serotype M3, M5, M18, and M49 isolates, and a SrtC2-type sortase of a GAS AI-4 serotype M12 isolate.

FIGURE 52: Amino acid alignment of the capsular polysaccharide adhesion proteins of GAS AI-4 serotype M12 (A735), GAS AI-3 serotype M5 (Manfredo), *S. pyogenes* strain MGAS315 serotype M3, *S. pyogenes* strain SSI-1 serotype M3, *S. pyogenes* strain MGAS8232 serotype M3, and GAS AI-2 serotype M1.

FIGURE 53: Amino acid alignment of F-like fibronectin-binding proteins of GAS AI-4 serotype M12 (A735) and *S. pyogenes* strain MGAS10394 serotype M6.

FIGURE 54: Amino acid alignment of F2-like fibronectin-binding proteins of GAS AI-4 serotype M12 (A735), *S. pyogenes* strain MGAS8232 serotype M3, GAS AI-3 strain M5 (Manfredo), *S. pyogenes* strain SSI serotype M3, and *S. pyogenes* strain MGAS315 serotype M3.

FIGURE 55: Amino acid alignment of fimbrial proteins of GAS AI-4 serotype M12 (A735), GAS AI-3 serotype M5 (Manfredo), *S. pyogenes* strain MGAS315 serotype M3, *S. pyogenes* strain SSI serotype M3, *S. pyogenes* strain MGAS8232 serotype M3, and *S. pyogenes* M1 GAS serotype M1.

FIGURE 56: Amino acid alignment of hypothetical proteins of GAS AI-4 serotype M12 (A735), *S.* pyogenes strain MGAS315 serotype M3, *S. pyogenes* strain SSI-1 serotype M3, GAS AI-3 serotype M5 (Manfredo), and *S. pyogenes* strain MGAS8232 serotype M3.

FIGURE 57: Results of FASTA homology search for amino acid sequences that align with the collagen adhesion protein of GAS AI-1 serotype M6 (MGAS10394).

FIGURE 58: Results of FASTA homology search for amino acid sequences that align with the fimbrial structural subunit of GAS AI-1 serotype M6 (MGAS10394).

FIGURE 59: Results of FASTA homology search for amino acid sequences that align with the hypothetical protein of GAS AI-2 serotype M1 (SF370).

FIGURE 60: Specifies pilin and E box motifs present in GAS type 3 and 4 adhesin islands.

FIGURE 61: Illustrates that surface expression of GBS 80 protein on GBS strains COH and JM9130013 correlates with formation of pili structures. Surface expression of GBS 80 was determined by FACS analysis using an antibody that cross-hybridizes with GBS 80. Formation of pili structures was determined by immunogold electron microscopy using gold-labelled anti-GBS 80 antibody.

FIGURE 62: Illustrates that surface exposure is capsule-dependent for GBS 322 but not for GBS 80.

FIGURE 63: Illustrates the amino acid sequence identity of GBS 59 proteins in GBS strains.

FIGURE 64: Western blotting of whole GBS cell extracts with anti-GBS 59 antibodies.

FIGURE 65: Western blotting of purified GBS 59 and whole GBS cell extracts with anti-GBS 59 antibodies.

FIGURE 66: FACS analysis of GBS strains CJB111, 7357B, 515 using GBS 59 antiserum.

FIGURE 67: Illustrates that anti-GBS 59 antibodies are opsonic for CJB111 GBS strain serotype V.

FIGURE 68: Western blotting of GBS strain JM9130013 total extracts.

FIGURE 69: Western blotting of GBS stain 515 total extracts shows that GBS 67 and GBS 150 are parts of a pilus.

FIGURE 70: Western blotting of GBS strain 515 knocked out for GBS 67 expression

FIGURE 71: FACS analysis of GBS strain 515 and GBS strain 515 knocked out for GBS 67 expression using GBS 67 and GBS 59 antiserum.

FIGURE 72: Illustrates complementation of GBS 515 knocked out for GBS 67 expression with a construct overexpressing GBS 80.

FIGURE 73: FACS analysis of GAS serotype M6 for spyM6_0159 surface expression.

FIGURE 74: FACS analysis of GAS serotype M6 for spyM6_0160 surface expression.

FIGURE 75: FACS analysis of GAS serotype M1 for GAS 15 surface expression.

FIGURE 76: FACS analysis of GAS serotype M1 for GAS 16 surface expression using a first anti-GAS 16 antiserum.

FIGURE 77: FACS analysis of GAS serotype M1 for GAS 18 surface expression using a first anti-GAS 18 antiserum.

FIGURE 78: FACS analysis of GAS serotype M1 for GAS 18 surface expression using a second anti-GAS 18 antiserum.

FIGURE 79: FACS analysis of GAS serotype M1 for GAS 16 surface expression using a second anti-GAS 16 antisera.

FIGURE 80: FACS analysis of GAS serotype M3 for spyM3_0098 surface expression.

FIGURE 81: FACS analysis of GAS serotype M3 for spyM3_0100 surface expression.

FIGURE 82: FACS analysis of GAS serotype M3 for spyM3_0102 surface expression.

FIGURE 83: FACS analysis of GAS serotype M3 for spyM3_0104 surface expression.

FIGURE 84: FACS analysis of GAS serotype M3 for spyM3_0106 surface expression.

FIGURE 85: FACS analysis of GAS serotype M12 for 19224134 surface expression.

FIGURE 86: FACS analysis of GAS serotype M12 for 19224135 surface expression.

FIGURE 87: FACS analysis of GAS serotype M12 for 19224137 surface expression.

FIGURE 88: FACS analysis of GAS serotype M12 for 19224141 surface expression.

FIGURE 89: Western blot analysis of GAS 15 expression on GAS M1 bacteria.

FIGURE 90: Western blot analysis of GAS 15 expression using GAS 15 immune sera.

FIGURE 91: Western blot analysis of GAS 15 expression using GAS 15 pre-immune sera.

FIGURE 92: Western blot analysis of GAS 16 expression on GAS M1 bacteria.

FIGURE 93: Western blot analysis of GAS 16 expression using GAS 16 immune sera.

FIGURE 94: Western blot analysis of GAS 16 expression using GAS 16 pre-immune sera.

FIGURE 95: Western blot analysis of GAS 18 on GAS M1 bacteria.

FIGURE 96: Western blot analysis of GAS 18 using GAS 18 immune sera.

FIGURE 97: Western blot analysis of GAS 18 using GAS 18 pre-immune sera.

FIGURE 98: Western blot analysis of M6_Spy0159 expression on GAS bacteria.

FIGURE 99: Western blot analysis of 19224135 expression on M12 GAS bacteria.

FIGURE 100: Western blot analysis of 19224137 expression on M12 GAS bacteria.

FIGURE 101: Full length nucleotide sequence of an *S. pneumoniae* strain 670 AI.

FIGURE 102: Western blot analysis of GAS 15, GAS 16, and GAS 18 in GAS M1 strain 2580.

FIGURE 103: Western blot analysis of GAS 15, GAS 16, and GAS 18 in GAS M1 strain 2913.

FIGURE 104: Western blot analysis of GAS 15, GAS 16, and GAS 18 in GAS M1 strain 3280.

FIGURE 105: Western blot analysis of GAS 15, GAS 16, and GAS 18 in GAS M1 strain 3348.

FIGURE 106: Western blot analysis of GAS 15, GAS 16, and GAS 18 in GAS M1 strain 2719.

FIGURE 107: Western blot analysis of GAS 15, GAS 16, and GAS 18 in GAS M1 strain SF370.

FIGURE 108: Western blot analysis of 19224135 and 19224137 in GAS M12 strain 2728.

FIGURE 109: Western blot analysis of 19224139 in GAS M12 strain 2728 using antisera raised against SpyM3_0102.

FIGURE 110: Western blot analysis of M6_Spy0159 and M6_Spy0160 in GAS M6 strain 2724.

FIGURE 111: Western blot analysis of M6_Spy0159 and M6_Spy0160 in GAS M6 strain SF370.

FIGURE 112: Western blot analysis of M6_Spy160 in GAS M6 strain 2724.

FIGURES 113-115: Electron micrographs of surface exposed GAS 15 on GAS M1 strain SF370.

FIGURES 116-121: Electron micrographs of surface exposed GAS 16 on GAS M1 strain SF370.

FIGURES 122-125: Electron micrographs of surface exposed GAS 18 on GAS M1 strain SF370 detected using anti-GAS 18 antisera.

FIGURE 126: IEM image of a hyperoligomer on GAS M1 strain SF370 detected using anti-GAS 18 antisera.

FIGURES 127-132: IEM images of oligomeric and hyperoligomeric structures containing M6_Spy0160 extending from the surface of GAS serotype M6 3650.

FIGURE 133A and B: Western blot analysis of *L. lactis* transformed to express GBS 80 with anti-GBS 80 antiserum.

FIGURES 134: Western blot analyses of *L. lactis* transformed to express GBS AI-1 with anti-GBS 80 antiserum.

FIGURE 135: Ponceau staining of same acrylamide gel as used in Figure 134.

FIGURE 136A: Western blot analysis of sonicated pellets and supernatants of cultured *L. lactis* transformed to express GBS AI-1 polypeptides using anti-GBS 80 antiserum.

FIGURE 136B: Polyacrylamide gel electrophoresis of sonicated pellets and supernatants of cultured *L. lactis* transformed to express GBS AI polypeptides.

FIGURE 137: Depiction of an example *S. pneumoniae* AI locus.

FIGURE 138: Schematic of primer hybridization sites within the *S. pneumoniae* AI locus of FIGURE 137.

FIGURE 139A: The set of amplicons produced from the *S. pneumoniae* strain TIGR4 AI locus.

FIGURE 139B: Base pair lengths of amplicons produced from FIGURE 139A primers in *S. pneumoniae* strain TIGR4.

FIGURE 140: CGH analysis of *S. pneumoniae* strains for the AI locus.

FIGURE 141: Amino acid sequence alignment of polypeptides encoded by AI orf 2 in *S. pneumoniae* AI-positive strains.

FIGURE 142: Amino acid sequence alignment of polypeptides encoded by AI orf 3 in *S. pneumoniae* AI-positive strains.

FIGURE 143: Amino acid sequence alignment of polypeptides encoded by AI orf 4 in *S. pneumoniae* AI-positive strains.

FIGURE 144: Amino acid sequence alignment of polypeptides encoded by AI orf 5 in *S. pneumoniae* AI-positive strains.

FIGURE 145: Amino acid sequence alignment of polypeptides encoded by AI orf 6 in *S. pneumoniae* AI-positive strains.

FIGURE 146: Amino acid sequence alignment of polypeptides encoded by AI orf 7 in *S. pneumoniae* AI-positive strains.

FIGURE 147: Amino acid sequence alignment of polypeptides encoded by AI orf 8 in *S. pneumoniae* AI-positive strains.

FIGURE 148: Diagram comparing amino acid sequences of RrgA in *S. pneumoniae* strains.

FIGURE 149: Amino acid sequence comparison of RrgB *S. pneumoniae* strains.

FIGURE 150A: Sp0462 amino acid sequence.

FIGURE 150B: Primers used to produce a clone encoding the Sp0462 polypeptide.

FIGURE 151A: Schematic depiction of recombinant Sp0462 polypeptide.

FIGURE 151B: Schematic depiction of full-length Sp0462 polypeptide.

FIGURE 152A: Western blot probed with serum obtained from *S. pneumoniae*–infected patients for Sp0462.

FIGURE 152B: Western blot probed with GBS 80 serum for Sp0462.

FIGURE 153A: Sp0463 amino acid sequence.

FIGURE 153B: Primers used to produce a clone encoding the Sp0463 polypeptide.

FIGURE 154A: Schematic depiction of recombinant Sp0463 polypeptide.

FIGURE 154B: Schematic depiction of full-length Sp0463 polypeptide.

FIGURE 155: Western blot detection of recombinant Sp0463 polypeptide.

FIGURE 156: Western blot detection of high molecular weight Sp0463 polymers.

FIGURE 157A: Sp0464 amino acid sequence.

FIGURE 157B: Primers used to produce a clone encoding the Sp0464 polypeptide.

FIGURE 158A: Schematic depiction of recombinant Sp0464 polypeptide.

FIGURE 158B: Schematic depiction of full-length Sp0464 polypeptide.

FIGURE 159: Western blot detection of recombinant Sp0464 polypeptide.

FIGURE 160: Amplification products prepared for production of Sp0462, Sp0463, and Sp0464 clones.

FIGURE 161: Opsonic killing by anti-sera raised against *L. lactis* expressing GBS AI

FIGURE 162: Schematic depicting GAS adhesin islands GAS AI-1, GAS AI-2, GAS AI-3 and GAS AI-4.

FIGURES 163 A-D: Immunoblots of cell-wall fractions of GAS strains with antisera specific for LPXTG proteins of M6_ ISS3650 (A), M1_SF370 (B), M5_ISS4883 (C) and M12_20010296 (D).

FIGURES 163 E-H: Immunoblots of cell-wall fractions of deletion mutants M1_SF370Δ128 (E) M1_SF370Δ130 (F) M1_SF370ΔSrtC1 (G) and the M1_128 deletion strain complemented with plasmid pAM::128 which contains the M1_128 gene (H) with antisera specific for the pilin components of M1_SF370.

FIGURES 163 I-N: Immunogold labelling and transmission electron microscopy of: T6 (I) and Cpa (J) in M6_ISS3650; M1_128 in M1_SF370 (K) and deletion strain M1_SF370Δ128 (N); M5_orf80 in M5_ISS4883 (L); M12_EftLSL.A in M12_20010296 (M). The strains used are indicated below the panels. Bars=200nm.

FIGURE 164: Schematic representation of the FCT region from 7 GAS strains

FIGURES 165 A-H: Flow cytometry of GAS bacteria treated or not with trypsin and stained with sera specific for the major pilus component. Preimmune staining; black lines, untreated bacteria; green lines and trypsin treated bacteria; blue lines. M6_ISS3650 stained with sera which recognize the M6 protein (A) or anti-M6_T6 (B), M1_SF370 stained with anti-M1 (C) or anti-M1_128 (D), M5_ISS4883 stained with anti-PrtF (E) or anti-M5_orf80 (F) and M12_20010296 with anti-M12 (G) or anti-EftLSL.A (H)

FIGURES 166 A-C: Immunoblots of recombinant pilin components with polyvalent Lancefield T-typing sera. The recombinant proteins are shown above the blot and the sera pool used is shown below the blot.

FIGURES 166 D-G: Immunoblots of pilin proteins with monovalent T-typing sera. The recombinant proteins are shown below the blot and the sera used above the blot.

Figure 166 H and I Flow cytometry analysis of strain M1_SF370 (H) and the deletion strain M1_SF370Δ128 (I) with T-typing antisera pool T.

FIGURE 167: Chart describing the number and type of sortase sequences identified within GAS AIs.

FIGURE 168 A: Immunogold-electronmicroscopy of *L. lactis* lacking an expression construct for GBS AI-1 using anti-GBS 80 antibodies.

FIGURE 168 B and C: Immunogold-electronmicroscopy detects GBS 80 in oligomeric (pilus) structures on surface of *L. lactis* transformed to express GBS AI-1

FIGURE 169: FACS analysis detects expression of GBS 80 and GBS 104 on the surface of *L. lactis* transformed to express GBS AI-1.

FIGURE 170: Phase contrast microscopy and immuno-electronmicroscopy shows that expression of GBS AI-1 in *L. lactis* induces *L. lactis* aggregation.

FIGURE 171: Purification of GBS pili from *L. lactis* transformed to express GBS AI-1.

PC FIGURE 172: Schematic depiction of GAS M6 (AI-1), M1 (AI-2), and M12 (AI-4) adhesin islands and portions of the adhesin islands inserted in the pAM401 construct for expression in *L. lactis*.

FIGURE 173 A-C: Western blot analysis showing assembly of GAS pili in *L. lactis* expressing GAS AI-2 (M1) (A), GAS AI-4 (M12) (B), and GAS AI-1 (M6) (C).

FIGURE 174: FACS analysis of GAS serotype M6 for M6_Spy0157 surface expression.

FIGURE 175: FACS analysis of GAS serotype M12 for 19224139 surface expression.

FIGURE 176 A-E: Immunogold electron microscopy using antibodies against M6_Spy0160 detects pili on the surface of M6 strain 2724.

FIGURE 176 F: Immunogold electron microscopy using antibodies against M6_Spy0159 detects M6_Spy0159 surface expression on M6 strain 2724.

FIGURE 177 A-C: Western blot analysis of M1 strain SF370 GAS bacteria individually deleted for M1_130, SrtC1, or M1_128 using anti-M1_130 serum (A), anti-M1_128 serum (B), and anti-M1_126 serum (C).

FIGURE 178 A-C: Immunogold electron microscopy using antibodies against M1_128 to detect surface expression on wildtype strain SF370 bacteria (A), M1_128 deleted SF370 bacteria (B), and SrtC1 deleted SF370 bacteria (C).

FIGURE 179 A-C: FACS analysis to detect expression of M1_126 (A), M1_128 (B), and M1_130 (C) on the surface of wildtype SF370 GAS bacteria.

FIGURE 179 D-F: FACS analysis to detect expression of M1_126 (D), M1_128 (E), and M1_130 (F) on the surface of M1_128 deleted SF370 GAS bacteria.

FIGURE 179 G-I: FACS analysis to detect expression of M1_126 (G), M1_128 (H), and M1_130 (I) on the surface of SrtC1 deleted SF370 GAS bacteria.

FIGURE 180 A and B: FACS analysis of wildtype (A) and LepA deletion mutant (B) strains of SF370 bacteria for M1 surface expression.

FIGURE 181: Western blot analysis detects high molecular weight polymers in *S. pneumoniae* TIGR4 using anti-RrgB antisera.

FIGURE 182: Detection of high molecular weight polymers in *S. pnuemoniae* rlrA positive strains.

FIGURE 183: Detection of high molecular weight polymers in *S. pneumoniae* TIGR4 by silver staining and Western blot analysis using anti-RrgB antisera.

FIGURE 184: Deletion of *S. pneumoniae* TIGR4 adhesin island sequences interferes with the ability of *S. pneumoniae* to adhere to A549 alveolar cells.

FIGURE 185: Negative staining of *S. pneumoniae* strain TIGR4 showing abundant pili on the bacterial surface.

FIGURE 186: Negative staining of strain TIGR4 deleted for rrgA-srtD adhesin island sequences showing no pili on the bacterial surface

FIGURE 187: Negative staining of the TIGR4 mgrA mutant showing abundant pili on the bacterial surface.

FIGURE 188: Negative staining of the negative control TIGR4 mgrA mutant deleted for adhesin island sequences rrgA-srtD showing no pili on the bacterial surface.

FIGURE 189: Immuno-gold labelling of *S. pneumoniae* strain TIGR4 grown on blood agar solid medium using α-RrgB (5nm) and α-RrgC (10nm). Bar represents 200nm.

FIGURE 190 A and B: Detection of expression and purification of *S. pneumoniae* RrgA protein by SDS-PAGE (A) and Western blot analysis (B).

FIGURE 191: Detection of RrgB by antibodies produced in mice.

FIGURE 192: Detection of RrgC by antibodies produced in mice.

FIGURE 193: Purification of *S. pneumoniae* TIGR 4 pili by a cultivation and digestion method and detection of the purified TIGR4 pili.

FIGURE 194: Purification of *S. pneumoniae* TIGR 4 pili by a sucrose gradient centrifugation method and detection of the purified TIGR4 pili.

FIGURE 195: Purification of *S. pneumoniae* TIGR 4 pili by a gel filtration method and detection of the purified TIGR4 pili.

FIGURE 196: Alignment of full length *S. pneumoniae* adhesin island sequences from ten *S. pneumoniae* strains.

FIGURE 197 A: Schematic of GBS AI-1 coding sequences.

FIGURE 197 B: Nucleotide sequence of intergenic region between AraC and GBS 80 (SEQ ID NO: 273.

FIGURE 197 C: FACS analysis results for GBS 80 expression in GBS strains having different length polyA tracts in the intergenic region between AraC and GBS 80.

FIGURE 198: Table comparing the percent identity of surface proteins encoded by a serotype M6 (harbouring a GAS AI-1) adhesin island relative to other GAS serotypes harbouring an adhesin island.

FIGURE 199: Table comparing the percent identity of surface proteins encoded by a serotype M1 (harbouring a GAS AI-2) adhesin island relative to other GAS serotypes harbouring an adhesin island.

FIGURE 200: Table comparing the percent identity of surface proteins encoded by serotypes M3, M18, M5, and M49 (harbouring GAS AI-3) adhesin islands relative to other GAS serotypes harbouring an adhesin island.

FIGURE 201: Table comparing the percent identity of surface proteins encoded by a serotype M12 (harbouring a GAS AI-1) adhesin island- relative to other GAS serotypes harbouring an adhesin island.

FIGURE 202: GBS 80 recombinant protein does not bind to epithelial cells.

FIGURE 203: Deletion of GBS 80 protein does not affect the ability of GBS to adhere and invade ME180 cervical epithelial cells.

FIGURE 204: GBS 80 binds to extracellular matrix proteins.

FIGURE 205: Deletion of GBS 104 protein, but not GBS 80, reduces the capacity of GBS to invade J774 macrophage-like cells

FIGURE 206: GBS 104 knockout mutant strains of bacteria translocate through an epithelial monolayer less efficiently that the isogenic wild type strain.

FIGURE 207: GBS 80 knockout mutant strains of bacteria partially lose the ability to translocate through an epithelial monolayer.

FIGURE 208: GBS adherence to HUVEC endothelial cells.

FIGURE 209: Strain growth rate of wildtype, GBS 80-deleted, or GBS 104 deleted COH1 GBS.

FIGURE 210: Binding of recombinant GBS 104 protein to epithelial cells by FACS analysis.

FIGURE 211: Deletion of GBS 104 proteinin the GBS strain COH1 reduces the ability of GBS to adhere to ME180 cervical epithelial cells.

FIGURE 212: COH1 strain GBS overexpressing GBS 80 protein has an impaired capacity to translocate through an epithelial monolayer.

FIGURE 213: Scanning electron microscopy shows that overexpression of GBS 80 protein on COH1 strain GBS enhances the capacity of the COH1 bacteria to form microcolonies on epithelial cells.

FIGURE 214: Confocal imaging shows that overexpression of GBS 80 proteins on COH1 strain GBS enhances the capacity of the COH1 bacteria to form microcolonies on epithelial cells.

FIGURE 215: Detection of GBS 59 on the surface of GBS strain 515 by immuno-electron microscopy.

FIGURE 216: Detection of GBS 67 on the surface of GBS strain 515 by immuno-electron microscopy.

FIGURE 217: GBS 67 binds to fibronectin.

FIGURE 218: Western blot analysis shows that deletion of both GBS AI-2 sortase genes abolishes assembly of the pilus.

FIGURE 219: FACS analysis shows that deletion of both GBS AI-2 sortase genes abolishes assembly of the pilus.

FIGURE 220 A-C: Western blot analysis shows that GBS 59, GBS 67, and GBS 150 form high molecular weight complexes.

FIGURE 221 A-C: Western blot analysis shows that GBS 59 is required for polymer formation of GBS 67 and GBS 150.

FIGURE 222: FACS analysis shows that GBS 59 is required for surface exposure of GBS 67.

FIGURE 223: Summary Western blots for detection of GBS 59, GBS 67, or GBS 150 in GBS 515 and GBS 515 mutant strain.

FIGURE 224: Description of GBS 59 allelic variants.

PC FIGURE 225: GBS 59 is opsonic only against a strain of GBS expressing a homologous GBS 59.

FIGURE 226 A and B: Results of FACS analysis for surface expression of GBS 59 using antibodies specific for different GBS 59isoforms.

FIGURE 227 A and B: Results of FACS analysis for surface expression of GBS 80, GBS 104, GBS 322, GBS 67, and GBS 59 on 41 various strains of GBS bacteria.

FIGURE 228: Results of FACS analysis for surface expression of GBS 80, GBS 104, GBS 322, GBS 67, and GBS 59 on 41 strains of GBS bacteria obtained from the CDC.

FIGURE 229: Expected immunogenicity coverage of different combinations of GBS 80, GBS 104, GBS 322, GBS 67, and GBS 59 across strains of GBS bacteria.

FIGURE 230: GBS 59 opsonophagocytic activity is comparable to that of a mixture of GBS 80, GBS 104, GBS 322 and GBS 67.

FIGURE 231 A-C: Schematic presentation of example hybrid GBS AIs.

FIGURE 232: Schematic presentation of an example hybrid GBS AI.

FIGURE 233 A and B: Western blot and FACS analysis detect expression of GBS 80 and GBS 67 on the surface of L. lactis transformed with a hybrid GBS AI.

FIGURE 234 A-E Hybrid GBS AI cloning strategy.

FIGURE 235: High magnification of *S. pneumoniae* strain TIGR4 pili double labeled with α-RrgB (5nm) and α-RrgC (10nm). Bar represents 100nm.

FIGURE 236: Immuno-gold labeling of the *S. pneumoniae* TIGR4 rrgA-srtD deletion mutant with no visible pili on the surface detectable by α-RrgB- and α-RrgC. Bar represents 200nm.

FIGURE 237: Variability in GBS 67 amino acid sequences between strains 2603 and H36B.

FIGURE 238: Strain variability in GBS 67 amino acid sequences of allele I (2603).

FIGURE 239: Stran variability in GBS 67 amino acid sequence of allele II (H36B).

BRIEF DESCRIPTION OF THE TABLES

TABLE 1: Active Maternal Immunization Assay for fragments of GBS 80

TABLE 2: Passive Maternal Immunization Assay for fragments of GBS 80

TABLE 3: Lethal dose 50% of AI-1 mutants from GBS strain isolate 2603.

TABLE 4: GAS AI-1 sequences from M6 isolate (MGAS10394).

TABLE 5: GAS AI-2 sequences from M1 isolate (SF370).

TABLE 6: GAS AI-3 sequences from M3 isolate (MGAS315).

TABLE 7: GAS AI-3 sequences from M3 isolate (SSI-1).

TABLE 8: GAS AI-3 sequences from M18 isolate (MGAS8232).

TABLE 9: S. pneumoniae AI sequences from TIGR4 sequence.

TABLE 10: GAS AI-3 sequences from M5 isolate (Manfredo).

PC TABLE 11: GAS A1-4 sequences from M12 isolate (A735).

TABLE 12: Conservation of GBS 80 and GBS 104 amino acid sequences.

TABLE 13: Conservation of GBS 322 and GBS 276 amino acid sequences.

TABLE 14: Active maternal immunization assay for a combination of fragments from GBS 322, GBS 80, GBS 104, and GBS 67.

TABLE 15: Antigen surface exposure of GBS 80, GBS 322, GBS 104, and GBS 67.

TABLE 16: Active maternal immunization assay for each of GBS 80 and GBS 322 antigens.

TABLE 17: Active maternal immunization assay for GBS 59.

TABLE 18: Summary of FACS values for surface expression of spyM6_0159.

TABLE 19: Summary of FACS values for surface expression of spyM6_0160.

TABLE 20: Summary of FACS values for surface expression of GAS 15.

TABLE 21: Summary of FACS values for surface expression of GAS 16.

TABLE 22: Summary of FACS values for surface expression of GAS 16 using a second antisera.

TABLE 23: Summary of FACS values for surface expression of GAS 18.

TABLE 24: Summary of FACS values for surface expression of GAS 18 using a second antisera.

TABLE 25: Summary of FACS values for surface expression of SpyM3_0098.

TABLE 26: Summary of FACS values for surface expression of SpyM3_0100.

TABLE 27: Summary of FACS values for surface expression of SpyM3_0102 in M3 serotypes.

TABLE 28: Summary of FACS values for surface expression of SpyM3_0102 in M6 serotypes.

TABLE 29: Summary of FACS values for surface expression of SpyM3_0104 in M3 serotypes.

TABLE 30: Summary of FACS values for surface expression of SpyM3_0104 in an M12 serotype.

TABLE 31: Summary of FACS values for surface expression of SPs_0106 in M3 serotypes.

TABLE 32: Summary of FACS values for surface expression of SPs_0106 in an M12 serotype.

TABLE 33: Summary of FACS values for surface expression of 19224134 in an M12 serotype.

TABLE 34: Summary of FACS values for surface expression of 19224134 in M6 serotypes.

TABLE 35: Summary of FACS values for surface expression of 19224135 in an M12 serotype.

TABLE 36: Summary of FACS values for surface expression of 19224137 in an M12 serotype.

PC TABLE 37: Summary of FACS values for surface expression of 19224141 in an M12 serotype.

TABLE 38: *S. pneumoniae* strain 670 AI sequences.

TABLE 39: Pecent identity comparison of *S. pneumoniae* strains AI sequences.

TABLE 40: FACS analysis of *L. lactis* and GBS bacteria strains expressing GBS AI-1.

TABLE 41: Sequences of primers used to amplify AI locus.

TABLE 42: Conservation of amino acid sequences encoded by the *S. pneumoniae* AI locus.

TABLE 43: Protection of Mice Immunized with *L. lactis* expressing GBS AI-1.

TABLE 44: GAS AI-3 sequences from M49 isolate (591).

TABLE 45: Comparison of Sequences Between the Four GAS AIs.

TABLE 46: Antibody Responses against GBS 80 in Serum of Mice Immunized with *L. lactis* Expressing GBS AI-1

TABLE 47: Anti-GBS 80 IgA Antibodies Detected in Mouse Tissues Following Immunization with *L. lactis* Expressing GBS AI-1

TABLE 48: GBS 67 Protects Mice in an Immunization Assay

TABLE 49: Exposure Levels of GBS 80, GBS 104, GBS 67, GBS 322, and GBS 59 on GBS Strains

TABLE 50: High Levels of Surface Protein Expression on GBS Serotypes

TABLE 51: Further Protection of Mice Immunized with *L. lactis* expressing GBS AI-1

DETAILED DESCRIPTION OF THE INVENTION

The practice of the present invention will employ, unless otherwise indicated, conventional methods of chemistry, biochemistry, molecular biology, immunology and pharmacology, within the skill of the art. Such techniques are explained fully in the literature. See, e.g., *Remington's Pharmaceutical Sciences*, Mack Publishing Company, Easton, Pa., 19th Edition (1995); *Methods In Enzymology* (S. Colowick and N. Kaplan, eds., Academic Press, Inc.); and *Handbook of Experimental Immunology*, Vols. I-IV (D.M. Weir and C.C. Blackwell, eds., 1986, Blackwell Scientific Publications); Sambrook, et al., *Molecular Cloning: A Laboratory Manual* (2nd Edition, 1989); *Handbook of Surface and Colloidal Chemistry* (Birdi, K.S. ed., CRC Press, 1997); *Short Protocols in Molecular Biology*, 4th ed. (Ausubel et al. eds., 1999, John Wiley & Sons); *Molecular Biology Techniques: An Intensive Laboratory Course*, (Ream et al., eds., 1998, Academic Press); *PCR (Introduction to Biotechniques Series)*, 2nd ed. (Newton & Graham eds., 1997, Springer Verlag); Peters and Dalrymple, *Fields Virology* (2d ed), Fields et al. (eds.), B.N. Raven Press, New York, NY.

All publications, patents and patent applications cited herein, are hereby incorporated by reference in their entireties.

As used herein, an "Adhesin Island" or "AI" refers to a series of open reading frames within a bacterial genome, such as the genome for Group A or Group B Streptococcus or other gram positive bacteria, that encodes for a collection of surface proteins and sortases. An Adhesin Island may

encode for amino acid sequences comprising at least one surface protein. The Adhesin Island may encode at least one surface protein. Alternatively, an Adhesin Island may encode for at least two surface proteins and at least one sortase. Preferably, an Adhesin Island encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. One or more AI surface proteins may participate in the formation of a pilus structure on the surface of the gram positive bacteria.

Adhesin Islands of the invention preferably include a divergently transcribed transcriptional regulator (*i.e.*, the transcriptional regulator is located near or adjacent to the AI protein open reading frames, but it transcribed in the opposite direction). The transcriptional regulator may regulate the expression of the AI operon.

GBS Adhesin Island 1

As discussed above, Applicants have identified a new adhesin island, "Adhesin Island 1", "AI-1", or "GBS AI-1", within the genomes of several Group B Streptococcus serotypes and isolates. AI-1 comprises a series of approximately five open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("AI-1 proteins"). Specifically, AI-1 includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4 or 5) of GBS 80, GBS 104, GBS 52, SAG0647 and SAG0648. One or more of the AI-1 open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the AI-1 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

A schematic of AI-1 is presented in Figure 1. AI-1 typically resides on an approximately 16.1 kb transposon-like element frequently inserted into the open reading frame for *trmA*. One or more of the AI-1 surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) motif or other sortase substrate motif. The AI surface proteins of the invention may affect the ability of the GBS bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of GBS to translocate through an epithelial cell layer. Preferably, one or more AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The AI-1 sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. AI-1 may encode for at least one surface protein. Alternatively, AI-1 may encode for at least two surface exposed proteins and at least one sortase. Preferably, AI-1 encodes for at least three surface exposed proteins and at least two sortases. The AI-1 protein preferably includes GBS 80 or a fragment thereof or a sequence having sequence identity thereto.

As used herein, an LPXTG motif represents an amino acid sequence comprising at least five amino acid residues. Preferably, the motif includes a leucine (L) in the first amino acid position, a proline (P) in the second amino acid position, a threonine (T) in the fourth amino acid position and a glycine (G) in the fifth amino acid position. The third position, represented by X, may be occupied by

any amino acid residue. Preferably, the X is occupied by lysine (K), Glutamate (E), Asparagine (N), Glutamine (Q) or Alanine (A). Preferably, the X position is occupied by lysine (K). In some embodiments, one of the assigned LPXTG amino acid positions is replaced with another amino acid. Preferably, such replacements comprise conservative amino acid replacements, meaning that the replaced amino acid residue has similar physiological properties to the removed amino acid residue. Genetically encoded amino acids may be divided into four families based on physiological properties: (1) acidic (asparatate and glutamate), (2) basic (lysine, arginine, histitidine), (3) non-polar (alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophane) and (4) uncharged polar (glycine, asparagines, glutamine, cysteine, serine, threonine, and tyrosine). Phenylalanine, tryptophan and tyrosine are sometimes classified jointly as aromatic amino acids. For example, it is reasonably predictable that an isolated replacement of a leucine with an isoleucine or valine, an asparate with a glutamate, a threonine with a serine, or a similar conservative replacement of an amino acid with a structurally related amino acid will not have a major effect on the biological activity.

The first amino acid position of the LPXTG motif may be replaced with another amino acid residue. Preferably, the first amino acid residue (leucine) is replaced with an alanine (A), valine (V), isoleucine (I), proline (P), phenylalanine (F), methionine (M), glutamic acid (E), glutamine (Q), or tryptophan (Y) residue. In one preferred embodiment, the first amino acid residue is replaced with an isoleucine (I).

The second amino acid residue of the LPXTG motif may be replaced with another amino acid residue. Preferably, the second amino acid residue praline (P) is replaced with a valine (V) residue.

The fourth amino acid residue of the LPXTG motif may be replaced with another amino acid residue. Preferably, the fourth amino acid residue (threonine) is replaced with a serine (S) or an alanine (A).

In general, an LPXTG motif may be represented by the amino acid sequence XXXXG, in which X at amino acid position 1 is an L, a V, an E, an I, an F, or a Q; X at amino acid position 2 is a P if X at amino acid position 1 is an L, an I, or an F; X at amino acid position 2 is a V if X at amino acid position 1 is a E or a Q; X at amino acid position 2 is a V or a P if X at amino acid position 1 is a V; X at amino acid position 3 is any amino acid residue; X at amino acid position 4 is a T if X at amino acid position 1 is a V, E, I, F, or Q; and X at amino acid position 4 is a T, S, or A if X at amino acid position 1 is an L.

Generally, the LPXTG motif of a GBS AI protein may be represented by the amino acid sequence XPXTG, in which X at amino acid position 1 is L, I, or F, and X at amino acid position 3 is any amino acid residue. Specific examples of LPXTG motifs in GBS AI proteins may include LPXTG (SEQ ID NO: 122) or IPXTG (SEQ ID NO: 133).

As discussed further below, the threonine in the fourth amino acid position of the LPXTG motif may be involved in the formation of a bond between the LPXTG containing protein and a cell wall precursor. Accordingly, in preferred LPXTG motifs, the threonine in the fourth amino acid

position is not replaced with another amino acid or, if the threonine is replaced, the replacement amino acid is preferably a conservative amino acid replacement, such as serine.

Instead of an LPXTG motif, the AI surface proteins of the invention may contain alternative sortase substrate motifs such as NPQTN (SEQ ID NO: 142), NPKTN (SEQ ID NO: 168), NPQTG (SEQ ID NO: 169), NPKTG (SEQ ID NO: 170), XPXTGG (SEQ ID NO: 143), LPXTAX (SEQ ID NO: 144), or LAXTGX (SEQ ID NO: 145). (Similar conservative amino acid substitutions can also be made to these membrane motifs).

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

The AI surface proteins may be polymerized into pili by sortase-catalysed transpeptidation. (See Figure 44.) Cleavage of AI surface proteins by sortase between the threonine and glycine residues of an LPXTG motif yields a thioester-linked acyl intermediate of sortase. Many AI surface proteins include a pilin motif amino acid sequence which interacts with the sortase and LPXTG amino acid sequence. The first lysine residue in a pilin motif can serve as an amino group acceptor of the cleaved LPXTG motif and thereby provide a covalent linkage between AI subunits to form pili. For example, the pilin motif can make a nucleophilic attack on the acyl enzyme providing a covalent linkage between AI subunits to form pili and regenerate the sortase enzyme. Examples of pilin motifs may include ((YPKN(X_{10})K; SEQ ID NO: 146), (YPKN(X_9)K; SEQ ID NO: 147), (YPK(X_7)K; SEQ ID NO: 148), (YPK(X_{11})K; SEQ ID NO: 149), or (PKN(X_9)K; SEQ ID NO: 150)). Preferably, the AI surface proteins of the invention include a pilin motif amino acid sequence.

Typically, AI surface proteins of the invention will contain an N-terminal leader or secretion signal to facilitate translocation of the surface protein across the bacterial membrane.

Group B Streptococci are known to colonize the urinary tract, the lower gastrointestinal tract and the upper respiratory tract in humans. Electron micrograph images of GBS infection of a cervical epithelial cell line (ME180) are presented in Figure 25. As shown in these images, the bacteria closely associate with tight junctions between the cells and appear to cross the monolayer by a paracellular route. Similar paracellular invasion of ME180 cells is also shown in the contrast images in Figure 26. The AI surface proteins of the invention may effect the ability of the GBS bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of GBS to translocate through an epithelial cell layer. Preferably, one or more AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface.

Applicants have discovered that AI-1 surface protein GBS 104 can bind epithelial cells such as ME180 human cervical cells, A549 human lung cells and Caco2 human intestinal cells (See

Figures 29 and 210). Further, deletion of the GBS 104 sequence in a GBS strain reduces the capacity of GBS to adhere to ME180 cervical epithelial cells. (See Figures 30 and 211). Deletion of GBS 104 also reduces the capacity of GBS to invade J774 macrophage-like cells. (See Figures 32 and 205). Deletion of GBS 104 also causes GBS to translocate through epithelial monolayers less efficiently. See Figure 206. GBS 104 protein therefore appears to bind to ME180 epithelial cells and to have a role in adhesion to epithelial cells and macrophage cell lines.

Similar to the GBS bacteria that are deletion mutants for GBS 104, GBS 80 knockout mutant strains also partially lose the ability to translocate through an epithelial monolayer. See Figure 207. Deletion of either GBS 80 or GBS 104 in COH1 cells diminishes adherence to HUVEC endothelial cells. See Figure 208. Deletion of GBS 80 or GBS 104 in COH1 does not, however, affect growth of COH1 either with ME180 cells or in incubation medium (IM). See Figure 209. Both GBS 80 and GBS 104, therefore, appear to be involved in translocation of GBS through epithelial cells.

GBS 80 does not appear to bind to epithelial cells. Incubation of epithelial cells in the presence of GBS 80 protein followed by FACS analysis using an anti-GBS 80 polyclonal antibody did not detect GBS 80 binding to the epithelial cells. See Figure 202. Furthermore, deletion of GBS 80 protein does not affect the ability of GBS to adhere and invade ME180 cervical epithelial cells. See Figure 203

Preferably, one or more of the surface proteins may bind to one or more extracellular matrix (ECM) binding proteins, such as fibrinogen, fibronectin, or collagen. As shown in Figures 5 and 204, and Example 1, GBS 80, one of the AI-1 surface proteins, can bind to the extracellular matrix binding proteins fibronectin and fibrinogen. While GBS 80 protein apparently does not bind to certain epithelial cells or affect the capacity of a GBS bacteria to adhere to or invade cervical epithelial cells (See Figures 27 and 28), removal of GBS 80 from a wild type strain decreases the ability of that strain to translocate through an epithelial cell layer (see Figure 31).

GBS 80 may also be involved in formation of biofilms. COH1 bacteria overexpressing GBS 80 protein have an impaired ability to translocate through an epithelial monolayer. See Figure 212. These COH1 bacteria overexpressing GBS 80 form microcolonies on epithelial cells. See Figures 213 and 214. These microcolonies may be the initiation of biofilm development.

AI Surface proteins may also demonstrate functional homology to previously identified adhesion proteins or extracellular matrix (ECM) binding proteins. For example, GBS 80, a surface protein in AI-1, exhibits some functional homology to *FimA*, a major fimbrial subunit of a Gram positive bacteria *A. naeslundii*. *FimA* is thought to be involved in binding salivary proteins and may be a component in a fimbrae on the surface of *A. naeslundii*. *See* Yeung *et al.* (1997) Infection & Immunity 65:2629-2639; Yeunge *et al* (1998) J. Bacteriol 66:1482-1491; Yeung *et al.* (1988) J. Bacteriol 170:3803 – 3809; and Li *et al.* (2001) Infection & Immunity 69:7224-7233.

A similar functional homology has also been identified between GBS 80 and proteins involved in pili formation in the Gram positive bacteria *Corynebacterium diphtheriae* (SpaA, SpaD, and SpaH). *See*, Ton-That *et al.* (2003) Molecular Microbiology 50(4):1429-1438 and Ton-That *et al.*

(2004) Molecular Microbiology 53(1):251-261. The *C. diphtheriae* proteins all included a pilin motif of WxxxVxVYPK (SEQ ID NO: 151; where x indicates a varying amino acid residue). The lysine (K) residue is particularly conserved in the *C. diphtheriae* pilus proteins and is thought to be involved in sortase catalized oligomerization of the subunits involved in the *C. diphtheriae* pilus structure. (The *C. diphtheriae* pilin subunit SpaA is thought to occur by sortase-catalyzed amide bond cross-linking of adjacent pilin subunits. As the thioester-linked acyl intermediate of sortase requires nucleophilic attack for release, the conserved lysine within the SpaA pilin motif might function as an amino group acceptor of cleaved sorting signals, thereby providing for covalent linkages of the *C. diphtheria* pilin subunits. *See* Figure 6(d) of Ton-That et al., Molecular Microbiology (2003) 50(4):1429-1438.)

In addition, an "E box" comprising a conserved glutamic acid residue has also been identified in the *C. diphtheria* pilin associated proteins as important in *C. diphtheria* pilin assembly. The E box motif generally comprises YxLxETxAPxGY (SEQ ID NO: 152; where x indicates a varying amino acid residue). In particular, the conserved glutamic acid residue within the E box is thought necessary for *C. diphtheria* pilus formation.

Preferably, the AI-1 polypeptides of the immunogenic compositions comprise an E box motif. Some examples of E box motifs in the AI-1 polypeptides may include the amino acid sequences YxLxExxxxxGY (SEQ ID NO: 153), YxLxExxxPxGY (SEQ ID NO: 154), or YxLxETxAPxGY (SEQ ID NO: 152). Specifically, the E box motif of the polypeptides may comprise the amino acid sequences YKLKETKAPEGY (SEQ ID NO: 155), YVLKEIETQSGY (SEQ ID NO: 156), or YKLYEISSPDGY (SEQ ID NO: 157).

As discussed in more detail below, a pilin motif containing a conserved lysine residue and an E box motif containing a conserved glutamic acid residue have both been identified in GBS 80.

While previous publications have speculated that pilus-like structures might be formed on the surface of streptococci, (*see, e.g.*, Ton-That et al., Molecular Microbiology (2003) 50(4): 1429 – 1438), these structures have not been previously visible in negative stain (non-specific) electron micrographs, throwing such speculations into doubt. For example, Figure 34 presents electron micrographs of GBS serotype III, strain isolate COH1 with a plasmid insert to facilitate the overexpression of GBS 80. This EM photo was produced with a standard negative stain – no pilus structures are distinguishable. In addition, the use of such AI surface proteins in immunogenic compositions for the treatment or prevention of infection against a Gram positive bacteria has not been previously described.

Surprisingly, Applicants have now identified the presence of GBS 80 in surface exposed pilus formations visible in electron micrographs. These structures are only visible when the electron micrographs are specifically stained against an AI surface protein such as GBS 80. Examples of these electron micrographs are shown in Figures 11, 16 and 17, which reveal the presence of pilus structures in wild type COH1 *Streptococcus agalactiae.* Other examples of these electron

micrographs are shown in Figure 49, which reveals that GBS 80 is associated with pili in a wild type clinical isolate of *S. agalactiae*, JM9030013. (See figure 49.)

Applicants have also constructed mutant GBS strains containing a plasmid comprising the GBS 80 sequence resulting in the overexpression of GBS 80 within this mutant. The electron micrographs of Figures 13 – 15 are also stained against GBS 80 and reveal long, oligomeric structures containing GBS 80 which appear to cover portions of the surface of the bacteria and stretch far out into the supernatant.

In some instances, the formation of pili structures on GBS appears to be correlated to surface expression of GBS 80. Figure 61 provides FAC analysis of GBS 80 surface levels on bacterial strains COH1 and JM9130013 using an anti-GBS 80 antisera. Immunogold electron microscopy of the COH1 and JM9130013 bacteria using anti-GBS 80 antisera demonstrates that JM9130013 bacteria, which have higher values for GBS 80 surface expression, also form longer pili structures.

The surface exposure of GBS 80 on GBS is generally not capsule-dependent. Figure 62 provides FACS analysis of capsulated and uncapsulated GBS analyzed with anti-GBS 80 and anti-GBS 322 antibodies. Surface exposure of GBS 80, unlike GBS 322, is not capsule dependent.

An Adhesin Island surface protein, such as GBS 80 appears to be required for pili formation, as well as an Adhesin Island sortase. Pili are formed in Coh1 bacterial clones that overexpress GBS 80, but lack GBS 104, or one of the AI-1 sortases sag0647 or sag0648. However, pili are not formed in Coh1 bacterial clones that overexpress GBS 80 and lack both sag0647 and sag0648. Thus, for example, it appears that at least GBS 80 and a sortase, sag0647 or sag0648, may be necessary for pili formation. (See Figure 48.) Overexpression of GBS 80 in GBS strain 515, which lacks an AI-1, also assembles GBS 80 into pili. GBS strain 515 contains an AI-2, and thus AI-2 sortases. The AI-2 sortases in GBS strain 515 apparently polymerize GBS 80 into pili. (See Figure 42.) Overexpression of GBS 80 in GBS strain 515 cell knocked out for GBS 67 expression also apparently polymerizes GBS 80 into pili. (See Figure 72.)

While GBS 80 appears to be required for GBS AI-1 pili formation, GBS 104 and sortase SAG0648 appears to be important for efficent AI-1 pili assembly. For example, high-molecular structures are not assembled in isogenic COH1 strains which lack expression of GBS 80 due to gene disruption and are less efficiently assembled in isogenic COH1 strains which lack the expression of GBS 104 (see Figure 41). This GBS strain comprises high molecular weight pili structures composed of covalently linked GBS 80 and GBS 104 subunits. In addition, deleting SAG0648 in COH1 bacteria interferes with assembly of some of the high molecular weight pili structures. Thus, indicating that SAG0648 plays a role in assembly of these pilin species. (See Figure 41).

EM photos confirm the involvement of AI surface protein GBS 104 within the hyperoligomeric structures of a GBS strain adapted for increased GBS 80 expression. (See Figures 34 – 41 and Example 6). In a wild type serotype VIII GBS strain, strain JM9030013, IEM identifies GBS 104 as forming clusters on the bacterial surface. (See Figure 50.)

PC GBS 52 also appears to be a component of the GBS pili. Immunoblots using an anti-GBS 80 antisera on total cell extracts of Coh1 and a GBS 52 null mutant Coh1 reveal a shift in detected proteins in the Coh1 wild type strain relative to the GBS 52 null mutant Coh1 strain. The shifted proteins were also detected in the wild type Coh1 bacteria with an anti-GBS 52 antisera, indicating that the GBS 52 may be present in the pilus. (See Figure 45.)

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising an AI surface protein such as GBS 80. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine amino acid residue.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include one or both of a pilin motif comprising a conserved lysine residue and an E box motif comprising a conserved glutamic acid residue.

More than one AI surface protein may be present in the oligomeric, pilus-like structures of the invention. For example, GBS 80 and GBS 104 may be incorporated into an oligomeric structure. Alternatively, GBS 80 and GBS 52 may be incorporated into an oligomeric structure, or GBS 80, GBS 104 and GBS 52 may be incorporated into an oligomeric structure.

In another embodiment, the invention includes compositions comprising two or more AI surface proteins. The composition may include surface proteins from the same adhesin island. For example, the composition may include two or more GBS AI-1 surface proteins, such as GBS 80, GBS 104 and GBS 52. The surface proteins may be isolated from Gram positve bacteria or they may be produced recombinantly.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a GBS Adhesin Island protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more GBS Adhesin Island 1 ("AI-1") proteins and one or more GBS Adhesin Island 2 ("AI-2") proteins, wherein one or more of the Adhesin Island proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

The oligomeric, pilus-like structures of the invention may be combined with one or more additional GBS proteins. In one embodiment, the oligomeric, pilus-like structures comprise one or more AI surface proteins in combination with a second GBS protein. The second GBS protein may be a known GBS antigen, such as GBS 322 (commonly referred to as "sip") or GBS 276. Nucleotide and amino acid sequences of GBS 322 sequenced from serotype V isolated strain 2603 V/R are set

forth in WO 02/34771 as SEQ ID 8539 and SEQ ID 8540 and in the present specification as SEQ ID NOs: 38 and 39. A particularly preferred GBS 322 polypeptide lacks the N-terminal signal peptide, amino acid residues 1-24. An example of a preferred GBS 322 polypeptide is a 407 amino acid fragment and is shown in SEQ ID NO: 40. Examples of preferred GBS 322 polypeptides are further described in PCTUS04/_____, attorney docket number PP20665.002 filed September 15, 2004, hereby incorporated by reference, published as WO 2005/002619.

Additional GBS proteins which may be combined with the GBS AI surface proteins of the invention are also described in WO 2005/002619. These GBS proteins include GBS 91, GBS 184, GBS 305, GBS 330, GBS 338, GBS 361, GBS 404, GBS 690, and GBS 691.

Additional GBS proteins which may be combined with the GBS AI surface proteins of the invention are described in WO 02/34771.

GBS polysaccharides which may be combined with the GBS AI surface proteins of the invention are described in WO 2004/041157. For example, the GBS AI surface proteins of the invention may be combined with a GBS polysaccharides selected from the group consisting of serotype Ia, Ib, Ia/c, II, III, IV, V, VI, VII and VIII.

The oligomeric, pilus-like structures may be isolated or purified from bacterial cultures in which the bacteria express an AI surface protein. The invention therefore includes a method for manufacturing an oligomeric AI surface antigen comprising culturing a GBS bacterium that expresses the oligomeric AI protein and isolating the expressed oligomeric AI protein from the GBS bacteria. The AI protein may be collected from secretions into the supernatant or it may be purified from the bacterial surface. The method may further comprise purification of the expressed AI protein. Preferably, the AI protein is in a hyperoligomeric form. Macromolecular structures associated with oligomeric pili are observed in the supernatant of cultured GBS strain Coh1. (See Figure 46.) These pili are found in the supernatant at all growth phases of the cultured Coh1 bacteria. (See Figure 47.)

The oligomeric, pilus-like structures may be isolated or purified from bacterial cultures overexpressing an AI surface protein. The invention therefore includes a method for manufacturing an oligomeric Adhesin Island surface antigen comprising culturing a GBS bacterium adapted for increased AI protein expression and isolation of the expressed oligomeric Adhesin Island protein from the GBS bacteria. The AI protein may be collected from secretions into the supernatant or it may be purified from the bacterial surface. The method may further comprise purification of the expressed Adhesin Island protein. Preferably, the Adhesin Island protein is in a hyperoligomeric form.

The GBS bacteria are preferably adapted to increase AI protein expression by at least two (*e.g.*, 2, 3, 4, 5, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 125, 150 or 200) times wild type expression levels.

GBS bacteria may be adapted to increase AI protein expression by any means known in the art, including methods of increasing gene dosage and methods of gene upregulation. Such means include, for example, transformation of the GBS bacteria with a plasmid encoding the AI protein. The plasmid may include a strong promoter or it may include multiple copies of the sequence encoding the

AI protein. Optionally, the sequence encoding the AI protein within the GBS bacterial genome may be deleted. Alternatively, or in addition, the promoter regulating the GBS Adhesin Island may be modified to increase expression.

GBS bacteria harbouring a GBS AI-1 may also be adapted to increase AI protein expression by altering the number adenosine nucleotides present at two sites in the intergenic region between AraC and GBS 80. See Figure 197 A, which is a schematic showing the organization of GBS AI-1 and Figure 197 B, which provides the sequence of the intergenic region between AraC and GBS 80 in the AI. The adenosine tracts which applicants have identified as influencing GBS 80 surface expression are at nucleotide positions 187 and 233 of the sequence shown in Figure 197 B (SEQ ID NO: 273). Applicants determined the influence of these adenosine tracts on GBS 80 surface expression in strains of GBS bacteria harboring four adenosines at position 187 and six adenosines at position 233, five adenosines at position 187 and six adenosines position 233, and five adenosines at position 187 and seven adenosines at position 233. FACS analysis of these strains using anti GBS 80 antiserum determined that an intergenic region with five adenosines at position 187 and six adenosines at position 233 had higher expression levels of GBS 80 on their surface than other stains. See Figure 197 C for results obtained from the FACS analysis. Therefore, manipulating the number of adenosines present at positions 187 and 233 of the AraC and GBS 80 intergenic region may further be used to adapt GBS to increase AI protein expression.

The invention further includes GBS bacteria which have been adapted to produce increased levels of AI surface protein. In particular, the invention includes GBS bacteria which have been adapted to produce oligomeric or hyperoligomeric AI surface protein, such as GBS 80. In one embodiment, the Gram positive bacteria of the invention are inactivated or attenuated to permit *in vivo* delivery of the whole bacteria, with the AI surface protein exposed on its surface.

The invention further includes GBS bacteria which have been adapted to have increased levels of expressed AI protein incorporated in pili on their surface. The GBS bacteria may be adapted to have increased exposure of oligomeric or hyperoligomeric AI proteins on its surface by increasing expression levels of a signal peptidase polypeptide. Increased levels of a local signal peptidase expression in Gram positive bacteria (such us LepA in GAS) are expected to result in increased exposure of pili proteins on the surface of Gram positive bacteria. Increased expression of a leader peptidase in GBS may be achieved by any means known in the art, such as increasing gene dosage and methods of gene upregulation. The GBS bacteria adapted to have increased levels of leader peptidase may additionally be adapted to express increased levels of at least one pili protein.

Alternatively, the AI proteins of the invention may be expressed on the surface of a non-pathogenic Gram positive bacteria, such as *Streptococus gordonii* (See, e.g., Byrd et al., "Biological consequences of antigen and cytokine co-expression by recombinant *Streptococcus gordonii* vaccine vectors", Vaccine (2002) 20:2197-2205) or *Lactococcus lactis* (See, e.g., Mannam et al., "Mucosal Vaccine Made from Live, Recombinant Lactococcus lactis Protects Mice against Pharangeal Infection with *Streptococcus pyogenes*" Infection and Immunity (2004) 72(6):3444-3450). As used herein,

SUBSTITUTE SHEET (RULE 26)

non-pathogenic Gram positive bacteria refer to Gram positive bacteria which are compatible with a human host subject and are not associated with human pathogenisis. Preferably, the non-pathogenic bacteria are modified to express the AI surface protein in oligomeric, or hyper-oligomeric form. Sequences encoding for an AI surface protein and, optionally, an AI sortase, may be integrated into the non-pathogenic Gram positive bacterial genome or inserted into a plasmid. The non-pathogenic Gram positive bacteria may be inactivated or attenuated to facilitate *in vivo* delivery of the whole bacteria, with the AI surface protein exposed on its surface. Alternatively, the AI surface protein may be isolated or purified from a bacterial culture of the non-pathogenic Gram positive bacteria. For example, the AI surface protein may be isolated from cell extracts or culture supernatants. Alternatively, the AI surface protein may be isolated or purified from the surface of the non-pathogenic Gram positive bacteria.

The non-pathogenic Gram positive bacteria may be used to express any of the Gram positive bacterial Adhesin Island proteins described herein, including proteins from a GBS Adhesin Island, a GAS Adhesin Island, or a S pneumo Adhesin Island. The non-pathogenic Gram positive bacteria are transformed to express an Adhesin Island surface protein. Preferably, the non-pathogenic Gram positive bacteria also express at least one Adhesin Island sortase. The AI transformed non-pathogenic Gram positive bacteria of the invention may be used to prevent or treat infection with a pathogenic Gram positive bacteria, such as GBS, GAS or *Streptococcus pneumoniae*. The non-pathogenic Gram positive bacteria may express the Gram positive bacterial Adheshin Island proteins in oligomeric forms that further comprise adhesin island proteins encoded within the genome of the non-pathogenic Gram positive bacteria.

Applicants modified *L. lactis* to demonstrate that it can express GBS AI polypeptides. *L. lactis* was transformed with a construct encoding GBS 80 under its own promoter and terminator sequences. The transformed *L. lactis* appeared to express GBS 80 as shown by Western blot analysis using anti-GBS 80 antiserum. See lanes 6 and 7 of the Western Blots provided in Figures 133A and 133B (133A and 133B are two different exposures of the same Western blot). See also Example 13.

Applicants also transformed *L. lactis* with a construct encoding GBS AI-1 polypeptides GBS 80, GBS 52, SAG0647, SAG0648, and GBS 104 under the GBS 80 promoter and terminator sequences. These *L. lactis* expressed high molecular weight structures that were immunoreactive with anti-GBS 80 in immunoblots. See Figure 134, lane 2, which shows detection of a GBS 80 monomer and higher molecular weight polymers in total transformed *L. lactis* extracts. Thus, it appeared that *L. lactis* is capable of expressing GBS 80 in oligomeric form. The high molecular weight polymers were not only detected in *L. lactis* extracts, but also in the culture supernatants. See Figure 135 at lane 4. See also Example 14. Thus, the GBS AI polypeptides in oligomeric form can be isolated and purified from either *L. lactis* cell extracts or culture supernatants. These oligomeric forms can, for instance, be isolated from cell extracts or culture supernatants by release by sonication. See Figure 136A and B. See also Figure 171, which shows purification of GBS pili from whole extracts of *L. lactis* expressing the GBS AI-1 following sonication and gel filtration on a Sephacryl HR 400 column.

Furthermore, the *L. lactis* transformed with the construct encoding GBS AI-1 polypeptides GBS 80, GBS 52, SAG0647, SAG0648, and GBS 104 under the GBS 80 promoter and terminator sequences expressed the GBS AI-1 polypeptides on its surface. FACS analysis of these transformed *L. lactis* detected cell surface expression of both GBS 80 and GBS 104. The surface expression levels of GBS 80 and GBS 104 on the transformed *L. lactis* were similar to the surface expression levels of GBS 80 and GBS 104 on GBS strains COH1 and JM9130013, which naturally express GBS AI-1. See Figure 169 for FACS analysis data for *L. lactis* transformed with GBS AI-1 and wildtype JM9130013 bacteria using anti-GBS 80 and GBS 104 antisera. Table 40 provides the results of FACS analysis of transformed *L. lactis*, COH1, and JM9130013 bacteria using anti-GBS 80 and anti-GBS 104 antisera. The numbers provided represent the mean fluorescence value difference calculated for immune versus pre-immune sera obtained for each bacterial strain.

Table 40: FACS analysis of *L. lactis* and GBS bacteria strains expressing GBS AI-1

	Anti-GBS 80 antiserum	Anti-GBS 104 antiserum
GBS AI-1 transformed *L. lactis*	298	251
GBS COH1	305	305
GBS JM9130013	461	355

Immunogold-electronmicroscopy performed with anti-GBS 80 primary antibodies detected the presence of pilus structures on the surface of the *L. lactis* bacteria expressing GBS AI-1, confirming the results of the FACS analysis. See Figure 168 B and C. Interestingly, this expression of GBS pili on the surface of the *L. lactis* induced *L. lactis* aggregation. See Figure 170. Thus, GBS AI polypeptides may also be isolated and purified from the surface of *L. lactis*. The ability of *L. lactis* to express GBS AI polypeptides on its surface also demonstrates that it may be useful as a host to deliver GBS AI antigens.

In fact, immunization of mice with *L. lactis* transformed with GBS AI-1 was protective in a subsequent challenge with GBS. Female mice were immunized with *L. lactis* transformed with GBS AI-1. The immunized female mice were bred and their pups were challenged with a dose of GBS sufficient to kill 90% of non-immunized pups. Detailed protocols for intranasal and subcutaneous immunization of mice with transformed *L. lactis* can be found in Examples 18 and 19, respectively. Table 43 provides data showing that immunization of the female mice with *L. lactis* expressing GBS AI-1 (LL-AI 1) greatly increased survival rate of challenged pups relative to both a negative PBS control (PBS) and a negative *L. lactis* control (LL 10 E9, which is wild type *L. lactis* not transformed to express GBS AI-1).

Table 43: Protection of Mice Immunized with *L. lactis* expressing GBS AI-1

Immunization Route	**Antigen**	**Alive/Treated**	**Survival %**	**Survival % Range**	**p value**
Intraperitoneum	Recombinant GBS 80	16/18	89	80-100	<0.001
Subcutaneous	LL-AI 1 10 E9	40/49	82	70-90	<0.001
	LL-AI 1 10 E10	50/60	83	60-100	<0.001
	PBS	4/30	13	0-30	
	LL 10 E9	3/57	5	0-20	
Intranasal	LL-AI 1 10 E9	22/60	37	0-100	0.02

PCT/US05	LL-AI 1 10 E10	31/49	63	30-90	<0.001
	LL 10 E9	2/27	7	0-20	

Table 51 provides further evidence that immunization of mice with *L. lactis* transformed with GBS AI-1 is protective against GBS.

Table 51: Further Protection of Mice Immunized with *L. lactis* expressing GBS AI-1

Antigen	Immunization route	Alive/Treated	Survival % (*P*val <0.0000001)
Recombinant GBS 80	IP	48/50	**92**
Recombinant GBS 80	SC	21/30	**70**
L.lactis+AI1 10^6 cfu	SC	6/66	9
L.lactis+AI1 10^7 cfu	SC	47/70	**73**
L.lactis+AI1 10^8 cfu	SC	116/153	**76**
L.lactis+AI1 10^9 cfu	SC	98/118	**83**
L.lactis+AI1 10^{10} cfu	SC	107/129	**83**
L.lactis 10^{10} cfu	SC	4/83	5
PBS	SC	6/110	5
L.lactis+AI1 10^{10} cfu	IN	51/97	**52**
L.lactis 10^{11} cfu	IN	1/40	7
PBS	IN	0/37	0

Protection of immunized mice with *L. lactis* expressing the GBS AI-1 is at least partly due to a newly raised antibody response. Table 46 provides anti-GBS 80 antibody titers detected in serum of the mice immunized with *L. lactis* expressing the GBS AI-1 as described above. Mice immunized with *L. lactis* expressing the GBS AI-1 have anti-GBS 80 antibody titres, which are not observed in mice immunized with *L. lactis* not transformed to express the GBS AI-1. Further, as expected from the survival data, mice subcutaneously immunized with *L. lactis* transformed to express the GBS AI-1 have significantly higher serum anti-GBS 80 antibody titers than mice intranasally immunized with *L. lactis* transformed to express the GBS AI-1.

Table 46: Antibody Responses against GBS 80 in Serum of Mice Immunized with *L. lactis* Expressing GBS AI-1

Antigen	Ab Titre Obtained Following		
	Subcutaneous Immunization	Intranasal Immunization	Intraperitoneal Immunization
LL 10 E9	0	0	
LL-AI 1 10 E9	14000	50	
LL-AI 1 10 E10	25000	406	
Recombinant GBS 80			120000

Anti-GBS 80 antibodies of the IgA isotype were specifically detected in various body fluids of the mice subcutaneously or intranasally immunized with *L. lactis* expressing the GBS AI-1.

Table 47: Anti-GBS 80 IgA Antibodies Detected in Mouse Tissues Following Immunization with *L. lactis* Expressing GBS AI-1

Antigen	Immunization route	Anti-GBS 80 IgA Antibodies Detected in		
		Serum	Vaginal Wash	Nasal Wash
LL 10 E9		0	0	0
LL-AI 1	Subcutaneous	0	25	20
LL-AI 1	Intranasal	140	0	150
GBS 80	Intraperitoneal	60	0	

Furthermore, opsonophagocytosis assays also demonstrated that at least some of the antiserum produced against the *L. lactis* expressing GBS AI 1 is opsonic for GBS. See Figure 161.

To obtain protection of against GBS across a greater number of strains and serotypes, it is possible to transform *L. lactis* with a recombinant GBS AI encoding both GBS AI-1 and AI-2, *i.e.*, a hybrid GBS AI. By way of example, a hybrid GBS AI may be a GBS AI-1 with a replacement of the GBS 104 gene with a GBS 67 gene. A schematic of such a hybrid GBS AI is depicted in Figure 231 A. A hybrid GBS AI may alternatively be a GBS AI-1 with a replacement of the GBS 52 gene with a GBS 59 gene. See the schematic at Figure 231 B. Alternatively, a hybrid GBS AI may be a GBS AI-1 with a substitution of a GBS 59 polypeptide for the GBS 52 gene and a substitution of the GBS 104 gene for genes encoding GBS 59 and the two GBS AI-2 sortases. Another example of a hybrid GBS AI is a GBS AI-1 with the substitution of a GBS 59 gene for the GBS 52 gene and a GBS 67 for the GBS 104 gene. See the schematic at Figure 232. A further example of a hybrid GBS AI is a GBS AI-1 having a GBS 59 gene and genes encoding the GBS AI-2 sortases in place of the GBS 52 gene. Yet another example of a hybrid GBS AI is a GBS AI-1 with a substitution of either GBS 52 or GBS 104 with a fusion protein comprising GBS 322 and one of GBS 59, GBS 67, or GBS 150. Some of these hybrid GBS AIs may be prepared as briefly outlined in Figure 234 A-F.

Applicants have prepared a hybrid GBS AI having a GBS AI-1 sequence with a substitution of a GBS 67 coding sequence for the GBS 104 gene as depicted in Figure 231 A. Transformation of *L. lactis* with the hybrid GBS AI-1 resulted in *L. lactis* expression of high molecular weight polymers containing the GBS 80 and GBS 67 proteins. See Figure 233 A, which provides Western blot analysis of *L. lactis* transformed with the hybrid GBS AI depicted in Figure 231 A. When *L. lactis* transformed with the hybrid GBS AI were probed with antibodies to GBS 80 or GBS 67, high molecular weight structures were detected. See lanes labelled LL + a) in both the α-80 and α-67 immunoblots. The GBS 80 and GBS 67 proteins were confirmed to be present on the surface of *L. lactis* by FACS analysis. See Figure 233 B, which shows a shift in fluorescence when GBS 80 and GBS 67 antibodies are used to detect GBS 80 and GBS 67 surface expression. The same shifts in fluorescence were not observed in *L. lactis* control cells, cells not transformed with the hybrid GBS AI.

Alternatively, the oligomeric, pilus-like structures may be produced recombinantly. If produced in a recombinant host cell system, the AI surface protein will preferably be expressed in coordination with the expression of one or more of the AI sortases of the invention. Such AI sortases will facilitate oligomeric or hyperoligomeric formation of the AI surface protein subunits.

PCT/US05/27239

AI-Sortases of the invention will typically have a signal peptide sequence within the first 70 amino acid residues. They may also include a transmembrane sequence within 50 amino acid residues of the C terminus. The sortases may also include at least one basic amino acid residue within the last 8 amino acids. Preferably, the sortases have one or more active site residues, such as a catalytic cysteine and histidine.

As shown in Figure 1, AI-1 includes the surface exposed proteins of GBS 80, GBS 52 and GBS 104 and the sortases SAG0647 and SAG0648. AI-1 typically appears as an insertion into the 3' end of the *trmA* gene.

In addition to the open reading frames encoding the AI-1 proteins, AI-1 may also include a divergently transcribed transcriptional regulator such as *araC* (*i.e.*, the transcriptional regulator is located near or adjacent to the AI protein open reading frames, but it transcribed in the opposite direction). It is believed that *araC* may regulate the expression of the AI operon. (*See* Korbel et al., Nature Biotechnology (2004) 22(7): 911 – 917 for a discussion of divergently transcribed regulators in *E. coli*).

AI-1 may also include a sequence encoding a *rho* independent transcriptional terminator (see hairpin structure in Figure 1). The presence of this structure within the adhesin island is thought to interrupt transcription after the GBS 80 open reading frame, leading to increased expression of this surface protein.

A schematic identifying AI-1 within several GBS serotypes is depicted in Figure 2. AI-1 sequences were identified in GBS serotype V, strain isolate 2603; GBS serotype III, strain isolate NEM316; GBS serotype II, strain isolate 18RS21; GBS serotype V, strain isolate CJB111; GBS serotype III, strain isolate COH1 and GBS serotype 1a, strain isolate A909. (Percentages shown are amino acid identity to the 2603 sequence). (An AI-1 was not identified in GBS serotype 1b, strain isolate H36B or GBS serotype 1a, strain isolate 515).

An alignment of AI-1 polynucleotide sequences from serotype V, strain isolates 2603 and CJB111; serotype II, strain isolate 18RS21; serotype III, strain isolates COH1 and NEM316; and serotype 1a, strain isolate A909 is presented in Figure 18. An alignment of amino acid sequences of AI-1 surface protein GBS 80 from serotype V, strain isolates 2603 and CJB111; serotype 1a, strain isolate A909; serotype III, strain isolates COH1 and NEM316 is presented in Figure 22. An alignment of amino acid sequences of AI-1 surface protein GBS 104 from serotype V, strain isolates 2603 and CJB111; serotype III, strain isolates COH1 and NEM316; and serotype II, strain isolate 18RS21 is presented in Figure 23. Preferred AI-1 polynucleotide and amino acid sequences are conserved among two or more GBS serotypes or strain isolates.

As shown in this figure, the full length of surface protein GBS 80 is particularly conserved among GBS serotypes V (strain isolates 2603 and CJBIII), III (strain isolates NEM316 and COH1), and Ia (strain isolate A909). The GBS 80 surface protein is missing or fragmented in serotypes II (strain isolate 18RS21), Ib (strain isolate H36B) and Ia (strain isolate 515).

Polynucleotide and amino acid sequences for *AraC* are set forth in FIGURE 30.

GBS Adhesin Island 2

A second adhesin island, "Adhesin Island 2" or "AI-2" or "GBS AI-2" has also been identified in numerous GBS serotypes. A schematic depicting the correlation between AI-1 and AI-2 within the GBS serotype V, strain isolate 2603 is shown in Figure 3. (Homology percentages in Figure 3 represent amino acid identity of the AI-2 proteins to the AI-1 proteins). Alignments of AI-2 polynucleotide sequences are presented in Figures 20 and 21 (Figure 20 includes sequences from serotype V, strain isolate 2603 and serotype III, strain isolate NEM316. Figure 21 includes sequences from serotype III, strain isolate COH1 and serotype Ia, strain isolate A909). An alignment of amino acid sequences of AI-2 surface protein GBS 067 from serotype V, strain isolates 2603 and CJB111; serotype 1a, strain isolate 515; serotype II, strain isolate 18RS21; serotype Ib, strain isolate H36B; and serotype III, strain isolate NEM316 is presented in Figure 24. Preferred AI-2 polynucleotide and amino acid sequences are conserved among two or more GBS serotypes or strain isolates.

AI-2 comprises a series of approximately five open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, AI-2 includes open reading frames encoding for two or more (i.e., 2, 3, 4, 5 or more) of GBS 67, GBS 59, GBS 150, SAG1405, SAG1406, 01520, 01521, 01522, 01523, 01523, 01524 and 01525. In one embodiment, AI-2 includes open reading frames encoding for two or more of GBS 67, GBS 59, GBS 150, SAG1405, and SAG1406. Alternatively, AI-2 may include open reading frames encoding for two or more of 01520, 01521, 01522, 01523, 01523, 01524 and 01525.

One or more of the surface proteins typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. The GBS AI-2 sortase proteins are thought to be involved in the secretion and anchoring of the LPXTG containing surface proteins. GBS AI-2 may encode for at least one surface protein. Alternatively, AI-2 may encode for at least two surface proteins and at least one sortase. Preferably, GBS AI-2 encodes for at least three surface proteins and at least two sortases. One or more of the AI-2 surface proteins may include an LPXTG or other sortase substrate motif.

One or more of the surface proteins may also typically include pilin motif. The pilin motif may be involved in pili formation. Cleavage of AI surface proteins by sortase between the threonine and glycine residue of an LPXTG motif yields a thioester-linked acyl intermediate of sortase. The first lysine residue in a pilin motif can serve as an amino group acceptor of the cleaved LPXTG motif and thereby provide a covalent linkage between AI subunits to form pili. For example, the pilin motif can make a nucleophilic attack on the acyl enzyme providing a covalent linkage between AI subunits to form pili and regenerate the sortase enzyme. Some examples of pilin motifs that may be present in the GBS AI-2 proteins include ((YPKN(X_8)K; SEQ ID NO: 158), (PK(X_8)K; SEQ ID NO: 159), (YPK(X_9)K;SEQ ID NO: 160), (PKN(X_8)K; SEQ ID NO: 161), or (PK(X_{10})K; SEQ ID NO: 162)).

One or more of the surface protein may also include an E box motif. The E box motif contains a conserved glutamic acid residue that is believed to be necessary for pilus formation. Some examples of E box motifs may include the amino acid sequences YxLxETxAPxG (SEQ ID NO: 163),

YXXXEXXAXXGY (SEQ ID NO: 164), YXLXEXXXPXDY (SEQ ID NO: 165), or YXLXETXAPXGY (SEQ ID NO: 152).

As shown in Figure 3, GBS AI-2 may include the surface exposed proteins of GBS 67, GBS 59 and GBS 150 and the sortases of SAG1406 and SAG1405. Alternatively, GBS AI-2 may include the proteins 01521, 01524 and 01525 and sortases 01520 and 01522. GBS 067 and 01524 are preferred AI-2 surface proteins.

AI-2 may also include a divergently transcribed transcriptional regulator such as a RofA like protein (for example *rogB*). As in AI-1, *rogB* is thought to regulate the expression of the AI-2 operon.

A schematic depiction of AI-2 within several GBS serotypes is depicted in Figure 4. (Percentages shown are amino acid identity to the 2603 sequence). While the AI-2 surface proteins GBS 59 and GBS 67 are more variable across GBS serotypes than the corresponding AI-1 surface proteins, AI-2 surface protein GBS 67 appears to be conserved in GBS serotypes where the AI-1 surface proteins are disrupted or missing.

For example, as discussed above and in Figure 2, the AI-1 GBS 80 surface protein is fragmented in GBS serotype II, strain isolate 18RS21. Within AI-2 for this same sequence, as shown in Figure 4, the GBS 67 surface protein has 99% amino acid sequence homology with the corresponding sequence in strain isolate 2603. Similarly, the AI-1 GBS 80 surface protein appears to be missing in GBS serotype Ib, strain isolate H36B and GBS serotype Ia, strain isolate 515. Within AI-2 for these sequences, however, the GBS 67 surface protein has 97 – 99 % amino acid sequence homology with the corresponding sequence in strain isolate 2603. GBS 67 appears to have two allelic variants, which can be divided according to percent homology with strains 2603 and H36B. See figures 237-239.

Unlike for GBS 67, amino acid sequence identity of GBS 59 is variable across different GBS strains. As shown in Figures 63 and 224, GBS 59 of GBS strain isolate 2603 shares 100% amino acid residue homology with GBS strain 18RS21, 62% amino acid sequence homology with GBS strain H36B, 48% amino acid residue homology with GBS strain 515 and GBS strain CJB111, and 47% amino acid residue homology with GBS strain NEM316. The amino acid sequence homologies of the different GBS strains suggest that there are two isoforms of GBS 59. The first isoform appears to include the GBS 59 protein of GBS strains CJB111, NEM316, and 515. The second isoform appears to include the GBS 59 protein of GBS strains 18RS21, 2603, and H36B. (See Figures 63 and 224.)

As expected from the variability in GBS 59 isoforms, antibodies specific for the first GBS 59 isoform detect the first but not the second GBS 59 isoform and antibodies specific for the second GBS 59 isoform detect the second but not the first GBS 59 isoform. See Figure 226A, which shows FACS analysis of 28 GBS strains having a GBS 59 gene detected using PCR for GBS 59 surface expression. For each of the 28 GBS strains, FACS analysis was performed using either an antibody for GBS 59 isoform 1 (α-cjb111) or GBS 59 isoform 2 (α -2603). Only one of the two antibodies detected GBS 59 surface expression on each GBS strain. As a negative control, GBS strains in which a GBS 59

DEMANDE/PCT

gene was not detectable by PCR did not have significant GBS 59 surface expression levels. Figure 226B.

Also, GBS 59 is opsonic only against GBS strains expressing a homologous GBS 59 protein. See Figure 225.

In one embodiment, the immunogenic composition of the invention comprises a first and a second isoform of the GBS 59 protein to provide protection across a wide range of GBS serotypes that express polypeptides from a GBS AI-2. The first isoform may be the GBS 59 protein of GBS strain CJB111, NEM316, or 515. The second isoform may be the GBS 59 protein of GBS strain 18RS21, 2603, or H36B.

The gene encoding GBS 59 has been identified in a high number of GBS isolates; the GBS 59 gene was detected in 31 of 40 GBS isolates tested (77.5%). The GBS 59 protein also appears to be present as part of a pilus in whole extracts derived from GBS strains. Figure 64 shows detection of high molecular weight GBS 59 polymers in whole extracts of GBS strains CJB111, 7357B, COH31, D1363C, 5408, 1999, 5364, 5518, and 515 using antiserum raised against GBS 59 of GBS strain CJB111. Figure 65 also shows detection of these high molecular weight GBS 59 polymers in whole extracts of GBS strains D136C, 515, and CJB111 with anti-GBS 59 antiserum. (See also Figure 220 A for detection of GBS 59 high molecular weight polymers in strain 515.) Figure 65 confirms the presence of different isoforms of GBS 59. Antisera raised against two different GBS 59 isoforms results in different patterns of immunoreactivity depending on the GBS strain origin of the whole extract. Figure 65 further shows detection of GBS 59 monomers in purified GBS 59 preparations.

GBS 59 is also highly expressed on the surface of GBS strains. GBS 59 was detected on the surface of GBS strains CJB111, DK1, DK8, Davis, 515, 2986, 5551, 1169, and 7357B by FACS analysis using mouse antiserum raised against GBS 59 of GBS CJB111. FACS analysis did not detect surface expression of GBS 59 in GBS strains SMU071, JM9130013, and COH1, which do not contain a GBS 59 gene. (See Figure 66.) Further confirmation that GBS 59 is expressed on the surface of GBS is detection of GBS 59 by immuno-electron microscopy on the surface of GBS strain 515 bacteria. See Figure 215.

GBS 67 and GBS 150 also appear to be included in high molecular weight structures, or pili. Figure 69 shows that anti-GBS 67 and anti-GBS 150 immunoreact with high molecular weight structures in whole GBS strain 515 extracts. (See also Figure 220 B and C.) It is also notable in Figure 69 that the anti-GBS 59 antisera, raised in a mouse following immunization with GBS 59 of GBS strain 2603, does not cross-hybridize with GBS 59 in GBS strain 515. GBS 59 of GBS stain 515 is of a different isotype than GBS 59 of GBS stain 2603. See Figure 63, which illustrates that the homology of these two GBS 59 polypeptides is 48%, and Figure 65, which confirms that GBS 59 antisera raised against GBS strain 2603 does not cross-hybridize with GBS 59 of GBS strain 515.

Formation of pili containing GBS 150 does not appear to require GBS 67 expression. Figure 70 provides Western blots showing that higher molecular weight structures in GBS strain 515 total

extracts immunoreact with anti-GBS 67 and anti-GBS 150 antiserum. In a GBS strain 515 lacking GBS 67 expression, anti-GBS 67 antiserum no longer immunoreacts with polypeptides in total extracts, while anti-GBS 150 antiserum is still able to cross-hybridze with high molecular weight structures.

Likewise, formation of pili containing GBS 59 does not appear to require GBS 67 expression. As expected, FACS detects GBS 67 cell surface expression on wildtype GBS strain 515, but not GBS strain 515 cells knocked out for GBS 67. FACS analysis using anti-GBS 59 antisera, however, detects GBS 59 expression on both the wildtype GBS strain 515 cells and the GBS strain 515 cells knocked out for GBS 67. Thus, GBS 59 cell surface expression is detected on GBS stain 515 cells regardless of GBS 67 expression.

GBS 67, while present in pili, appears to be localized around the surface of GBS strain 515 cells. See the immuno-electron micrographs presented in Figure 216. GBS 67 binds to fibronectin. See Figure 217.

Formation of pili encoded by GBS AI-2 does require expression of GBS 59. Deletion of GBS 59 from strain 515 bacteria eliminates detection of high molecular weight structures by antibodies that bind to GBS 59 (Figure 221 A, lane 3), GBS 67 (Figure 221 B, lane 3), and GBS 150 (Figure 221 C, lane 3). By contrast, Western blot analysis of 515 bacteria with a deletion of the GBS 67 gene detects high molecular weight structures using GBS 59 (Figure 221 A, lane 2) and GBS 150 (Figure 221 C, lane 2) antisera. Similarly, Western blot analysis of 515 bacteria with a deletion of the GBS 150 gene detects high molecular weight structures using GBS 59 (Figure 221 A, lane 4) and GBS 67 (Figure 221 B, lane 4). See also Figure 223, which provides Western blots of each of the 515 strains interrogated with antibodies for GBS 59, GBS 67, and GBS 150. FACS analysis of strain 515 bacteria deleted for either GBS 59 or GBS 67 confirms these results. See Figure 222, which shows that only deletion of GBS 59 abolishes surface expression of both GBS 59 and GBS 67.

Formation of pili encoded by GBS AI-2 also requires expression of both GBS adhesin island-2 encoded sortases. See Figure 218, which provides Western blot analysis of strain 515 bacteria lacking Srt1, Srt2, or both Srt1 and Srt2. Only deletion of both Srt1 and Srt2 abolishes pilus assembly as detected by antibodies that cross-hybridize with each of GBS 59, GBS 67 and GBS 150. The results of the Western blot analysis were verified by FACS, which provided similar results. See Figure 219.

As shown in Figure 4, two of the GBS strain isolates (COH 1 and A909) do not appear to contain homologues to the surface proteins GBS 59 and GBS 67. For these two strains, the percentages shown in Figure 4 are amino acid identity to the COH1 protein). Notwithstanding the difference in the surface protein lengths for these two strains, AI-2 within these sequences still contains two sortase proteins and three LPXTG containing surface proteins, as well as a signal peptidase sequence leading into the first surface protein. One of the surface proteins in this variant of AI-2, spb1, has previously been identified as a potential adhesion protein. (See Adderson et al., Infection and Immunity (2003) 71(12):6857 – 6863). Alternatively, because of the lack of GBS 59

and GBS 67 sequences, this variant of AI-2 may be a third type of AI (Adhesin Island-3, AI-3, or GBS AI-3).

More than one AI surface protein may be present in the oligomeric, pilus-like structures of the invention. For example, GBS 59 and GBS 67 may be incorporated into an oligomeric structure. Alternatively, GBS 59 and GBS 150 may be incorporated into an oligomeric structure, or GBS 59, GBS 150 and GBS 67 may be incorporated into an oligomeric structure.

In another embodiment, the invention includes compositions comprising two or more AI surface proteins. The composition may include surface proteins from the same adhesin island. For example, the composition may include two or more GBS AI-2 surface proteins, such as GBS 59, GBS 67 and GBS 150. The surface proteins may be isolated from Gram positve bacteria or they may be produced recombinantly.

GAS Adhesin Islands

As discussed above, Applicants have identified at least four different GAS Adhesin Islands. These adhesion islands are thought to encode surface proteins which are important in the bacteria's virulence, and Applicants have obtained the first electron micrographs revealing the presence of these adhesin island proteins in hyperoligomeric pilus structures on the surface of Group A Streptococcus.

Group A Streptococcus is a human specific pathogen which causes a wide variety of diseases ranging from pharyngitis and impetigo through life threatening invasive disease and necrotizing fascilitis. In addition, post-streptococcal autoimmune responses are still a major cause of cardiac pathology in children.

Group A Streptococcal infection of its human host can generally occur in three phases. The first phase involves attachment and/or invasion of the bacteria into host tissue and multiplication of the bacteria within the extracellular spaces. Generally this attachment phase begins in the throat or the skin. The deeper the tissue level infected, the more severe the damage that can be caused. In the second stage of infection, the bacteria secretes a soluble toxin that diffuses into the surrounding tissue or even systemically through the vasculature. This toxin binds to susceptible host cell receptors and triggers innappropropriate immune responses by these host cells, resulting in pathology. Because the toxin can diffuse throughout the host, the necrosis directly caused by the GAS toxins may be physically located in sites distant from the bacterial infection. The final phase of GAS infection can occur long after the original bacteria have been cleared from the host system. At this stage, the host's previous immune response to the GAS bacteria due to cross reactivity between epitopes of a GAS surface protein, M, and host tissues, such as the heart. A general review of GAS infection can be found in Principles of Bacterial Pathogeneis, Groisman ed., Chapter 15 (2001).

In order to prevent the pathogenic effects associated with the later stages of GAS infection, an effective vaccine against GAS will preferably facilitate host elimination of the bacteria during the initial attachment and invasion stage.

PC Isolates of Group A Streptococcus are historically classified according to the M surface protein described above. The M protein is surface exposed trypsin-sensitive protein generally comprising two polypeptide chains complexed in an alpha helical formation. The carboxyl terminus is anchored in the cytoplasmic membrane and is highly conserved among all group A streptococci. The amino terminus, which extend through the cell wall to the cell surface, is responsible for the antigenic variability observed among the 80 or more serotypes of M proteins.

A second layer of classification is based on a variable, trypsin-resistant surface antigen, commonly referred to as the T-antigen. Decades of epidemiology based on M and T serological typing have been central to studies on the biological diversity and disease causing potential of Group A Streptococci. While the M-protein component and its inherent variability have been extensively characterized, even after five decades of study, there is still very little known about the structure and variability of T-antigens. Antisera to define T types is commercially available from several sources, including Sevapharma (http://www.sevapharma.cz/en).

The gene coding for one form of T-antigen, T-type 6, from an M6 strain of GAS (D741) has been cloned and characterized and maps to an approximately 11 kb highly variable pathogenicity island. Schneewind et al., J Bacteriol. (1990) 172(6):3310 – 3317. This island is known as the Fibronectin-binding, Collagen-binding T-antigen (FCT) region because it contains, in addition to the T6 coding gene (*tee6*), members of a family of genes coding for Extra Cellular Matrix (ECM) binding proteins. Bessen et al., Infection & Immunity (2002) 70(3):1159-1167. Several of the protein products of this gene family have been shown to directly bind either fibronectin and/or collagen. *See* Hanski et al., Infection & Immunity (1992) 60(12):5119-5125; Talay et al., Infection & Immunity (1992(60(9):3837-3844; Jaffe et al. (1996) 21(2):373-384; Rocha et al., Adv Exp Med Biol. (1997) 418:737-739; Kreikemeyer et al., J Biol Chem (2004) 279(16):15850-15859; Podbielski et al., Mol. Microbiol. (1999) 31(4):1051-64; and Kreikemeyer et al., Int. J. Med Microbiol (2004) 294(2-3):177-88. In some cases direct evidence for a role of these proteins in adhesion and invasion has been obtained.

Applicants raised antiserum against a recombinant product of the *tee6* gene and used it to explore the expression of T6 in M6 strain 2724. In immunoblot of mutanolysin extracts of this strain, the antiserum recognized, in addition to a band corresponding to the predicted molecular mass of the product, very high molecular weight ladders ranging in mobility from about 100 kDa to beyond the resolution of the 3-8% gradient gels used.

This pattern of high molecular weight products is similar to that observed in immunoblots of the protein components of the pili identified in *Streptococcus agalactiae* (described above) and previously in *Corynebacterium diphtheriae*. Electron microscropy of strain M6_2724 with antisera specific for the product of *tee6* revealed abundant surface staining and long pilus like structures extending up to 700 nanometers from the bacterial surface, revealing that the T6 protein, one of the antigens recognized in the original Lancefiled serotyping system, is located within a GAS Adhesin Island (GAS AI-1) and forms long covalently linked pilus structures.

PCT/US05/27239

Applicants have identified at least four different Group A Streptococcus Adhesin Islands. While these GAS AI sequences can be identified in numerous M types, Applicants have surprisingly discovered a correlation between the four main pilus subunits from the four different GAS AI types and specific T classifications. While other trypsin-resistant surface exposed proteins are likely also implicated in the T classification designations, the discovery of the role of the GAS adhesin islands (and the associated hyper-oligomeric pilus like structures) in T classification and GAS serotype variance has important implications for prevention and treatment of GAS infections. Applicants have identified protein components within each of the GAS adhesin islands which are associated with the pilus formation. These proteins are believed to be involved in the bacteria's initial adherence mechanisms. Immunological recognition of these proteins may allow the host immune response to slow or prevent the bacteria's transition into the more pathogenic later stages of infection.

In addition, Applicants have discovered that the GBS pili structures appear to be implicated in the formation of biofilms (populations of bacteria growing on a surface, often enclosed in an exopolysaccharide matrix). Biofilms are generally associated with bacterial resistance, as antibiotic treatments and host immune response are frequently unable to erradicate all of the bacteria components of the biofilm. Direction of a host immune response against surface proteins exposed during the first steps of bacterial attachment (i.e., before complete biofilm formation) is preferable.

The invention therefore provides for improved immunogenic compositions against GAS infection which may target GAS bacteria during their initial attachment efforts to the host epithelial cells and may provide protection against a wide range of GAS serotypes. The immunogenic compositions of the invention include GAS AI surface proteins which may be formulated in an oligomeric, or hyperoligomeric (pilus) form. The invention also includes combinations of GAS AI surface proteins. Combinations of GAS AI surface proteins may be selected from the same adhesin island or they may be selected from different GAS adhesin islands.

While there is surprising variability in the number and sequence of the GAS AI components across isolates, GAS AI sequences may be generally characterized as Type 1, Type 2, Type 3, and Type 4, depending on the number and type of sortase sequence within the island and the percentage identity of other proteins within the island. Schematics of the GAS adhesin islands are set forth in FIGURE 51A and FIGURE 162. In all strains identified so far, the adhesin island region is flanked by highly conserved open reading frames M1_123 and M1_136. Between three and five genes in each GAS adhesin island code for ECM binding adhesin proteins containing LPXTG motifs.

<u>GAS Adhesin Island 1</u>

As discussed above, Applicants have identified adhesin islands, "GAS Adhesin Island 1" or "GAS AI-1", within the genome Group A Streptococcus serotypes and isolates. GAS AI-1 comprises a series of approximately five open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("GAS AI-1 proteins"). GAS AI-1 preferably comprises surface proteins, a srtB sortase, and a rofA divergently transcribed transcriptional regulator. GAS AI-1 surface proteins may include a fibronectin binding protein, a collagen adhesion protein and a

fimbrial structural subunit. Preferably, each of these GAS AI-1 surface proteins includes an LPXTG sortase substrate motif, such as LPXTG (SEQ ID NO: 122) or LPXSG (SEQ ID NO: 134) (conservative replacement of threonine with serine). Specifically, GAS AI-1 includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4 or 5) of M6_Spy0157, M6_Spy0158, M6_Spy0159, M6_Spy0160, M6_Spy0161.

Applicants have also identified open reading frames encoding fimbrial structural subunits in other GAS bacteria harbouring an AI-1. These open reading frames encode fimbrial structural subunits CDC SS 410_fimbrial, ISS3650_fimbrial, and DSM2071_fimbrial. A GAS AI-1 may comprise a polynucleotide encoding any one of CDC SS 410_fimbrial, ISS3650_fimbrial, and DSM2071_fimbrial.

As discussed above, the hyper-oligomeric pilus structure of GAS AI-1 appears to be responsible for the T-antigen type 6 classification, and GAS AI-1 corresponds to the FCT region previously identified for *tee6*. As in GAS AI-1, the *tee6* FCT region includes open reading frames encoding for a collagen adhesion protein (cpa, capsular polysaccharide adhesion) and a fibronectin binding protein (prtF1). Immunoblots of *tee6*, a GAS AI-1 fimbrial structural subunit corresponding to M6_Spy160, reveal high molecular weight structures indicative of the hyper-oligomeric pilus structures. Immunoblots with antiserum specific for Cpa also recognize a high molecular weight ladder structure, indicating Cpa involvement in the GAS AI-1 pilus structure or formation. In EM photos of GAS bacteria, Cpa antiserum reveals abundant staining on the surface of the bacteria and occasional gold particles extended from the surface of the bacteria. In contrast, immunoblots with antiserum specific for PrtF1 recognize only a single molecular species with electrophoretic mobility corresponding to its predicted molecular mass, indicating that PrtF1 may not be associated with the oligomeric pilus structure. A preferred immunogenic composition of the invention comprises a GAS AI-1 surface protein which may be formulated or purified in an oligomeric (pilis) form. In a preferred embodiment, the oligomeric form is a hyperoligomer. Another preferred immunogenic composition of the invention comprises a GAS AI-1 surface protein which has been isolated in an oligomeric (pilis) form. The oligomer or hyperoligomeric pilus structures comprising the GAS AI-1 surface proteins may be purified or otherwise formulate for use in immunogenic compositions.

One or more of the GAS AI-1 open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the GAS AI-1 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the GAS AI-1 surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif.

The LPXTG sortase substrate motif of a GAS AI surface protein may be generally represented by the formula XXXXG, wherein X at amino acid position 1 is an L, a V, an E, or a Q, wherein X at amino acid position 2 is a P if X at amino acid position 1 is an L, wherein X at amino acid position 2 is a V if X at amino acid position 1 is a E or a Q, wherein X at amino acid position 2 is

a V or a P if X at amino acid position 1 is a V, wherein X at amino acid position 3 is any amino acid residue, wherein X at amino acid position 4 is a T if X at amino acid position 1 is a V, E, or Q, and wherein X at amino acid position 4 is a T, S, or A if X at amino acid position 1 is an L. Some examples of LPXTG motifs present in GAS AI surface proteins include LPSXG (SEQ ID NO: 134), VVXTG (SEQ ID NO: 135), EVXTG (SEQ ID NO: 136), VPXTG (SEQ ID NO: 137), QVXTG (SEQ ID NO: 138), LPXAG (SEQ ID NO: 139), QVPTG (SEQ ID NO: 140), and FPXTG (SEQ ID NO: 141).

The GAS AI surface proteins of the invention may affect the ability of the GAS bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of GAS to translocate through an epithelial cell layer. Preferably, one or more GAS AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. GAS AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The GAS AI-1 sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. GAS AI-1 may encode for at least one surface protein. Alternatively, GAS AI-1 may encode for at least two surface exposed proteins and at least one sortase. Preferably, GAS AI-1 encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

GAS AI-1 preferably includes a srtB sortase. GAS srtB sortases may preferably anchor surface proteins with an LPSTG motif (SEQ ID NO: 166), particularly where the motif is followed by a serine.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a GAS AI-1 surface protein such as M6_Spy0157, M6_Spy0159, M6_Spy0160, CDC SS 410_fimbrial, ISS3650_fimbrial, or DSM2071_fimbrial. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine or serine amino acid residue, respectively.

PCT/US05/27239

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a GAS Adhesin Island protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more GAS Adhesin Island 1 ("GAS AI-1") proteins and one or more GAS Adhesin Island 2 ("GAS AI-2"), GAS Adhesin Island 3 ("GAS AI-3"), or GAS Adhesin Island 4 ("GAS AI-4") proteins, wherein one or more of the GAS Adhesin Island proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the GAS AI-1 proteins, GAS AI-1 may also include a divergently transcribed transcriptional regulator such as *RofA* (*i.e.*, the transcriptional regulator is located near or adjacent to the AI protein open reading frames, but it transcribed in the opposite direction).

GAS Adhesin Island 2

A second adhesin island, "GAS Adhesin Island 2" or "GAS AI-2" has also been identified in Group A Streptococcus serotypes and isolates. GAS AI-2 comprises a series of approximately eight open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("GAS AI-2 proteins"). Specifically, GAS AI-2 includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, 7, or 8) of GAS15, Spy0127, GAS16, GAS17, GAS18, Spy0131, Spy0133, and GAS20.

A preferred immunogenic composition of the invention comprises a GAS AI-2 surface protein which may be formulated or purified in an oligomeric (pilis) form. In a preferred embodiment, the oligomeric form is a hyperoligomer. Another preferred immunogenic composition of the invention comprises a GAS AI-2 surface protein which has been isolated in an oligomeric (pilis) form. The oligomer or hyperoligomeric pilus structures comprising the GAS AI-2 surface proteins may be purified or otherwise formulate for use in immunogenic compositions.

One or more of the GAS AI-2 open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the GAS AI-2 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the GAS AI-2 surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. The AI surface proteins of the invention may affect the ability of the GAS bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of GAS to translocate through an epithelial cell layer. Preferably, one or more AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

PCT/US2005/027239

The GAS AI-2 sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. GAS AI-2 may encode for at least one surface protein. Alternatively, GAS AI-2 may encode for at least two surface exposed proteins and at least one sortase. Preferably, GAS AI-2 encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising an AI surface protein such as GAS15, GAS16, or GAS18. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine amino acid residue.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a GAS Adhesin Island protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more GAS Adhesin Island 2 ("GAS AI-2") proteins and one or more GAS Adhesin Island 1 ("GAS AI-1"), GAS Adhesin Island 3 ("GAS AI-3"), or GAS Adhesin Island 4 ("GAS AI-4") proteins, wherein one or more of the Adhesin Island proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the GAS AI-2 proteins, GAS AI-2 may also include a divergently transcribed transcriptional regulator such as *rofA* (*i.e.*, the transcriptional regulator is located near or adjacent to the AI protein open reading frames, but it transcribed in the opposite direction).

GAS Adhesin Island 3

A third adhesin island, "GAS Adhesin Island 3" or "GAS AI-3" has also been identified in several Group A Streptococcus serotypes and isolates. GAS AI-3 comprises a series of approximately

PCT/US05/27239

seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("GAS AI-3 proteins"). Specifically, GAS AI-3 includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of SpyM3_0098, SpyM3_0099, SpyM3_0100, SpyM3_0101, SpyM3_0102, SpyM3_0103, SpyM3_0104, SPs0100, SPs0101, SPs0102, SPs0103, SPs0104, SPs0105, SPs0106, orf78, orf79, orf80, orf81, orf82, orf83, orf84, spyM18_0126, spyM18_0127, spyM18_0128, spyM18_0129, spyM18_0130, spyM18_0131, spyM18_0132, SpyoM01000156, SpyoM01000155, SpyoM01000154, SpyoM01000153, SpyoM01000152, SpyoM01000151, SpyoM01000150, and SpyoM01000149. In one embodiment, GAS AI-3 includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of SpyM3_0098, SpyM3_0099, SpyM3_0100, SpyM3_0101, SpyM3_0102, SpyM3_0103, and SpyM3_0104. In another embodiment, GAS AI-3 includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of SPs0100, SPs0101, SPs0102, SPs0103, SPs0104, SPs0105, and SPs0106. In a further embodiment, GAS AI-3 includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of orf78, orf79, orf80, orf81, orf82, orf83, and orf84. In yet another embodiment, GAS AI-3 includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of spyM18_0126, spyM18_0127, spyM18_0128, spyM18_0129, spyM18_0130, spyM18_0131, and spyM18_0132. In yet another embodiment, GAS AI-3 includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of SpyoM01000156, SpyoM01000155, SpyoM01000154, SpyoM01000153, SpyoM01000152, SpyoM01000151, SpyoM01000150, and SpyoM01000149.

Applicants have also identified open reading frames encoding fimbrial structural subunits in other GAS bacteria harbouring an AI-3. These open reading frames encode fimbrial structural subunits ISS3040_fimbrial, ISS3776_fimbrial, and ISS4959_fimbrial. A GAS AI-3 may comprise a polynucleotide encoding any one of ISS3040_fimbrial, ISS3776_fimbrial, and ISS4959_fimbrial.

One or more of the GAS AI-3 open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the GAS AI-3 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

A preferred immunogenic composition of the invention comprises a GAS AI-3 surface protein which may be formulated or purified in an oligomeric (pilis) form. In a preferred embodiment, the oligomeric form is a hyperoligomer. Another preferred immunogenic composition of the invention comprises a GAS AI-3 surface protein which has been isolated in an oligomeric (pilis) form. The oligomer or hyperoligomeric pilus structures comprising the GAS AI-3 surface proteins may be purified or otherwise formulate for use in immunogenic compositions.

One or more of the GAS AI-3 surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. The AI surface proteins of the invention may affect the ability of the GAS bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of GAS to translocate through an epithelial cell layer. Preferably, one or more AI surface proteins are capable of binding to or otherwise associating with an

PCT/US05/27239

epithelial cell surface. AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The GAS AI-3 sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. GAS AI-3 may encode for at least one surface protein. Alternatively, GAS AI-3 may encode for at least two surface exposed proteins and at least one sortase. Preferably, GAS AI-3 encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine or alanine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

The invention includes a composition comprising oligomeric, pilus-like structures comprising an AI surface protein such as SpyM3_0098, SpyM3_0100, SpyM3_0102, SpyM3_0104, SPs0100, SPs0102, SPs0104, SPs0106, orf78, orf80, orf82, orf84, spyM18_0126, spyM18_0128, spyM18_0130, spyM18_0132, SpyoM01000155, SpyoM01000153, SpyoM01000151, SpyoM01000149, ISS3040_fimbrial, ISS3776_fimbrial, and ISS4959_fimbrial. In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising an AI surface protein such as SpyM3_0098, SpyM3_0100, SpyM3_0102, and SpyM3_0104. In another embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising an AI surface protein such as SPs0100, SPs0102, SPs0104, and SPs0106. In another embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising an AI surface protein such as orf78, orf80, orf82, and orf84. In yet another embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising an AI surface protein such as spyM18_0126, spyM18_0128, spyM18_0130, and spyM18_0132. In a further embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising an AI surface protein such as SpyoM01000155, SpyoM01000153, SpyoM01000151, and SpyoM01000149. In yet a further embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising an AI surface protein such as ISS3040_fimbrial, ISS3776_fimbrial, and ISS4959_fimbrial. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric

PCT/US05/27239

subunits may be covalently associated via an LPXTG motif, preferably, via the threonine amino acid residue.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a GAS Adhesin Island protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more GAS Adhesin Island 3 ("GAS AI-3") proteins and one or more GAS Adhesin Island 1 ("GAS AI-1"), GAS Adhesin Island 2 ("GAS AI-2"), or GAS Adhesin Island 4 ("GAS AI-4") proteins, wherein one or more of the Adhesin Island proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the GAS AI-3 proteins, GAS AI-3 may also include a transcriptional regulator such as *Nra*.

GAS Adhesin Island 4

A fourth adhesin island, "GAS Adhesin Island 4" or "GAS AI-4" has also been identified in Group A Streptococcus serotypes and isolates. GAS AI-4 comprises a series of approximately eight open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases ("GAS AI-4 proteins"). Specifically, GAS AI-4 includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, 7, or 8) of 19224134, 19224135, 19223136, 19223137, 19224138, 19224139, 19224140, and 19224141.

Applicants have also identified open reading frames encoding fimbrial structural subunits in other GAS bacteria harbouring an AI-4. These open reading frames encode fimbrial structural subunits 20010296_fimbrial, 20020069_fimbrial, CDC SS 635_fimbrial, ISS4883_fimbrial, and ISS4538_fimbrial. A GAS AI-4 may comprise a polynucleotide encoding any one of 20010296_fimbrial, 20020069_fimbrial, CDC SS 635_fimbrial, ISS4883_fimbrial, and ISS4538_fimbrial.

One or more of the GAS AI-4 open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the GAS AI-4 open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

A preferred immunogenic composition of the invention comprises a GAS AI-4 surface protein which may be formulated or purified in an oligomeric (pilis) form. In a preferred embodiment, the oligomeric form is a hyperoligomer. Another preferred immunogenic composition of the invention comprises a GAS AI-4 surface protein which has been isolated in an oligomeric (pilis) form. The oligomer or hyperoligomeric pilus structures comprising the GAS AI-4 surface proteins may be purified or otherwise formulate for use in immunogenic compositions.

One or more of the GAS AI-4 surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. The AI surface proteins of the

PCT/US2005/027239

invention may effect the ability of the GAS bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of GAS to translocate through an epithelial cell layer. Preferably, one or more AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The GAS AI-4 sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. GAS AI-4 may encode for at least one surface protein. Alternatively, GAS AI-4 may encode for at least two surface exposed proteins and at least one sortase. Preferably, GAS AI-4 encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising an AI surface protein such as 19224134, 19224135, 19224137, 19224139, 19224141, 20010296_fimbrial, 20020069_fimbrial, CDC SS 635_fimbrial, ISS4883_fimbrial, and ISS4538_fimbrial. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine amino acid residue.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a GAS Adhesin Island protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more GAS Adhesin Island 4 ("GAS AI-4") proteins and one or more GAS Adhesin Island 1 ("GAS AI-1"), GAS Adhesin Island 2 ("GAS AI-2"), or GAS Adhesin Island 3 ("GAS AI-3") proteins, wherein one or more of the Adhesin Island proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

PCT/US05/27239

In addition to the open reading frames encoding the GAS AI-4 proteins, GAS AI-4 may also include a divergently transcribed transcriptional regulator such as *rofA* (*i.e.*, the transcriptional regulator is located near or adjacent to the AI protein open reading frames, but it transcribed in the opposite direction).

The oligomeric, pilus-like structures of the invention may be combined with one or more additional GAS proteins. In one embodiment, the oligomeric, pilus-like structures comprise one or more AI surface proteins in combination with a second GAS protein.

The oligomeric, pilus-like structures may be isolated or purified from bacterial cultures in which the bacteria express an AI surface protein. The invention therefore includes a method for manufacturing an oligomeric AI surface antigen comprising culturing a GAS bacterium that expresses the oligomeric AI protein and isolating the expressed oligomeric AI protein from the GAS bacteria. The AI protein may be collected from secretions into the supernatant or it may be purified from the bacterial surface. The method may further comprise purification of the expressed AI protein. Preferably, the AI protein is in a hyperoligomeric form.

The oligomeric, pilus-like structures may be isolated or purified from bacterial cultures overexpressing an AI surface protein. The invention therefore includes a method for manufacturing an oligomeric Adhesin Island surface antigen comprising culturing a GAS bacterium adapted for increased AI protein expression and isolation of the expressed oligomeric Adhesin Island protein from the GAS bacteria. The AI protein may be collected from secretions into the supernatant or it may be purified from the bacterial surface. The method may further comprise purification of the expressed Adhesin Island protein. Preferably, the Adhesin Island protein is in a hyperoligomeric form.

The GAS bacteria are preferably adapted to increase AI protein expression by at least two (*e.g.*, 2, 3, 4, 5, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 125, 150 or 200) times wild type expression levels.

GAS bacteria may be adapted to increase AI protein expression by any means known in the art, including methods of increasing gene dosage and methods of gene upregulation. Such means include, for example, transformation of the GAS bacteria with a plasmid encoding the AI protein. The plasmid may include a strong promoter or it may include multiple copies of the sequence encoding the AI protein. Optionally, the sequence encoding the AI protein within the GAS bacterial genome may be deleted. Alternatively, or in addition, the promoter regulating the GAS Adhesin Island may be modified to increase expression.

The invention further includes GAS bacteria which have been adapted to produce increased levels of AI surface protein. In particular, the invention includes GAS bacteria which have been adapted to produce oligomeric or hyperoligomeric AI surface protein. In one embodiment, the Gram positive bacteria of the invention are inactivated or attenuated to permit *in vivo* delivery of the whole bacteria, with the AI surface protein exposed on its surface.

The invention further includes GAS bacteria which have been adapted to have increased levels of expressed AI protein incorporated in pili on their surface. The GAS bacteria may be adapted

PCT/US2005/027239

to have increased exposure of oligomeric or hyperoligomeric AI proteins on its surface by increasing expression levels of LepA polypeptide, or an equivalent signal peptidase, in the GAS bacteria. Applicants have shown that deletion of LepA in strain SF370 bacteria, which harbour a GAS AI-2, abolishes surface exposure of M and pili proteins on the GAS. Increased levels of LepA expression in GAS are expected to result in increased exposure of M and pili proteins on the surface of GAS. Increased expression of LepA in GAS may be achieved by any means known in the art, such as increasing gene dosage and methods of gene upregulation. The GAS bacteria adapted to have increased levels of LepA expression may additionally be adapted to express increased levels of at least one pili protein.

Alternatively, the AI proteins of the invention may be expressed on the surface of a non-pathogenic Gram positive bacteria, such as *Streptococus gordonii* (See, e.g., Byrd et al., "Biological consequences of antigen and cytokine co-expression by recombinant *Streptococcus gordonii* vaccine vectors", Vaccine (2002) 20:2197-2205) or *Lactococcus lactis* (See, e.g., Mannam et al., "Mucosal Vaccine Made from Live, Recombinant Lactococcus lactis Protects Mice against Pharangeal Infection with *Streptococcus pyogenes*" Infection and Immunity (2004) 72(6):3444-3450). As used herein, non-pathogenic Gram positive bacteria refer to Gram positive bacteria which are compatible with a human host subject and are not associated with human pathogenisis. Preferably, the non-pathogenic bacteria are modified to express the AI surface protein in oligomeric, or hyper-oligomeric form. Sequences encoding for an AI surface protein and, optionally, an AI sortase, may be integrated into the non-pathogenic Gram positive bacterial genome or inserted into a plasmid. The non-pathogenic Gram positive bacteria may be inactivated or attenuated to facilitate *in vivo* delivery of the whole bacteria, with the AI surface protein exposed on its surface. Alternatively, the AI surface protein may be isolated or purified from a bacterial culture of the non-pathogenic Gram positive bacteria. For example, the AI surface protein may be isolated from cell extracts or culture supernatants. Alternatively, the AI surface protein may be isolated or purified from the surface of the non-pathogenic Gram positive bacteria.

The non-pathogenic Gram positive bacteria may be used to express any of the GAS Adhesin Island proteins described herein. The non-pathogenic Gram positive bacteria are transformed to express an Adhesin Island surface protein. Preferably, the non-pathogenic Gram positive bacteria also express at least one Adhesin Island sortase. The AI transformed non-pathogenic Gram positive bacteria of the invention may be used to prevent or treat infection with pathogenic GAS.

Applicants modified *L. lactis* to demonstrate that, like GBS polypeptides, it can express GAS AI polypeptides. *L. lactis* was transformed with pAM401 constructs encoding entire pili gene clusters of AI-1, AI-2, and AI-4 adhesin islands. Briefly, the pAM401 is a promoterless high-copy plasmid. The entire pili gene clusters of an M6 (AI-1), M1 (AI-2), and M12 (AI-4) bacteria were inserted into the pAM401 construct. The gene clusters were transcribed under the control their own (M6, M1, or M12) promoter or the GBS promoter that successfully initiated expression of the GBS AI-1 adhesin islands in *L. lactis*, described above. Figure 172 provides a schematic depiction of GAS M6 (AI-1),

PCT/US05/27239

M1 (AI-2), and M12 (AI-4) adhesin islands and indicates the portions of the adhesin island sequences inserted in the pAM401 construct.

Each of the *L. lactis* transformed with one of the M6, M1, or M12 adhesin island gene clusters expressed high molecular weight structures that were immunoreactive with antibodies that bind to polypeptides present in their respective pili. Figures 173 A-C provide results of Western blot analysis of surface protein-enriched extracts of *L. lactis* transformed with M6 (Figure 173 A), M1 (Figure 173 B), or M12 (Figure 173 C) adhesin island gene clusters using antibodies that bind to the fimbrial structural subunit encoded by each cluster. Figure 173A at lanes 3 and 4 shows detection of high molecular structures in *L. lactis* transformed with an adhesin island pilus gene cluster from an M1 AI-2 using an antibody that binds to fimbrial structural subunit Spy0128. Figure 173B at lanes 3 and 4 shows detection of high molecular weight structures in *L. lactis* transformed with an adhesin island pilus gene cluster from an M12 AI-4 using an antibody that binds to fimbrial structural subunit EftLSL.A. Figure 173C at lane 3 shows detection of high molecular weight structures in *L. lactis* transformed with an adhesin island pilus gene cluster from an M6 AI-1 using an antibody that binds to fimbrial structural subunit M6_Spy0160. In figures 173 A-C, "p1" immediately following the notation of AI subtype indicates that the promoter present in the Adhesin Island is used to drive transcription of the adhesin island gene cluster and "p2" indicates that the promoter was the GBS promoter described above. Thus, it appears that *L. lactis* is capable of expressing the fimbrial structural subunits encoded by GAS adhesin islands in an oligomeric form.

Alternatively, the oligomeric, pilus-like structures may be produced recombinantly. If produced in a recombinant host cell system, the AI surface protein will preferably be expressed in coordination with the expression of one or more of the AI sortases of the invention. Such AI sortases will facilitate oligomeric or hyperoligomeric formation of the AI surface protein subunits.

S. pneumoniae from TIGR4 Adhesin Island

As discussed above, Applicants have identified adhesin islands within the genome of *S. pneumoniae* from TIGR4. The *S. pneumoniae* from TIGR4 Adhesin Island comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, the *S. pneumoniae* from TIGR4 AI proteins includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of SP0462, SP0463, SP0464, SP0465, SP0466, SP0467, and SP0468.

A preferred immunogenic composition of the invention comprises a *S. pneumoniae* from TIGR4 AI surface protein which may be formulated or purified in an oligomeric (pilis) form. In a preferred embodiment, the oligomeric form is a hyperoligomer. Another preferred immunogenic composition of the invention comprises a *S. pneumoniae* from TIGR4 AI surface protein which has been isolated in an oligomeric (pilis) form. The oligomer or hyperoligomer pilus structures comprising *S. pneumoniae* surface proteins may be purified or otherwise formulated for use in immunogenic compositions.

PCT/US05/27239

One or more of the *S. pneumoniae* from TIGR4 AI open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* from TIGR4 AI open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the *S. pneumoniae* from TIGR4 AI surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif.

The *S. pneumoniae* from TIGR4 AI surface proteins of the invention may affect the ability of the *S. pneumoniae* bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of *S. pneumoniae* to translocate through an epithelial cell layer. Preferably, one or more *S. pneumoniae* from TIGR4 AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. *S. pneumoniae* from TIGR4 AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The *S. pneumoniae* from TIGR4 AI sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. *S. pneumoniae* from TIGR4 AI may encode for at least one surface protein. Alternatively, *S. pneumoniae* from TIGR4 AI may encode for at least two surface exposed proteins and at least one sortase. Preferably, *S. pneumoniae* from TIGR4 AI encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a *S. pneumoniae* from TIGR4 AI surface protein such as SP0462, SP0463, SP0464, or SP0465. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine or serine amino acid residue, respectively.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

PCT/US05/27239

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a *S. pneumoniae* from TIGR4 AI protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more *S. pneumoniae* from TIGR4 AI proteins and one or more *S. pneumoniae* strain 670 AI proteins, wherein one or more of the *S. pneumoniae* AI proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the *S. pneumoniae* from TIGR4 AI proteins, *S. pneumoniae* from TIGR4 AI may also include a transcriptional regulator.

S. pneumoniae strain 670 Adhesin Island

As discussed above, Applicants have identified adhesin islands within the genome of *S. pneumoniae* strain 670. The *S. pneumoniae* strain 670 Adhesin Island comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, the *S. pneumoniae* strain 670 AI proteins includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of orf1_670, orf3_670, orf4_670, orf5_670, orf6_670, orf7_670, orf8_670.

A preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 670 AI surface protein which may be formulated or purified in an oligomeric (pilis) form. Another preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 670 AI surface protein which has been isolated in an oligomeric (pilis) form.

One or more of the *S. pneumoniae* strain 670 AI open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* strain 670 AI open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the *S. pneumoniae* strain 670 AI surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif.

The *S. pneumoniae* strain 670 AI surface proteins of the invention may affect the ability of the *S. pneumoniae* bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of *S. pneumoniae* to translocate through an epithelial cell layer. Preferably, one or more *S. pneumoniae* strain 670 AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. *S. pneumoniae* strain 670 AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The *S. pneumoniae* strain 670 AI sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. *S. pneumoniae* strain 670 AI may encode for at least one surface protein. Alternatively, *S. pneumoniae* strain 670 AI may encode for at least two surface exposed proteins and at least one sortase. Preferably, *S. pneumoniae* strain 670 AI encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface

PCT/US2005/027239

protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a *S. pneumoniae* strain 670 AI surface protein such as orf3_670, orf4_670, or orf5_670. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine or serine amino acid residue, respectively.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a *S. pneumoniae* strain 670 AI protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more *S. pneumoniae* strain 670 AI proteins and one or more *S. pneumoniae* from TIGR4 AI proteins, wherein one or more of the *S. pneumoniae* AI proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the *S. pneumoniae* strain 670 AI proteins, *S. pneumoniae* strain 670 AI may also include a transcriptional regulator.

<u>*S. pneumoniae* strain 14 CSR 10 Adhesin Island</u>

As discussed above, Applicants have identified adhesin islands within the genome of *S. pneumoniae* strain 14 CSR 10. The *S. pneumoniae* strain 14 CSR 10 Adhesin Island comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, the *S. pneumoniae* strain 14 CSR 10 AI proteins includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of ORF2_14CSR, ORF3_14CSR, ORF4_14CSR, ORF5_14CSR, ORF6_14CSR, ORF7_14CSR, ORF8_14CSR.

A preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 14 CSR 10 AI surface protein which may be formulated or purified in an oligomeric (pilis) form. Another preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 14 CSR 10 AI surface protein which has been isolated in an oligomeric (pilis) form.

PCT/US2005/027239

One or more of the *S. pneumoniae* strain 14 CSR 10 AI open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* strain 14 CSR 10 AI open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the *S. pneumoniae* strain 14 CSR 10 AI surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif.

The *S. pneumoniae* strain 14 CSR 10 AI surface proteins of the invention may affect the ability of the *S. pneumoniae* bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of *S. pneumoniae* to translocate through an epithelial cell layer. Preferably, one or more *S. pneumoniae* strain 14 CSR 10 AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. *S. pneumoniae* strain 14 CSR 10 AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The *S. pneumoniae* strain 14 CSR 10 AI sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. *S. pneumoniae* strain 14 CSR 10 AI may encode for at least one surface protein. Alternatively, *S. pneumoniae* strain 14 CSR 10 AI may encode for at least two surface exposed proteins and at least one sortase. Preferably, *S. pneumoniae* strain 14 CSR 10 AI encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a *S. pneumoniae* strain 14 CSR 10 AI surface protein such as orf3_CSR, orf4_CSR, or orf5_CSR. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine or serine amino acid residue, respectively.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

PCT/US2005/027239

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a *S. pneumoniae* strain 14 CSR 10 AI protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more *S. pneumoniae* strain 14 CSR 10 AI proteins, and one or more AI proteins of any of *S. pneumoniae* from TIGR4, 670, 19A Hungary 6, 6B Finland 12, 6B Spain 2, 9V Spain 3, 19F Taiwan 14, 23F Taiwan 15, or 23F Poland 16, wherein one or more of the *S. pneumoniae* AI proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the *S. pneumoniae* strain 14 CSR 10AI proteins, *S. pneumoniae* strain 14 CSR 10 AI may also include a transcriptional regulator.

<u>*S. pneumoniae* strain 19A Hungary 6 Adhesin Island</u>

As discussed above, Applicants have identified adhesin islands within the genome of *S. pneumoniae* strain 19A Hungary 6. The *S. pneumoniae* strain 19A Hungary 6 Adhesin Island comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, the *S. pneumoniae* strain 19A Hungary 6 AI proteins includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of ORF2_19AH, ORF3_19AH, ORF4_19AH, ORF5_19AH, ORF6_19AH, ORF7_19AH, ORF8_19AH.

A preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 19A Hungary 6 AI surface protein which may be formulated or purified in an oligomeric (pilis) form. Another preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 19A Hungary 6 AI surface protein which has been isolated in an oligomeric (pilis) form.

One or more of the *S. pneumoniae* strain 19A Hungary 6 AI open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* strain 19A Hungary 6 AI open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the *S. pneumoniae* strain 19A Hungary 6 AI surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif.

The *S. pneumoniae* strain 19A Hungary 6 AI surface proteins of the invention may affect the ability of the *S. pneumoniae* bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of *S. pneumoniae* to translocate through an epithelial cell layer. Preferably, one or more *S. pneumoniae* strain 19A Hungary 6 AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. *S. pneumoniae* strain 19A Hungary 6 AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The *S. pneumoniae* strain 19A Hungary 6 AI sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. *S. pneumoniae* strain 19A Hungary 6 AI may encode for at least one surface protein. Alternatively, *S. pneumoniae* strain 19A Hungary 6 AI may encode for at least two surface exposed proteins and at least one sortase.

Preferably, *S. pneumoniae* strain 19A Hungary 6 AI encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a *S. pneumoniae* strain 19A Hungary 6 AI surface protein such as orf3_19AH, orf4_19AH, or orf5_19AH. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine or serine amino acid residue, respectively.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a *S. pneumoniae* strain 19A Hungary 6 AI protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more *S. pneumoniae* strain 19A Hungary 6 AI proteins and one or more AI proteins from one of any one of *S. pneumoniae* from TIGR4, 670, 14 CSR 10, 6B Finland 12, 6B Spain 2, 9V Spain 3, 19F Taiwan 14, 23F Taiwan 15, or 23F Poland 16 AI GR4 AI proteins, wherein one or more of the *S. pneumoniae* AI proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the *S. pneumoniae* strain 19A Hungary 6 AI proteins, *S. pneumoniae* strain 19A Hungary 6 AI may also include a transcriptional regulator.

<u>*S. pneumoniae* strain 19F Taiwan 14 Adhesin Island</u>

As discussed above, Applicants have identified adhesin islands within the genome of *S. pneumoniae* strain 19F Taiwan 14. The *S. pneumoniae* strain 19F Taiwan 14 Adhesin Island comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, the *S. pneumoniae* strain 19F Taiwan 14 AI proteins includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7)

of ORF2_19FTW, ORF3_19FTW, ORF4_19FTW, ORF5_19FTW, ORF6_19FTW, ORF7_19FTW, ORF8_19FTW.

A preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 19F Taiwan 14 AI surface protein which may be formulated or purified in an oligomeric (pilis) form. Another preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 19F Taiwan 14 AI surface protein which has been isolated in an oligomeric (pilis) form.

One or more of the *S. pneumoniae* strain 19F Taiwan 14 AI open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* strain 19F Taiwan 14 AI open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the *S. pneumoniae* strain 19F Taiwan 14 AI surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif.

The *S. pneumoniae* strain 19F Taiwan 14 AI surface proteins of the invention may affect the ability of the *S. pneumoniae* bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of *S. pneumoniae* to translocate through an epithelial cell layer. Preferably, one or more *S. pneumoniae* strain 19F Taiwan 14 AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. *S. pneumoniae* strain 19F Taiwan 14 AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The *S. pneumoniae* strain 19F Taiwan 14 AI sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. *S. pneumoniae* strain 19F Taiwan 14 AI may encode for at least one surface protein. Alternatively, *S. pneumoniae* strain 19F Taiwan 14 AI may encode for at least two surface exposed proteins and at least one sortase. Preferably, *S. pneumoniae* strain 19F Taiwan 14 AI encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a *S. pneumoniae* strain 19F Taiwan 14 AI surface protein such as orf3_19FTW, orf4_19FTW, or orf5_19FTW. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30,

35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine or serine amino acid residue, respectively.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a *S. pneumoniae* strain 19F Taiwan 14 AI protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more *S. pneumoniae* strain 19F Taiwan 14 AI proteins and one or more AI proteins of any one or more of *S. pneumoniae* from TIGR4, 670, 19A Hungary 6, 6B Finland 12, 6B Spain 2, 9V Spain 3, 14 CSR 10, 23F Taiwan 15, or 23F Poland 16, wherein one or more of the *S. pneumoniae* AI proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the *S. pneumoniae* strain 19F Taiwan 14 AI proteins, *S. pneumoniae* strain 19F Taiwan 14 AI may also include a transcriptional regulator.

S. pneumoniae strain 23F Poland 16 Adhesin Island

As discussed above, Applicants have identified adhesin islands within the genome of *S. pneumoniae* strain 23F Poland 16. The *S. pneumoniae* strain 23F Poland 16 Adhesin Island comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, the *S. pneumoniae* strain 23F Poland 16 AI proteins includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of ORF2_23FP, ORF3_23FP, ORF4_23FP, ORF5_23FP, ORF6_23FP, ORF7_23FP, and ORF8_23FP.

A preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 23F Poland 16 AI surface protein which may be formulated or purified in an oligomeric (pilis) form. Another preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 23F Poland 16 AI surface protein which has been isolated in an oligomeric (pilis) form.

One or more of the *S. pneumoniae* strain 23F Poland 16 AI open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* strain 23F Poland 16 AI open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the *S. pneumoniae* strain 23F Poland 16 AI surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif.

The *S. pneumoniae* strain 23F Poland 16 AI surface proteins of the invention may affect the ability of the *S. pneumoniae* bacteria to adhere to and invade epithelial cells. AI surface proteins may

also affect the ability of *S. pneumoniae* to translocate through an epithelial cell layer. Preferably, one or more *S. pneumoniae* strain 23F Poland 16 AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. *S. pneumoniae* strain 23F Poland 16 AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The *S. pneumoniae* strain 23F Poland 16 AI sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. *S. pneumoniae* strain 23F Poland 16 AI may encode for at least one surface protein. Alternatively, *S. pneumoniae* strain 23F Poland 16 AI may encode for at least two surface exposed proteins and at least one sortase. Preferably, *S. pneumoniae* strain 23F Poland 16 AI encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a *S. pneumoniae* strain 23F Poland 16 AI surface protein such as orf3_23FP, orf4_23FP, or orf5_23FP. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine or serine amino acid residue, respectively.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a *S. pneumoniae* strain 23F Poland 16 AI protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more *S. pneumoniae* strain 23F Poland 16 AI proteins and one or more AI proteins from any one or more *S. pneumoniae* strains of TIGR4, 670, 19A Hungary 6, 6B Finland 12, 6B Spain 2, 9V Spain 3, 19F Taiwan 14, 23F Taiwan 15, or 14 CSR 10, wherein one or more of the *S. pneumoniae* AI proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the *S. pneumoniae* strain 23F Poland 16 AI proteins, *S. pneumoniae* strain 23F Poland 16 AI may also include a transcriptional regulator.

S. pneumoniae strain 23F Taiwan 15 Adhesin Island

As discussed above, Applicants have identified adhesin islands within the genome of *S. pneumoniae* strain 23F Taiwan 15. The *S. pneumoniae* strain 23F Taiwan 15 Adhesin Island comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, the *S. pneumoniae* strain 23F Taiwan 15 AI proteins includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of ORF2_23FTW, ORF3_23FTW, ORF4_23FTW, ORF5_23FTW, ORF6_23FTW, ORF7_23FTW, ORF8_23FTW.

A preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 23F Taiwan 15 AI surface protein which may be formulated or purified in an oligomeric (pilis) form. Another preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 23F Taiwan 15 AI surface protein which has been isolated in an oligomeric (pilis) form.

One or more of the *S. pneumoniae* strain 23F Taiwan 15 AI open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* strain 23F Taiwan 15 AI open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the *S. pneumoniae* strain 23F Taiwan 15 AI surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif.

The *S. pneumoniae* strain 23F Taiwan 15 AI surface proteins of the invention may affect the ability of the *S. pneumoniae* bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of *S. pneumoniae* to translocate through an epithelial cell layer. Preferably, one or more *S. pneumoniae* strain 23F Taiwan 15 AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. *S. pneumoniae* strain 23F Taiwan 15 AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The *S. pneumoniae* strain 23F Taiwan 15 AI sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. *S. pneumoniae* strain 23F Taiwan 15 AI may encode for at least one surface protein. Alternatively, *S. pneumoniae* strain 23F Taiwan 15 AI may encode for at least two surface exposed proteins and at least one sortase. Preferably, *S. pneumoniae* strain 23F Taiwan 15 AI encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the

transglycosylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a *S. pneumoniae* strain 23F Taiwan 15 AI surface protein such as orf3_23FTW, orf4_23FTW, or orf5_23FTW. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine or serine amino acid residue, respectively.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a *S. pneumoniae* strain 23F Taiwan 15 AI protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more *S. pneumoniae* strain 23F Taiwan 15 AI proteins and one or more AI proteins from any one or more of *S. pneumoniae* from TIGR4, 670, 19A Hungary 6, 6B Finland 12, 6B Spain 2, 9V Spain 3, 19F Taiwan 14, 14 CSR 10, or 23F Poland 16 AI, wherein one or more of the *S. pneumoniae* AI proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the *S. pneumoniae* strain 23F Taiwan 15 AI proteins, *S. pneumoniae* strain 23F Taiwan 15 AI may also include a transcriptional regulator.

<u>*S. pneumoniae* strain 6B Finland 12 Adhesin Island</u>

As discussed above, Applicants have identified adhesin islands within the genome of *S. pneumoniae* strain 6B Finland 12. The *S. pneumoniae* strain 6B Finland 12 Adhesin Island comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, the *S. pneumoniae* strain 6B Finland 12 AI proteins includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of ORF2_6BF, ORF3_6BF, ORF4_6BF, ORF5_6BF, ORF6_6BF, ORF7_6BF, ORF8_6BF.

A preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 6B Finland 12 AI surface protein which may be formulated or purified in an oligomeric (pilis) form. Another preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 6B Finland 12 AI surface protein which has been isolated in an oligomeric (pilis) form.

One or more of the *S. pneumoniae* strain 6B Finland 12 AI open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF.

Alternatively one or more of the *S. pneumoniae* strain 6B Finland 12 AI open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the *S. pneumoniae* strain 6B Finland 12 AI surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif.

The *S. pneumoniae* strain 6B Finland 12 AI surface proteins of the invention may affect the ability of the *S. pneumoniae* bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of *S. pneumoniae* to translocate through an epithelial cell layer. Preferably, one or more *S. pneumoniae* strain 6B Finland 12 AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. *S. pneumoniae* strain 6B Finland 12 AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The *S. pneumoniae* strain 6B Finland 12 AI sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. *S. pneumoniae* strain 6B Finland 12 AI may encode for at least one surface protein. Alternatively, *S. pneumoniae* strain 6B Finland 12 AI may encode for at least two surface exposed proteins and at least one sortase. Preferably, *S. pneumoniae* strain 6B Finland 12 AI encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a *S. pneumoniae* strain 6B Finland 12 AI surface protein such as orf3_6BF, orf4_6BF, or orf5_6BF. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine or serine amino acid residue, respectively.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric pilus-like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a *S. pneumoniae* strain 6B Finland 12 AI protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more *S. pneumoniae* strain 6B Finland 12 AI proteins and one or more AI proteins of any one or more of *S. pneumoniae* from TIGR4, 670, 19A Hungary 6, 6B Finland 12, 6B Spain 2, 9V Spain 3, 19F Taiwan 14, 23F Taiwan 15, or 23F Poland 16 AI, wherein one or more of the *S. pneumoniae* AI proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the *S. pneumoniae* strain 6B Finland 12 AI proteins, *S. pneumoniae* strain 6B Finland 12 AI may also include a transcriptional regulator.

S. pneumoniae strain 6B Spain 2 Adhesin Island

As discussed above, Applicants have identified adhesin islands within the genome of *S. pneumoniae* strain 6B Spain 2. The *S. pneumoniae* strain 6B Spain 2 Adhesin Island comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, the *S. pneumoniae* strain 6B Spain 2 AI proteins includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of ORF2_6BSP, ORF3_6BSP, ORF4_6BSP, ORF5_6BSP, ORF6_6BSP, ORF7_6BSP, and ORF8_6BSP.

A preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 6B Spain 2 AI surface protein which may be formulated or purified in an oligomeric (pilis) form. Another preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 6B Spain 2 AI surface protein which has been isolated in an oligomeric (pilis) form.

One or more of the *S. pneumoniae* strain 6B Spain 2 AI open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* strain 6B Spain 2 AI open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the *S. pneumoniae* strain 6B Spain 2 AI surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif.

The *S. pneumoniae* strain 6B Spain 2 AI surface proteins of the invention may affect the ability of the *S. pneumoniae* bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of *S. pneumoniae* to translocate through an epithelial cell layer. Preferably, one or more *S. pneumoniae* strain 6B Spain 2 AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. *S. pneumoniae* strain 6B Spain 2 AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The *S. pneumoniae* strain 6B Spain 2 AI sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. *S. pneumoniae* strain 6B Spain 2 AI may encode for at least one surface protein. Alternatively, *S. pneumoniae* strain 6B Spain 2 AI may encode for at least two surface exposed proteins and at least one sortase. Preferably, *S.*

PCT/US2005/027239

pneumoniae strain 6B Spain 2 AI encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a *S. pneumoniae* strain 6B Spain 2 AI surface protein such as orf3_6BSP, orf4_6BSP, or orf5_6BSP. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine or serine amino acid residue, respectively.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a *S. pneumoniae* strain 6B Spain 2 AI protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more *S. pneumoniae* strain 6B Spain 2 AI proteins and one or more AI proteins of any one or more of *S. pneumoniae* from TIGR4, 670, 19A Hungary 6, 6B Finland 12, 14 CSR 10, 9V Spain 3, 19F Taiwan 14, 23F Taiwan 15, or 23F Poland 16 AI, wherein one or more of the *S. pneumoniae* AI proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the *S. pneumoniae* strain 6B Spain 2 AI proteins, *S. pneumoniae* strain 6B Spain 2 AI may also include a transcriptional regulator.

S. pneumoniae strain 9V Spain 3 Adhesin Island

As discussed above, Applicants have identified adhesin islands within the genome of *S. pneumoniae* strain 9V Spain 3. The *S. pneumoniae* strain 9V Spain 3 Adhesin Island comprises a series of approximately seven open reading frames encoding for a collection of amino acid sequences comprising surface proteins and sortases. Specifically, the *S. pneumoniae* strain 9V Spain 3 AI proteins includes open reading frames encoding for two or more (*i.e.*, 2, 3, 4, 5, 6, or 7) of

PCT/US05/27239

ORF2_9VSP, ORF3_9VSP, ORF4_9VSP, ORF5_9VSP, ORF6_9VSP, ORF7_9VSP, and ORF8_9VSP.

A preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 9V Spain 3 AI surface protein which may be formulated or purified in an oligomeric (pilis) form. Another preferred immunogenic composition of the invention comprises a *S. pneumoniae* strain 9V Spain 3 AI surface protein which has been isolated in an oligomeric (pilis) form.

One or more of the *S. pneumoniae* strain 9V Spain 3 AI open reading frame polynucleotide sequences may be replaced by a polynucleotide sequence coding for a fragment of the replaced ORF. Alternatively, one or more of the *S. pneumoniae* strain 9V Spain 3 AI open reading frames may be replaced by a sequence having sequence homology to the replaced ORF.

One or more of the *S. pneumoniae* strain 9V Spain 3 AI surface protein sequences typically include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif.

The *S. pneumoniae* strain 9V Spain 3 AI surface proteins of the invention may affect the ability of the *S. pneumoniae* bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of *S. pneumoniae* to translocate through an epithelial cell layer. Preferably, one or more *S. pneumoniae* strain 9V Spain 3 AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. *S. pneumoniae* strain 9V Spain 3 AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

The *S. pneumoniae* strain 9V Spain 3 AI sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. *S. pneumoniae* strain 9V Spain 3 AI may encode for at least one surface protein. Alternatively, *S. pneumoniae* strain 9V Spain 3 AI may encode for at least two surface exposed proteins and at least one sortase. Preferably, *S. pneumoniae* strain 9V Spain 3 AI encodes for at least three surface exposed proteins and at least two sortases.

The AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a *S. pneumoniae* strain 9V Spain 3 AI surface protein such as orf3_9VSP, orf4_9VSP, or orf5_9VSP. The oligomeric, pilus-like structure may comprise numerous units of AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each

PCT/US05/27239
subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine or serine amino acid residue, respectively.

AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include a pilin motif.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a *S. pneumoniae* strain 9V Spain 3 AI protein in oligomeric form, preferably in a hyperoligomeric form. In one embodiment, the invention comprises a composition comprising one or more *S. pneumoniae* strain 9V Spain 3 AI proteins and one or more AI proteins from any one or more of *S. pneumoniae* from TIGR4, 670, 19A Hungary 6, 6B Finland 12, 6B Spain 2, 14 CSR 10, 19F Taiwan 14, 23F Taiwan 15, or 23F Poland 16 AI, wherein one or more of the *S. pneumoniae* AI proteins is in the form of an oligomer, preferably in a hyperoligomeric form.

In addition to the open reading frames encoding the *S. pneumoniae* strain 9V Spain 3 AI proteins, *S. pneumoniae* strain 9V Spain 3 AI may also include a transcriptional regulator.

The *S. pneumoniae* oligomeric, pilus-like structures may be isolated or purified from bacterial cultures in which the bacteria express an *S. pneumoniae* AI surface protein. The invention therefore includes a method for manufacturing an oligomeric AI surface antigen comprising culturing a *S. pneumoniae* bacterium that expresses the oligomeric AI protein and isolating the expressed oligomeric AI protein from the *S. pneumoniae* bacteria. The AI protein may be collected from secretions into the supernatant or it may be purified from the bacterial surface. The method may further comprise purification of the expressed AI protein. Preferably, the AI protein is in a hyperoligomeric form.

The oligomeric, pilus-like structures may be isolated or purified from bacterial cultures overexpressing an AI surface protein. The invention therefore includes a method for manufacturing an *S. pneumoniae* oligomeric Adhesin Island surface antigen comprising culturing a *S. pneumoniae* bacterium adapted for increased AI protein expression and isolation of the expressed oligomeric Adhesin Island protein from the *S. pneumoniae* bacteria. The AI protein may be collected from secretions into the supernatant or it may be purified from the bacterial surface. The method may further comprise purification of the expressed Adhesin Island protein. Preferably, the Adhesin Island protein is in a hyperoligomeric form.

The *S. pneumoniae* bacteria are preferably adapted to increase AI protein expression by at least two (*e.g.*, 2, 3, 4, 5, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 125, 150 or 200) times wild type expression levels.

S. pneumoniae bacteria may be adapted to increase AI protein expression by any means known in the art, including methods of increasing gene dosage and methods of gene upregulation. Such means include, for example, transformation of the *S. pneumoniae* bacteria with a plasmid

PCT/US2005/027239

encoding the AI protein. The plasmid may include a strong promoter or it may include multiple copies of the sequence encoding the AI protein. Optionally, the sequence encoding the AI protein within the *S. pneumoniae* bacterial genome may be deleted. Alternatively, or in addition, the promoter regulating the *S. pneumoniae* Adhesin Island may be modified to increase expression.

The invention further includes *S. pneumoniae* bacteria which have been adapted to produce increased levels of AI surface protein. In particular, the invention includes *S. pneumoniae* bacteria which have been adapted to produce oligomeric or hyperoligomeric AI surface protein. In one embodiment, the *S. pneumoniae* of the invention are inactivated or attenuated to permit *in vivo* delivery of the whole bacteria, with the AI surface protein exposed on its surface.

The invention further includes *S. pneumoniae* bacteria which have been adapted to have increased levels of expressed AI protein incorporated in pili on their surface. The *S. pneumoniae* bacteria may be adapted to have increased exposure of oligomeric or hyperoligomeric AI proteins on its surface by increasing expression levels of a signal peptidase polypeptide. Increased levels of a local signal peptidase expression in Gram positive bacteria (such us LepA in GAS) are expected to result in increased exposure of pili proteins on the surface of Gram positive bacteria. Increased expression of a leader peptidase in *S. pneumoniae* may be achieved by any means known in the art, such as increasing gene dosage and methods of gene upregulation. The *S. pneumoniae* bacteria adapted to have increased levels of leader peptidase may additionally be adapted to express increased levels of at least one pili protein.

Alternatively, the AI proteins of the invention may be expressed on the surface of a non-pathogenic Gram positive bacteria, such as *Streptococus gordonii* (See, e.g., Byrd et al., "Biological consequences of antigen and cytokine co-expression by recombinant *Streptococcus gordonii* vaccine vectors", Vaccine (2002) 20:2197-2205) or *Lactococcus lactis* (See, e.g., Mannam et al., "Mucosal Vaccine Made from Live, Recombinant Lactococcus lactis Protects Mice against Pharangeal Infection with *Streptococcus pyogenes*" Infection and Immunity (2004) 72(6):3444-3450). As used herein, non-pathogenic Gram positive bacteria refer to Gram positive bacteria which are compatible with a human host subject and are not associated with human pathogenisis. Preferably, the non-pathogenic bacteria are modified to express the AI surface protein in oligomeric, or hyper-oligomeric form. Sequences encoding for an AI surface protein and, optionally, an AI sortase, may be integrated into the non-pathogenic Gram positive bacterial genome or inserted into a plasmid. The non-pathogenic Gram positive bacteria may be inactivated or attenuated to facilitate *in vivo* delivery of the whole bacteria, with the AI surface protein exposed on its surface. Alternatively, the AI surface protein may be isolated or purified from a bacterial culture of the non-pathogenic Gram positive bacteria. For example, the AI surface protein may be isolated from cell extracts or culture supernatants. Alternatively, the AI surface protein may be isolated or purified from the surface of the non-pathogenic Gram positive bacteria.

The non-pathogenic Gram positive bacteria may be used to express any of the *S. pneumoniae* Adhesin Island proteins described herein. The non-pathogenic Gram positive bacteria are transformed

PCT/US2005/027239
to express an Adhesin Island surface protein. Preferably, the non-pathogenic Gram positive bacteria also express at least one Adhesin Island sortase. The AI transformed non-pathogenic Gram positive bacteria of the invention may be used to prevent or treat infection with pathogenic *S. pneumoniae*.

Figures 190 A and B, and 193-195 provide examples of three methods successfully practiced by applicants to purify pili from *S. pneumoniae* TIGR4.

Immunogenic Compositions

The Gram positive bacteria AI proteins described herein are useful in immunogenic compositions for the prevention or treatment of Gram positive bacterial infection. For example, the GBS AI surface proteins described herein are useful in immunogenic compositions for the prevention or treatment of GBS infection. As another example, the GAS AI surface proteins described herein may be useful in immunogenic compositions for the prevention or treatment of GAS infection. As another example, the *S. pneumoniae* AI surface proteins may be useful in immunogenic cojmpositions for the prevention or treatment of *S. pneumoniae* infection.

Gram positive bacteria AI surface proteins that can provide protection across more than one serotype or strain isolate may be used to increase immunogenic effectiveness. For example, a particular GBS AI surface protein having an amino acid sequence that is at least 50% (*i.e.*, at least 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100%) homologous to the particular GBS AI surface protein of at least 2 (*i.e.*, at least 3, 4, 5, 6, 7, 8, 9, 10, or more) other GBS serotypes or strain isolates may be used to increase the effectiveness of such compositions.

As another example, fragments of Gram positive bacteria AI surface proteins that can provide protection across more than one serotype or strain isolate may be used to increase immunogenic effectiveness. Such a fragment may be identified within a consensus sequence of a full length amino acid sequence of a Gram positive bacteria AI surface protein. Such a fragment can be identified in the consensus sequence by its high degree of homology or identity across multiple (*i.e*, at least 3, 4, 5, 6, 7, 8, 9, 10, or more) Gram positive bacteria serotypes or strain isolates. Preferably, a high degree of homology is a degree of homology of at least 90% (*i.e.*, at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100%) across Gram positive bacteria serotypes or strain isolates. Preferably, a high degree of identity is a degree of identity of at least 90% (*i.e.*, at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100%) across Gram positive bacteria serotypes or strain isolates. In one embodiment of the invention, such a fragment of a Gram positive bacteria AI surface protein may be used in the immunogenic compositions.

In addition, the AI surface protein oligomeric pilus structures may be formulated or purified for use in immunization. Isolated AI surface protein oligomeric pilus structures may also be used for immunization.

The invention includes an immunogenic composition comprising a first Gram positive bacteria AI protein and a second Gram positive bacterial AI protein. One or more of the AI proteins

PCT/US2005/027239

may be a surface protein. Such surface proteins may contain an LPXTG motif or other sortase substrate motif.

The first and second AI proteins may be from the same or different genus or species of Gram positive bacteria. If within the same species, the first and second AI proteins may be from the same or different AI subtypes. If two AIs are of the same subtype, the AIs have the same numerical designation. For example, all AIs designated as AI-1 are of the same AI subtype. If two AIs are of a different subtype, the AIs have different numerical designations. For example, AI-1 is of a different AI subtype from AI-2, AI-3, AI-4, etc. Likewise, AI-2 is of a different AI subtype from AI-1, AI-3, and AI-4, etc.

For example, the invention includes an immunogenic composition comprising one or more GBS AI-1 proteins and one or more GBS AI-2 proteins. One or more of the AI proteins may be a surface protein. Such surface proteins may contain an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) and may bind fibrinogen, fibronectin, or collagen. One or more of the AI proteins may be a sortase. The GBS AI-1 proteins may be selected from the group consisting of GBS 80, GBS 104, GBS 52, SAG0647 and SAG0648. Preferably, the GBS AI-1 proteins include GBS 80 or GBS 104.

The GBS AI-2 proteins may be selected from the group consisting of GBS 67, GBS 59, GBS 150, SAG1405, SAG1406, 01520, 01521, 01522, 01523, 01523, 01524 and 01525. In one embodiment, the GBS AI-2 proteins are selected from the group consisting of GBS 67, GBS 59, GBS 150, SAG1405, and SAG1406. In another embodiment, the GBS AI-2 proteins may be selected from the group consisting of 01520, 01521, 01522, 01523, 01523, 01524 and 01525. Preferably, the GBS AI-2 protein includes GBS 59 or GBS 67.

As another example, the invention includes an immunogenic composition comprising one or more of any combination of GAS AI-1, GAS AI-2, GAS AI-3, or GAS AI-4 proteins. One or more of the GAS AI proteins may be a sortase. The GAS AI-1 proteins may be selected from the group consisting of M6_Spy0156, M6_Spy0157, M6_Spy0158, M6_Spy0159, M6_Spy0160, M6_Spy0161, CDC SS 410_fimbrial, ISS3650_fimbrial, and DSM2071_fimbrial. Preferably, the GAS AI-1 proteins are selected from the group consisting of M6_Spy0157, M6_Spy0159, M6_Spy0160, CDC SS 410_fimbrial, ISS3650_fimbrial, and DSM2071_fimbrial.

The GAS AI-2 proteins may be selected from the group consisting of Spy0124, GAS15, Spy0127, GAS16, GAS17, GAS18, Spy0131, Spy0133, and GAS20. Preferably, the GAS AI-2 proteins are selected from the group consisting of GAS15, GAS16, and GAS18.

The GAS AI-3 proteins may be selected from the group consisting of SpyM3_0097, SpyM3_0098, SpyM3_0099, SpyM3_0100, SpyM3_0101, SpyM3_0102, SpyM3_0103, SpyM3_0104, SPs0099, SPs0100, SPs0101, SPs0102, SPs0103, SPs0104, SPs0105, SPs0106, orf77, orf78, orf79, orf80, orf81, orf82, orf83, orf84, spyM18_0125, spyM18_0126, spyM18_0127, spyM18_0128, spyM18_0129, spyM18_0130, spyM18_0131, spyM18_0132, SpyoM01000156, SpyoM01000155, SpyoM01000154, SpyoM01000153, SpyoM01000152, SpyoM01000151, SpyoM01000150, SpyoM01000149, ISS3040_fimbrial, ISS3776_fimbrial, and ISS4959_fimbrial. In

PCT/US05/27239

one embodiment the GAS AI-3 proteins are selected from the group consisting of SpyM3_0097, SpyM3_0098, SpyM3_0099, SpyM3_0100, SpyM3_0101, SpyM3_0102, SpyM3_0103, and SpyM3_0104. In another embodiment, the GAS AI-3 proteins are selected from the group consisting of SPs0099, SPs0100, SPs0101, SPs0102, SPs0103, SPs0104, SPs0105, and SPs0106. In yet another embodiment, the GAS AI-3 proteins are selected from the group consisting of orf77, orf78, orf79, orf80, orf81, orf82, orf83, and orf84. In a further embodiment, the GAS AI-3 proteins are selected from the group consisting of spyM18_0125, spyM18_0126, spyM18_0127, spyM18_0128, spyM18_0129, spyM18_0130, spyM18_0131, and spyM18_0132. In yet another embodiment the GAS AI-3 proteins are selected from the group consisting of SpyoM01000156, SpyoM01000155, SpyoM01000154, SpyoM01000153, SpyoM01000152, SpyoM01000151, SpyoM01000150, and SpyoM01000149.

The GAS AI-4 proteins may be selected from the group consisting of 19224133, 19224134, 19224135, 19224136, 19224137, 19224138, 19224139, 19224140, 19224141, 20010296_fimbrial, 20020069_fimbrial, CDC SS 635_fimbrial, ISS4883_fimbrial, and ISS4538_fimbrial. Preferably, the GAS-AI4 proteins are selected from the group consisting of 19224134, 19224135, 19224137, 19224139, 19224141, 20010296_fimbrial, 20020069_fimbrial, CDC SS 635_fimbrial, ISS4883_fimbrial, and ISS4538_fimbrial.

As yet another example, the invention includes an immunogenic composition comprising one or more of any combination of *S. pneumonaie* from TIGR4, *S. pneumonaie* strain 670, *S. pneumonaie* from 19A Hungary 6, *S. pneumonaie* from 6B Finland 12, *S. pneumonaie* from 6B Spain 2, *S. pneumonaie* from 9V Spain 3, *S. pneumonaie* from 14 CSR 10, *S. pneumonaie* from 19F Taiwan 14, *S. pneumonaie* from 23F Taiwan 15, or *S. pneumonaie* from 23F Poland 16 AI proteins. One or more of the AI proteins may be a surface protein. Such surface proteins may contain an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) and may bind fibrinogen, fibronectin, or collagen. One or more of the AI proteins may be a sortase.

The *S. pneumonaie* from TIGR4 AI proteins may be selected from the group consisting of SP0462, SP0463, SP0464, SP0465, SP0466, SP0467, SP0468. Preferably, the *S. pneumonaie* from TIGR4 AI proteins include SP0462, SP0463, or SP0464.

The *S. pneumonaie* strain 670 AI proteins may be selected from the group consisting of Orf1_670, Orf3_670, Orf4_670, Orf5_670, Orf6_670, Orf7_670, and Orf8_670. Preferably, the *S. pneumonaie* strain 670 AI proteins include Orf3_670, Orf4_670, or Orf5_670.

The *S. pneumonaie* from 19A Hungary 6 AI proteins may be selected from the group consisting of ORF2_19AH, ORF3_19AH, ORF4_19AH, ORF5_19AH, ORF6_19AH, ORF7_19AH, or ORF8_19AH.

The *S. pneumonaie* from 6B Finland 12 AI proteins may be selected from the group consisting of ORF2_6BF, ORF3_6BF, ORF4_6BF, ORF5_6BF, ORF6_6BF, ORF7_6BF , or ORF8_6BF.

PCT/US2005/027239

The *S. pneumonaie* from 6B Spain 2 AI proteins may be selected from the group consisting of ORF2_6BSP, ORF3_6BSP, ORF4_6BSP, ORF5_6BSP, ORF6_6BSP, ORF7_6BSP , or ORF8_6BSP.

The *S. pneumonaie* from 9V Spain 3 AI proteins may be selected from the group consisting of ORF2_9VSP, ORF3_9VSP, ORF4_9VSP, ORF5_9VSP, ORF6_9VSP, ORF7_9VSP , or ORF8_9VSP.

The *S. pneumonaie* from 14 CSR 10 AI proteins may be selected from the group consisting of ORF2_14CSR, ORF3_14CSR, ORF4_14CSR, ORF5_14CSR, ORF6_14CSR, ORF7_14CSR , or ORF8_14CSR.

The *S. pneumonaie* from 19F Taiwan 14 AI proteins may be selected from the group consisting of ORF2_19FTW, ORF3_19FTW, ORF4_19FTW, ORF5_19FTW, ORF6_19FTW, ORF7_19FTW , or ORF8_19FTW.

The *S. pneumonaie* from 23F Taiwan 15 AI proteins may be selected from the group consisting of ORF2_23FTW, ORF3_23FTW, ORF4_23FTW, ORF5_23FTW, ORF6_23FTW, ORF7_23FTW, or ORF8_23FTW.

The *S. pneumonaie* from 23F Poland 16 AI proteins may be selected from the group consisting of ORF2_23FP, ORF3_23FP, ORF4_23FP, ORF5_23FP, ORF6_23FP, ORF7_23FP , or ORF8_23FP.

Preferably, the Gram positive bacteria AI proteins included in the immunogenic compositions of the invention can provide protection across more than one serotype or strain isolate. For example, the immunogenic composition may comprise a first AI protein, wherein the amino acid sequence of said AI protein is at least 90% (*i.e.*, at least 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 or 100%) homologous to the amino acid sequence of a second AI protein, and wherein said first AI protein and said second AI protein are derived from the genomes of different serotypes of a Gram positive bacteria. The first AI protein may also be homologous to the amino acid sequence of a third AI protein, such that the first AI protein, the second AI protein and the third AI protein are derived from the genomes of different serotypes of a Gram positive bacteria. The first AI protein may also be homologous to the amino acid sequence of a fourth AI protein, such that the first AI protein, the second AI protein and the third AI protein are derived from the genomes of different serotypes of a Gram positive bacteria.

For example, preferably, the GBS AI proteins included in the immunogenic compositions of the invention can provide protection across more than one GBS serotype or strain isolate. For example, the immunogenic composition may comprise a first GBS AI protein, wherein the amino acid sequence of said AI protein is at least 90% (*i.e.*, at least 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 or 100%) homologous to the amino acid sequence of a second GBS AI protein, and wherein said first AI protein and said second AI protein are derived from the genomes of different GBS serotypes. The first GBS AI protein may also be homologous to the amino acid sequence of a third GBS AI protein, such that the first AI protein, the second AI protein and the third AI protein are derived from the genomes of

different GBS serotypes. The first AI protein may also be homologous to the amino acid sequence of a fourth GBS AI protein, such that the first AI protein, the second AI protein and the third AI protein are derived from the genomes of different GBS serotypes.

The first AI protein may be selected from an AI-1 protein or an AI-2 protein. For example, the first AI protein may be a GBS AI-1 surface protein such as GBS 80. The amino acid sequence of GBS 80 from GBS serotype V, strain isolate 2603 is greater than 90% homologous to the GBS 80 amino acid sequence from GBS serotype III, strain isolates NEM316 and COH1 and the GBS 80 amino acid sequence from GBS serotype 1a, strain isolate A909.

As another example, the first AI protein may be GBS 104. The amino acid sequence of GBS 104 from GBS serotype V, strain isolate 2603 is greater than 90% homologous to the GBS 104 amino acid sequence from GBS serotype III, strain isolates NEM316 and COH1, the GBS 104 amino acid sequence from GBS serotype 1a, strain isolate A909, and the GBS 104 amino acid sequence serotype II, strain isolate 18RS21.

Table 12 provides the amino acid sequence identity of GBS 80 and GBS 104 across GBS serotypes Ia, Ib, II, III, V, and VIII. The GBS strains in which genes encoding GBS 80 and GBS 104 were identified share, on average, 99.88 and 99.96 amino acid sequence identity, respectively. This high degree of amino acid identity indicates that an immunogenic composition comprising a first protein of GBS 80 or GBS 104 may provide protection across more than one GBS serotype or strain isolate.

Table 12. Conservation of GBS 80 and GBS 104 amino acid sequences

Serotype	Strains	GBS 80		GBS 104	
		cGH	%AA identity	cGH	%AA identity
Ia	**090**	+	99.79	+	100.00
	A909	+	100.00	+	100.00
	515	-		-	
	DK1	-		-	
	DK8	-		-	
	Davis	-		-	
Ib	**7357b**	+	100.00	+	
	H36B	-		-	
II	**18RS21**	-		+	100.00
	DK21	-		-	
III	**NEM316**	+	100.00	+	100.00
	COH31	+	100.00	+	
	D136	+	100.00	+	
	M732	+	100.00	+	99.88
	COH1	+	99.79	+	99.88
	M781	+	99.79	+	99.88
No type	**CJB110**	+	99.37	+	100.00
	1169NT	-		-	
V	**CJB111**	+	100.00	+	100.00
	2603	+	100.00	+	100.00
VIII	**JM130013**	+	99.79	+	100.00
	SMU014	+	100.00	+	

Serotype	Strains	GBS 80		GBS 104	
		cGH	%AA identity	cGH	%AA identity
total		14/22	99.88+/-0.19	15/22	99.96 +/-0.056

As another example, the first AI protein may be an AI-2 protein such as GBS 67. The amino acid sequence of GBS 67 from GBS serotype V, strain isolate 2603 is greater than 90% homologous to the GBS 67 amino acid sequence from GBS serotype III, strain isolate NEM316, the GBS 67 amino acid sequence from GBS serotype 1b, strain isolate H36B, and the GBS 67 amino acid sequence from GBS serotype II, strain isolate 17RS21.

As another example, the first AI protein may be an AI-2 protein such as spb1. The amino acid sequence of spb1 from GBS serotype III, strain isolate COH1 is greater than 90% homologous to the spb1 amino acid sequence from GBS serotype Ia, strain isolate A909.

As yet another example, the first AI protein may be an AI-2 protein such as GBS 59. The amino acid sequence of GBS 59 from GBS serotype II, strain isolate 18RS21 is 100% homologous to the GBS 59 amino acid sequence from GBS serotype V, strain isolate 2603. The amino acid sequence of GBS 59 from GBS serotype V, strain isolate CJB111 is 98% homologous to the GBS 59 amino acid sequence from GBS serotype III, strain isolate NEM316.

The compositions of the invention may also be designed to include Gram positive AI proteins from divergent serotypes or strain isolates, *i.e.*, to include a first AI protein which is present in one collection of serotypes or strain isolates of a Gram positive bacteria and a second AI protein which is present in those serotypes or strain isolates not represented by the first AI protein.

For example, the invention may include an immunogenic composition comprising a first and second Gram positive bacteria AI protein, wherein a polynucleotide sequence encoding for the full length sequence of the first AI protein is not present in a similar Gram positive bacterial genome comprising a polynucleotide sequence encoding for the second AI protein.

The compositions of the invention may also be designed to include AI proteins from divergent GBS serotypes or strain isolates, *i.e.*, to include a first AI protein which is present in one collection of GBS serotypes or strain isolates and a second AI protein which is present in those serotypes or strain isolates not represented by the first AI protein.

For example, the invention may include an immunogenic composition comprising a first and second GBS AI protein, wherein a polynucleotide sequence encoding for the full length sequence of the first GBS AI protein is not present in a genome comprising a polynucleotide sequence encoding for the second GBS AI protein. For example, the first AI protein could be GBS 80 (such as the GBS 80 sequence from GBS serotype V, strain isolate 2603). As previously discussed (and depicted in Figure 2), the sequence for GBS 80 in GBS sertoype II, strain isolate 18RS21 is disrupted. In this instance, the second AI protein could be GBS 104 or GBS 67 (sequences selected from the GBS serotype II, strain isolate 18RS21).

PCT/US2005/027239

Further, the the invention may include an immunogenic composition comprising a first and second GBS AI protein, wherein the first GBS AI protein has detectable surface exposure on a first GBS strain or serotype but not a second GBS strain or serotype and the second GBS AI protein has detectable surface exposure on a second GBS strain or serotype but not a first GBS strain or serotype. For example, the first AI protein could be GBS 80 and the second AI protein could be GBS 67. As seen in Table 15, there are some GBS serotypes and strains that have surface exposed GBS 80 but that do not have surface exposed GBS 67 and vice versa. An immunogenic composition comprising a GBS 80 and a GBS 67 protein may provide protection across a wider group of GBS strains and serotypes.

PCT/US05/27239

Table 15: Antigen surface exposure of GBS 80, GBS 322, GBS 104, and GBS 67

GBS strains	Type	GBS 80	GBS 322	GBS 104	GBS 67
DK1*	Ia	0	nd	237	478
DK8*		0	213	151	475
Davis*		0	86	271	430
515*		0	227	262	409
090		0	0	0	0
A909		0	0	0	0
2986		0	0	157	397
5551		0	36	384	485
2177	Ib	477	323	328	66
H36B*		0	105	518	444
7357b-		91	102	309	316
2129		57	71	132	0
5518		31	nd	60	28
COH1	III	305	130	305	0
D136C		16	460	226	406
COH31		0	479	71	273
M732		105	292	101	0
M781		65	224	136	0
1998		95	288	205	350
5376		165	76	156	0
5435		93	88	100	0
18RS21	II	0	471	50	103
DK21*		0	342	419	331
3050		43	188	289	460
5401		170	135	494	618
2141		0	76	0	69
CJB111	V	365	58	355	481
2603		62	293	100	105
5364		454	463	379	394
2110		0	11	345	589
2274	IV	113	161	465	484
1999		0	55	492	453
2210		0	0	363	574
2928	VII	0	0	0	0
SMU071	VIII	556	170	393	79
JM9130013		587	133	436	83
2189		0	0	0	0
5408		0	0	159	433
CJB110	NT	71	587	169	245
1169*		0	213	371	443
Δ Mean > 100		9/40	22/38	32/40	25/40
		22%	58%	80%	62%

Alternatively, the invention may include an immunogenic composition comprising a first and second Gram positive bacteria AI protein, wherein the polynucleotide sequence encoding the sequence of the first AI protein is less than 90 % (*i.e.*, less than 90, 88, 86, 84, 82, 80, 78, 76, 74, 72, 70, 65, 60, 55, 50, 45, 40, 35 or 30 percent) homologous than the corresponding sequence in the genome of the second AI protein.

PCT/US05/27239

The invention may include an immunogenic composition comprising a first and second GBS AI protein, wherein the polynucleotide sequence encoding the sequence of the first GBS AI protein is less than 90 % (*i.e.*, less than 90, 88, 86, 84, 82, 80, 78, 76, 74, 72, 70, 65, 60, 55, 50, 45, 40, 35 or 30 percent) homologous than the corresponding sequence in the genome of the second GBS AI protein. For example, the first GBS AI protein could be GBS 67 (such as the GBS 67 sequence from GBS serotype 1b, strain isolate H36B). As shown in Figures 2 and 4, the GBS 67 sequence for this strain is less than 90% homologous (87%) to the corresponding GBS 67 sequence in GBS serotype V, strain isolate 2603. In this instance, the second GBS AI protein could then be the GBS 80 sequence from GBS serotype V, strain isolate 2603.

An example immunogenic composition of the invention may comprise adhesin island proteins GBS 80, GBS 104, GBS 67, and GBS 59, and non-AI protein GBS 322. FACS analysis of different GBS strains demonstrates that at least one of these five proteins is always found to be expressed on the surface of GBS bacteria. An initial FACS analysis of 70 strains of GBS bacteria, obtained from the CDC in the United States (33 strains), ISS in Italy (17 strains), and Houston/Harvard (20 strains), detected surface exposure of at least one of GBS 80, GBS 104, GBS 322, GBS 67, or GBS 59 on the surface of the GBS bacteria. Figure 227 provides the FACS data obtained for surface exposure of GBS 80, GBS 104, GBS 67, GBS 322, and GBS 59 on each of 37 GBS strains. Figure 228 provides the FACS data obtained for surface exposure of GBS 80, GBS 104, GBS 67, GBS 322, and GBS 59 on each of 41 GBS strains obtained from the CDC. As can be seen from Figures 227 and 228, each GBS strain had surface expression of at least one of GBS 80, GBS 104, GBS 67, GBS 322, and GBS 59. The surface exposure of at least one of these proteins on each bacterial strain indicates that an immunogenic composition comprising these proteins will provide wide protection across GBS strains and serotypes.

The surface exposed GBS 80, GBS 104, GBS 67, GBS 322, and GBS 59 proteins are also present at high levels as determined by FACS. Table 49 summarizes the FACS results for the initial 70 GBS strains examined for GBS 80, GBS 104, GBS 67, GBS 322, and GBS 59 surface expression. A protein was designated as having high levels of surface expression of a protein if a five-fold shift in fluorescence was observed when using antibodies for the protein relative to preimmune control serum.

Table 49: Exposure Levels of GBS 80, GBS 104, GBS 67, GBS 322, and GBS 59 on GBS Strains

5-fold shift in fluorescence by FACS	GBS 80	GBS 104	GBS 67	GBS 59	GBS 322
	17/70	14/70	49/70	46/70	33/70
	24%	20%	70%	66%	47%

Table 50 details which of the surface proteins is highly expressed on the different GBS serotype.

Table 50: High Levels of Surface Protein Expression on GBS Serotypes

5-fold shift in fluorescence by FACS	GBS 80	GBS 104	GBS 67	GBS 59	GBS 322
Ia + Ib + III	4/36	2/36	22/36	20/36	18/36
II + V	11/25	9/25	21/25	21/25	13/25
Others	2/9	3/9	6/9	5/9	2/9

Alternatively, the immunogenic composition of the invention may include GBS 80, GBS 104, GBS 67, and GBS 322. Assuming that protein antigens that are highly accessible to antibodies confer 100% protection with suitable adjuvants, an immunogenic composition containing GBS 80, GBS 104, GBS 67, GBS 59 and GBS 322 will provide protection for 89% of GBS strains and serotypes, the same percentage as an immunogenic composition containing GBS 80, GBS 104, GBS 67, and GBS 322 proteins. See Figure 229. However, it may be preferable to include GBS 59 in the composition to increase its immunogenic strength. As seen from Table 50, GBS 59 is highly expressed on the surface two-thirds of GBS bacteria examined by FACS analysis, unlike GBS 80, GBS 104, and GBS 322, which are highly expressed in less than half of GBS bacteria examined. GBS 59 opsonophagocytic activity is also comparable to that of a mix of GBS 322, GBS 104, GBS 67, and GBS 80 proteins. See Figure 230.

By way of another example, preferably, the GAS AI proteins included in the immunogenic compositions of the invention can provide protection across more than one GAS serotype or strain isolate. For example, the immunogenic composition may comprise a first GAS AI protein, wherein the amino acid sequence of said AI protein is at least 90% (*i.e.*, at least 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 or 100%) homologous to the amino acid sequence of a second GAS AI protein, and wherein said first AI protein and said second AI protein are derived from the genomes of different GAS serotypes. The first GAS AI protein may also be homologous to the amino acid sequence of a third GAS AI protein, such that the first AI protein, the second AI protein and the third AI protein are derived from the genomes of different GAS serotypes. The first AI protein may also be homologous to the amino acid sequence of a fourth GAS AI protein, such that the first AI protein, the second AI protein and the third AI protein are derived from the genomes of different GAS serotypes.

The compositions of the invention may also be designed to include GAS AI proteins from divergent serotypes or strain isolates, *i.e.*, to include a first AI protein which is present in one collection of serotypes or strain isolates of a GAS bacteria and a second AI protein which is present in those serotypes or strain isolates not represented by the first AI protein.

For example, the first AI protein could be a prtF2 protein (such as the 19224141 protein from GAS serotype M12, strain isolate A735). As previously discussed (and depicted in Figure 164), the sequence for a prtF2 protein is not present in GAS AI types 1 or 2. In this instance, the second AI protein could be collagen binding protein M6_Spy0159 (from M6 isolate (MGAS10394), which comprises an AI-1) or GAS15 (from M1 isolate (SF370), which comprises an AI-2).

Further, the invention may include an immunogenic composition comprising a first and second GAS AI protein, wherein the first GAS AI protein has detectable surface exposure on a first GAS strain or serotype but not a second GAS strain or serotype and the second GAS AI protein has detectable surface exposure on a second GAS strain or serotype but not a first GAS strain or serotype.

The invention may include an immunogenic composition comprising a first and second GAS AI protein, wherein the polynucleotide sequence encoding the sequence of the first GAS AI protein is

PCT/US2005/027239

less than 90 % (*i.e.*, less than 90, 88, 86, 84, 82, 80, 78, 76, 74, 72, 70, 65, 60, 55, 50, 45, 40, 35 or 30 percent) homologous than the corresponding sequence in the genome of the second GAS AI protein. Preferably the first and second GAS AI proteins are subunits of the pilus. More preferably the first and second GAS AI proteins are selected from the major pilus forming proteins (i.e., M6_Spy0160 from M6 strain 10394, SPy0128 from M1 strain SF370, SpyM3_0100 from M3 strain 315, SPs0102 from M3 strain SSI, orf80 from M5 isolate Manfredo, spyM18_0128 from M18 strain 8232, SpyoM01000153 from M49 strain 591, 19224137 from M12 strain A735, fimbrial structural subunit from M77 strain ISS4959, fimbrial structural subunit from M44 strain ISS3776, fimbrial structural subunit from M50 strain ISS3776 ISS 4538, fimbrial structural subunit from M12strain CDC SS635, fimbrial structural subunit from M23 strain DSM2071, fimbrial structural subunit from M6 strain CDC SS410). Table 45 provides the percent identity between the amino acidic sequences of each of the main pilus forming subunits from GAS AI-1, AI-2, AI-3, and AI-4 representative strains (i.e., M6_Spy0160 from M6 strain 10394, SPy0128 from M1 strain SF370, SpyM3_0100 from M3 strain 315, SPs0102 from M3 strain SSI, orf80 from M5 isolate Manfredo, spyM18_0128 from M18 strain 8232, SpyoM01000153 from M49 strain 591, 19224137 from M12 strain A735, Fimbrial structural subunit from M77 strain ISS4959, fimbrial structural subunit from M44 strain ISS3776, fimbrial structural subunit from M50 strain ISS3776 ISS 4538, fimbrial structural subunit from M12strain CDC SS635, fimbrial structural subunit from M23 strain DSM2071, fimbrial structural subunit from M6 strain CDC SS410).

Table 45: Comparison of Amino Acid Sequences of Major Pilus Proteins in the Four GAS AIs

		AI-1	**AI-2**	**AI-3**			**AI-4**
		M6-10394	M1-370	M3-315	M5-Manfredo	M18-8232	M12-A735
AI-1	M6-10394	**100%** M6-10394 M6-SS-410 **56%** M23-DSM2071	23%	25%	23%	24%	26%
AI-2	M1-370	23%	**100%**	40%	41%	38%	40%
AI-3	M3-315	25%	40%	**100%** M3-315 M3-SSI-1	64%	67%	61%
	M5-Manfredo	23%	39%	64%	**100%** M5-Manfredo M44-3776 M77-4959	60%	65%

	M18-8232	24%	38%	67%	60%	**90%** M18-8232 M49-591	62%
AI-4	M12-A735	26%	40%	61%	65%	62%	**97-100%** M12-A735 M12-SS-635 M50-4538

For example, the first main pilus subunit may be selected from bacteria of GAS serotype M6 strain 10394 and the second main pilus subunit may be selected from bacteria of GAS serotype M1 strain 370. As can be seen from Table 45, the main pilus subunits encoded by these strains of bacteria share only 23% nucleotide identity. An immunogenic composition comprising pilus main subunits from each of these strains of bacteria is expected to provide protection across a wider group of GAS strains and serotypes. Other examples of main pilus subunits that can be used in combination to provide increased protection across a wider range of GAS strains and serotypes include proteins encoded by GAS serotype M5 Manfredo isolate and serotype M6 strain 10394, which share 23% sequence identity, GAS serotype M18 strain 8232 and serotype M1 strain 370, which share 38% sequence identity, GAS serotype M3 strain 315 and serotype M12 strain A735, which share 61% sequence identity, and GAS serotype M3 strain 315 and serotype M6 strain 10394 which share 25% sequence identity.

As also can be seen from Table 45, the amino acid sequences of the four types of main pilus subunits present in GAS are relatively divergent. Figures 198-201 provide further tables comparing the percent identity of adhesin island-encoded surface exposed proteins for different GAS serotypes relative to other GAS serotypes harbouring an adhesin island of the same or a different subtype (GAS AI-1, GAS AI-2, GAS AI-3, and GAS AI-4). See also further discussion below.

Immunizations with the Adhesin Island proteins of the invention are discussed further in the Examples.

Co-expression of GBS Adhesin Island proteins and role of GBS AI proteins in surface presentation

In addition to the use of the GBS adhesin island proteins for cross strain and cross serotype protection, Applicants have identified interactions between adhesin island proteins which appear to affect the delivery or presentation of the surface proteins on the surface of the bacteria.

In particular, Applicants have discovered that surface exposure of GBS 104 is dependent on the concurrent expression of GBS 80. As discussed further in Example 2, reverse transcriptase PCR analysis of AI-1 shows that all of the AI genes are co-transcribed as an operon. Applicants constructed a series of mutant GBS containing in frame deletions of various AI-1 genes. (A schematic of the GBS mutants is presented in Figure 7). FACS analysis of the various mutants comparing mean shift values using anti-GBS 80 versus anti-GBS 104 antibodies is presented in Figure 8. Removal of the GBS 80 operon prevented surface exposure of GBS 104; removal of the

PCT/US05/27239

GBS 104 operon did not affect surface exposure of GBS 80. While not being limited to a specific theory, it is thought that GBS 80 is involved in the transport or localization of GBS 104 to the surface of the bacteria. The two proteins may be oligomerized or otherwise associated. It is possible that this association involves a conformational change in GBS 104 that facilitates its transition to the surface of the GBS bacteria.

Pili structures that comprise GBS 104 appear to be of a lower molecular weight than pili structures lacking GBS 104. Figure 68 shows that polyclonal anti-GBS 104 antibodies (see lane marked α-104 POLIC.) cross-hybridize with smaller structures than do polyclonal anti-GBS 80 antibodies (see lane marked α-GBS 80 POLIC.).

In addition, Applicants have shown that removal of GBS 80 can cause attenuation, further suggesting the protein contributes to virulence. As described in more detail in Example 3, the LD_{50}'s for the Δ80 mutant and the Δ80, Δ104 double mutant were reduced by an order of magnitude compared to wildtype and Δ104 mutant.

The sortases within the adhesin island also appear to play a role in localization and presentation of the surface proteins. As discussed further in Example 4, FACS analysis of various sortase deletion mutants showed that removal of sortase SAG0648 prevented GBS 104 from reaching the surface and slightly reduced the surface exposure of GBS 80. When sortase SAG0647 and sortase SAG0648 were both knocked out, neither GBS 80 nor GBS 104 were surface exposed. Expression of either sortase alone was sufficient for GBS 80 to arrive at the bacterial surface. Expression of SAG0648, however, was required for GBS 104 surface localization.

Accordingly, the compositions of the invention may include two or more AI proteins, wherein the AI proteins are physically or chemically associated. For example, the two AI proteins may form an oligomer. In one embodiment, the associated proteins are two AI surface proteins, such as GBS 80 and GBS 104. The associated proteins may be AI surface proteins from different adhesin islands, including host cell adhesin island proteins if the AI surface proteins are expressed in a recombinant system. For example, the associated proteins may be GBS 80 and GBS 67.

Adhesin Island proteins from other Gram positive bacteria

Applicants' identification and analysis of the GBS adhesin islands and the immunological and biological functions of these AI proteins and their pilus structures provides insight into similar structures in other Gram positive bacteria.

As discussed above, "Adhesin Island" or "AI" refers to a series of open reading frames within a bacterial genome that encode for a collection of surface proteins and sortases. An Adhesin Island may encode for amino acid sequences comprising at least one surface protein. The Adhesin Island may encode at least one surface protein. Alternatively, an Adhesin Island may encode for at least two surface proteins and at least one sortase. Preferably, an Adhesin Island encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. One or more AI

PCT/US05/27239

surface proteins may participate in the formation of a pilus structure on the surface of the Gram positive bacteria.

Gram positive adhesin islands of the invention preferably include a divergently transcribed transcriptional regulator. The transcriptional regulator may regulate the expression of the AI operon.

The invention includes a composition comprising one or more Gram positive bacteria AI surface proteins. Such AI surface proteins may be associated in an oligomeric or hyperoligomeric structure.

Preferred Gram positive adhesin island proteins for use in the invention may be derived from *Staphylococcus* (such as *S. aureus*), *Streptococcus* (such as *S. agalactiae* (GBS), *S. pyogenes* (GAS), *S. pneumonaie*, *S. mutans*), *Enterococcus* (such as *E. faecalis* and *E. faecium*), *Clostridium* (such as *C. difficile*), *Listeria* (such as *L. monocytogenes*) and *Corynebacterium* (such as *C. diphtheria*).

One or more of the Gram positive AI surface protein sequences typically include an LPXTG motif or other sortase substrate motif. Gram positive AI surface proteins of the invention may affect the ability of the Gram positive bacteria to adhere to and invade epithelial cells. AI surface proteins may also affect the ability of Gram positive bacteria to translocate through an epithelial cell layer. Preferably, one or more AI surface proteins are capable of binding to or otherwise associating with an epithelial cell surface. Gram positive AI surface proteins may also be able to bind to or associate with fibrinogen, fibronectin, or collagen.

Gram positive AI sortase proteins are predicted to be involved in the secretion and anchoring of the LPXTG containing surface proteins. A Gram positive bacteria AI may encode for at least one surface exposed protein. The Adhesin Island may encode at least one surface protein. Alternatively, a Gram positive bacteria AI may encode for at least two surface exposed proteins and at least one sortase. Preferably, a Gram positive AI encodes for at least three surface exposed proteins and at least two sortases.

Gram positive AI surface proteins may be covalently attached to the bacterial cell wall by membrane-associated transpeptidases, such as an AI sortase. The sortase may function to cleave the surface protein, preferably between the threonine and glycine residues of an LPXTG motif. The sortase may then assist in the formation of an amide link between the threonine carboxyl group and a cell wall precursor such as lipid II. The precursor can then be incorporated into the peptidoglycan via the transglycoslylation and transpeptidation reactions of bacterial wall synthesis. *See* Comfort et al., Infection & Immunity (2004) 72(5): 2710 – 2722. Typically, Gram positive bacteria AI surface proteins of the invention will contain an N-terminal leader or secretion signal to facilitate translocation of the surface protein across the bacterial membrane.

Gram positive bacteria AI surface proteins of the invention may affect the ability of the Gram positive bacteria to adhere to and invade target host cells, such as epithelial cells. Gram positive bacteria AI surface proteins may also affect the ability of the gram positive bacteria to translocate through an epithelial cell layer. Preferably, one or more of the Gram positive AI surface proteins are

PCT/US05/27239
capable of binding to or other associating with an epithelial cell surface. Further, one or more Gram positive AI surface proteins may bind to fibrinogen, fibronectin, or collagen protein.

In one embodiment, the invention includes a composition comprising oligomeric, pilus-like structures comprising a Gram positive bacteria AI surface protein. The oligomeric, pilus-like structure may comprise numerous units of the AI surface protein. Preferably, the oligomeric, pilus-like structures comprise two or more AI surface proteins. Still more preferably, the oligomeric, pilus-like structure comprises a hyper-oligomeric pilus-like structure comprising at least two (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 150, 200 or more) oligomeric subunits, wherein each subunit comprises an AI surface protein or a fragment thereof. The oligomeric subunits may be covalently associated via a conserved lysine within a pilin motif. The oligomeric subunits may be covalently associated via an LPXTG motif, preferably, via the threonine amino acid residue.

Gram positive bacteria AI surface proteins or fragments thereof to be incorporated into the oligomeric, pilus-like structures of the invention will preferably include one or both of a pilin motif comprising a conserved lysine residue and an E box motif comprising a conserved glutamic acid residue.

The oligomeric, pilus like structures may be used alone or in the combinations of the invention. In one embodiment, the invention comprises a Gram positive bacteria Adhesin Island in oligomeric form, preferably in a hyperoligomeric form.

The oligomeric, pilus-like structures of the invention may be combined with one or more additional Gram positive AI proteins (from the same or a different Gram positive species or genus). In one embodiment, the oligomeric, pilus-like structures comprise one or more Gram positive bacteria AI surface proteins in combination with a second Gram positive bacteria protein. The second Gram positive bacteria protein may be a known antigen, and need not normally be associated with an AI protein.

The oligomeric, pilus-like structures may be isolated or purified from bacterial cultures overexpressing a Gram positive bacteria AI surface protein. The invention therefore includes a method for manufacturing an oligomeric Adhesin Island surface antigen comprising culturing a Gram positive bacteria adapted for increased AI protein expression and isolation of the expressed oligomeric Adhesin Island protein from the Gram positive bacteria. The AI protein may be collected from secretions into the supernatant or it may be purified from the bacterial surface. The method may further comprise purification of the expressed Adhesin Island protein. Preferably, the Adhesin Island protein is in a hyperoligomeric form.

Gram positive bacteria are preferably adapted to increase AI protein expression by at least two (*e.g.*, 2, 3, 4, 5, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 125, 150 or 200) times wild type expression levels.

Gram positive bacteria may be adapted to increase AI protein expression by means known in the art, including methods of increasing gene dosage and methods of gene upregulation. Such means

PCT/US05/27239
include, for example, transformation of the Gram positive bacteria with a plasmid encoding the AI protein. The plasmid may include a strong promoter or it may include multiple copies of the sequence encoding the AI protein. Optionally, the sequence encoding the AI protein within the Gram positive bacterial genome may be deleted. Alternatively, or in addition, the promoter regulating the Gram positive Adhesin Island may be modified to increase expression.

The invention further includes Gram positive bacteria which have been adapted to produce increased levels of AI surface protein. In particular, the invention includes Gram positive bacteria which have been adapted to produce oligomeric or hyperoligomeric AI surface protein. In one embodiment, the Gram positive bacteria of the invention are inactivated or attenuated to permit *in vivo* delivery of the whole bacteria, with the AI surface protein exposed on its surface.

The invention further includes Gram positive bacteria which have been adapted to have increased levels of expressed AI protein incorporated in pili on their surface. The Gram positive bacteria may be adapted to have increased exposure of oligomeric or hyperoligomeric AI proteins on its surface by increasing expression levels of a signal peptidase polypeptide. Increased levels of a local signal peptidase expression in Gram positive bacteria (such us LepA in GAS) are expected to result in increased exposure of pili proteins on the surface of Gram positive bacteria. Increased expression of a leader peptidase in Gram positive may be achieved by any means known in the art, such as increasing gene dosage and methods of gene upregulation. The Gram positive bacteria adapted to have increased levels of leader peptidase may additionally be adapted to express increased levels of at least one pili protein.

Alternatively, the AI proteins of the invention may be expressed on the surface of a non-pathogenic Gram positive bacteria, such as *Streptococus gordonii* (See, e.g., Byrd et al., "Biological consequences of antigen and cytokine co-expression by recombinant *Streptococcus gordonii* vaccine vectors", Vaccine (2002) 20:2197-2205) or *Lactococcus lactis* (See, e.g., Mannam et al., "Mucosal VaccineMade from Live, Recombinant Lactococcus lactis Protects Mice against Pharangeal Infection with *Streptococcus pyogenes*" Infection and Immunity (2004) 72(6):3444-3450). It has already been demonstrated, above, that *L. lactis* expresses GBS and GAS AI polypeptides in oligomeric form and on its surface.

Alternatively, the oligomeric, pilus-like structures may be produced recombinantly. If produced in a recombinant host cell system, the Gram positive bacteria AI surface protein will preferably be expressed in coordination with the expression of one or more of the AI sortases of the invention. Such AI sortases will facilitate oligomeric or hyperoligomeric formation of the AI surface protein subunits.

Gram positive AI Sortases of the invention will typically have a signal peptide sequence within the first 70 amino acid residues. They may also include a transmembrane sequence within 50 amino acid residues of the C terminus. The sortases may also include at least one basic amino acid residue within the last 8 amino acids. Preferably, the sortases have one or more active site residues, such as a catalytic cysteine and histidine.

PCT/US2005/027239

Adhesin island surface proteins from two or more Gram positive bacterial genus or species may be combined to provide an immunogenic composition for prophylactic or therapeutic treatment of disease or infection of two more Gram positive bacterial genus or species. Optionally, the adhesin island surface proteins may be associated together in an oligomeric or hyperoligomeric structure.

In one embodiment, the invention comprises an adhesin island surface proteins from two or more *Streptococcus* species. For example, the invention includes a composition comprising a GBS AI surface protein and a GAS adhesin island surface protein. As another example, the invention includes a composition comprising a GAS adhesin island surface protein and a *S. pneumoniae* adhesin island surface protein.

In one embodiment, the invention comprises an adhesin island surface protein from two or more Gram positive bacterial genus. For example, the invention includes a composition comprising a *Streptococcus* adhesin island protein and a *Corynebacterium* adhesin island protein.

Examples of AI sequences in several Gram positive bacteria are discussed further below.

Streptococcus pyogenes (GAS)

As discussed above, Applicants have identified at least four different GAS Adhesin Islands. These adhesion islands are thought to encode surface proteins which are important in the bacteria's virulence, and Applicants have obtained the first electron micrographs revealing the presence of these adhesin island proteins in hyperoligomeric pilus structures on the surface of Group A Streptococcus.

Group A Streptococcus is a human specific pathogen which causes a wide variety of diseases ranging from pharyngitis and impetigo through life threatening invasive disease and necrotizing fasciitis. In addition, post-streptococcal autoimmune responses are still a major cause of cardiac pathology in children.

Group A Streptococcal infection of its human host can generally occur in three phases. The first phase involves attachment and/or invasion of the bacteria into host tissue and multiplication of the bacteria within the extracellular spaces. Generally this attachment phase begins in the throat or the skin. The deeper the tissue level infected, the more severe the damage that can be caused. In the second stage of infection, the bacteria secrete a soluble toxin that diffuses into the surrounding tissue or even systemically through the vasculature. This toxin binds to susceptible host cell receptors and triggers innappropropriate immune responses by these host cells, resulting in pathology. Because the toxin can diffuse throughout the host, the necrosis directly caused by the GAS toxins may be physically located in sites distant from the bacterial infection. The final phase of GAS infection can occur long after the original bacteria have been cleared from the host system. At this stage, the host's previous immune response to the GAS bacteria due to cross reactivity between epitopes of a GAS surface protein, M, and host tissues, such as the heart. A general review of GAS infection can be found in Principles of Bacterial Pathogeneis, Groisman ed., Chapter 15 (2001).

In order to prevent the pathogenic effects associated with the later stages of GAS infection, an effective vaccine against GAS will preferably facilitate host elimination of the bacteria during the initial attachment and invasion stage.

PCT/US05/27239

Isolates of Group A Streptococcus are historically classified according to the M surface protein described above. The M protein is surface exposed trypsin-sensitive protein generally comprising two polypeptide chains complexed in an alpha helical formation. The carboxyl terminus is anchored in the cytoplasmic membrane and is highly conserved among all group A streptococci. The amino terminus, which extends through the cell wall to the cell surface, is responsible for the antigenic variability observed among the 80 or more serotypes of M proteins.

A second layer of classification is based on a variable, trypsin-resistant surface antigen, commonly referred to as the T-antigen. Decades of epidemiology based on M and T serological typing have been central to studies on the biological diversity and disease causing potential of Group A Streptococci. While the M-protein component and its inherent variability have been extensively characterized, even after five decades of study, there is still very little known about the structure and variability of T-antigens. Antisera to define T types are commercially available from several sources, including Sevapharma (http://www.sevapharma.cz/en).

The gene coding for one form of T-antigen, T-type 6, from an M6 strain of GAS (D741) has been cloned and characterized and maps to an approximately 11 kb highly variable pathogenicity island. Schneewind et al., J Bacteriol. (1990) 172(6):3310 – 3317. This island is known as the Fibronectin-binding, Collagen-binding T-antigen (FCT) region because it contains, in addition to the T6 coding gene (*tee6*), members of a family of genes coding for Extra Cellular Matrix (ECM) binding proteins. Bessen et al., Infection & Immunity (2002) 70(3):1159-1167. Several of the protein products of this gene family have been shown to directly bind either fibronectin and/or collagen. *See* Hanski et al., Infection & Immunity (1992) 60(12):5119-5125; Talay et al., Infection & Immunity (1992(60(9):3837-3844; Jaffe et al. (1996) 21(2):373-384; Rocha et al., Adv Exp Med Biol. (1997) 418:737-739; Kreikemeyer et al., J Biol Chem (2004) 279(16):15850-15859; Podbielski et al., Mol. Microbiol. (1999) 31(4):1051-64; and Kreikemeyer et al., Int. J. Med Microbiol (2004) 294(2-3):177-88. In some cases direct evidence for a role of these proteins in adhesion and invasion has been obtained.

Applicants raised antiserum against a recombinant product of the *tee6* gene and used it to explore the expression of T6 in M6 strain ISS3650. In immunoblot of mutanolysin extracts of this strain, the antiserum recognized, in addition to a band corresponding to the predicted molecular mass of the *tee6* gene product, very high molecular weight ladders ranging in mobility from about 100 kDa to beyond the resolution of the 3-8% gradient gels used. See Figure 163A, last lane labeled "M6_Tee6."

This pattern of high molecular weight products is similar to that observed in immunoblots of the protein components of the pili identified in *Streptococcus agalactiae* (described above) and previously in *Corynebacterium diphtheriae*. Electron microscropy of strain M6 ISS3650 with antisera specific for the product of *tee6* revealed abundant surface staining and long pilus like structures extending up to 700 nanometers from the bacterial surface, revealing that the T6 protein, one of the

PCT/US2005/027239
antigens recognized in the original Lancefield serotyping system, is located within a GAS Adhesin Island (GAS AI-1) and forms long covalently linked pilus structures. See Figure 163I.

In addition to the *tee6* gene, the FCT region in M6_ISS3650 (GAS AI-1) contains two other genes (*prtF1* and *cpa*) predicted to code for surface exposed proteins; these proteins are characterized as containing the cell wall attachment motif LPXTG. Western blot analysis using antiserum specific for PrtF1 detected a single molecular species with electrophoretic mobility corresponding to the predicted molecular mass of the protein and one smaller band of unknown origin. Western blot analysis using antisera specific for Cpa recognized a high molecular weight covalently linked ladder (Fig 163A, second lane). Immunogold labelling of Cpa with specific antiserum followed by transmission electron microscopy detected an abundance of Cpa at the cell surface and only occasional structures extending from the cell surface (Fig. 163J).

Four classes of FCT region can be discerned by the types and order of the genes contained within the region. The FCT region of strains of types M3, M5, M18 and M49 have a similar organization whereas those of M6, M1 and M12 differ. See Figure 164. As discussed below, these four FCT regions correlate to four GAS Adhesin Island types (AI-1, AI-2, AI-3 and AI-4).

Applicants discovery of genes coding for pili in the FCT region of strain M6_ISS3650 prompted them to examine the predicted surface exposed proteins in the variant FCT regions of three other GAS strains of having different M-type (M1_SF370, M5_ISS4883 and M12_20010296) representing the other three FCT variants. Each gene present in the FCT region of each bacteria was cloned and expressed. Antisera specific for each recombinant protein was then used to probe mutanolysin extracts of the respective strains (*6*). In M1 strain SF370, there are three predicted surface proteins (Cpa (also referred to as M1_126 and GAS 15), M1_128 (a fimbrial protein also referred to as Spy0128 and GAS 16), and M1_130 (also referred to as Spy0130 and GAS 18)) (GAS AI-2). Antisera specific for each surface protein reacted with a ladder of high molecular weight material (Fig. 163B). Immunogold staining of M1 strain SF370 with antiserum specific for M1_128 revealed pili structures similar to those seen when M6 strain ISS3650 was immunogold stained with antiserum specific for tee6 (See Fig 1163K). Antisera specific for surface proteins Cpa and M1_130 revealed abundant surface staining and occasional structures extending from the surface of M1 strain SF370 bacteria (Fig. 163S).

The M1_128 protein appears to be necessary for polymerization of Cpa and M1_130 proteins. If the M1_128 gene in M1_SF370 was deleted, Western blot analysis using antibodies that hybridize to Cpa and M1_130 no longer detected high molecular weight ladders comprising the Cpa and M1_130 proteins (Fig. 163 E). See also Figures 177 A-C which provide the results of Western blot analysis of the M1_128 (Δ128) deleted bacteria using anti-M1_130 antiserum (Figure 177 A), anti-M1_128 antiserum (Figure 177 B), and anti-M1_126 antiserum (Figure 177

PCT/US05/27239

C). High molecular weight ladders, indicative of pilus formation on the surface of M1 strain SF370, could not be detected by any of the three antisera in Δ128 bacteria. If the Δ128 bacteria were transformed with a plasmid containing the gene for M1_128, Western blot analysis using antisera specific for Cpa and M1_130 again detected high molecular weight ladders (Figure 163 H).

In agreement with the Western blot analysis, immunoelectron microscopy failed to detect pilus assembly on the Δ128 strain SF370 bacteria using M1_128 antisera (Figure 178 B). Although Δ128 SF370 bacteria were unable to form pili, M1_126 (cpa) and M1_130, which contain sortase substrate motifs, were present on the bacteria's surface. FACS analysis of the M1_128 deleted (Δ128) strain SF370 bacteria also detected both M1_126 and M1_130 on the surface of the Δ128 strain SF370 bacteria. See Figure 179 D and F, which show a shift in fluorescence when antibodies immunoreactive to M1_126 and M1_130 are used on Δ128 bacteria. As expected, virtually no shift in fluorescence is observed when antibodies immunoreactive to M1_128 are used with the Δ128 bacteria (Figure 179 E).

By contrast, deletion of the M1_130 gene did not effect polymerization of M1_128 (Fig. 163 F). See also Figures 177 A-C, which provide Western blot analysis results of the M1_130 deleted (Δ130) strain SF370 bacteria using anti-M1_130 (Figure 177 A), anti-M1_128 (Figure 177 B), and anti-M1_126 antiserum (Figure 177 C). The anti-M1_128 and anti-M1_126 antiserum both detected the presence of high molecular weight ladders in the Δ130 strain SF370 bacteria, indicating that the Δ130 bacteria form pili that comprise M1_126 and M1_128 polypeptides in the absence of M1_130. As expected, the Western blot probed with antiserum immunoreactive with M1_130 did not detect any proteins for the Δ130 bacteria(Figure 177A).

Hence, the composition of the pili in GAS resembles that previously described for both *C. diphtheria* (*7, 8*) and *S. agalactiae* (described above) (*9*) in that each pilus is formed by a backbone component which abundantly stains the pili in EM and is essential for the incorporation of the other components.

Also similar to *C. diphtheria*, elimination of the srtC1 gene from the FCT region of M1_SF370 abolished polymerization of all three proteins and assembly of pili (Fig. 163 G). See also Figures 177 A-C, which provide Western blot analysis of the SrtC1 deleted (ΔSrtC1) strain SF370 bacteria using anti-M1_130 (Figure 177 A), anti-M1_128 (Figure 177 B), and anti-M1_126 antiserum (Figure 177 C). None of the three antisera immunoreacted with high molecular weight structures (pili) in the ΔSrtC1 bacteria. Confirming that deletion of the SrtC1 gene abrogates pilus assembly in strain SF370, immunoelectron microscopy using antisera against M1_128 failed to detect pilus formation on the bacteria surface. See Figure 178 C. Although no assembled pili were detected on ΔSrtC1 SF370, M1_128 proteins could be detected on the surface of SF370. Thus, it appeared that SrtC1 deletion prevented pilus assembly on the

PCT/US2005/027239

surface of the SF370 bacteria, but not anchoring of the proteins that comprise pili to the bacterial cell wall. FACS analysis of the ΔSrtC1 strain SF370 confirmed that deletion of SrtC1 does not eliminate cell surface expression of M1_126, M1_128 or M1_130. See Figure 179 G-I, which show a shift in fluorescence when antibodies immunoreactive to M1_126 (Figure 179 G), M1_128 (Figure 179 H), and M1_130 (Figure 179 I) are used to detect cell surface protein expression on ΔSrtC1 bacteria. Thus, SrtC1 deletion prevents pilus formation, but not surface anchoring of proteins involved in pilus formation on the surface of bacteria. Another sortase is possibly involved in anchoring of the proteins to the bacteria surface. Pilus polymerization in *C. diphtheriae* is also dependent on particular sortase enzyme whose gene resides at the same genetic locus as the pilus components (*7, 8*).

The LepA signal peptidase, Spy0127, also appears to be essential for pilus assembly in strain SF370. LepA deletion mutants (ΔLepA) of strain SF370 fail to assemble pili on the cell surface. Not only are the ΔLepA mutants unable to assemble pili, they are also deficient at cell surface M1 expression. See Figure 180, which provides a FACS analysis of the wildtype (A) and ΔLepA mutant (B) SF370 bacteria using M1 antisera. No shift in fluorescence is observed for the ΔLepA mutant bacteria in the presence of M1 immune serum. It is possible that these deletion mutants of LepA will be useful for detecting non-M, non-pili, surface exposed antigens on the surface of GAS, or any Gram positive bacteria. These antigens may also be useful in immunogenic compositions.

Pili were also observed in M5 strain ISS4882 and M12 strain 20010296. The M5 strain ISS4882 contains genes for four predicted surface exposed proteins (GAS AI-3). Antisera against three of the four products of the FCT region (GAS AI-3) of M5_ISS4883 (Cpa, M5_orf80, M5_orf82) stained high molecular weight ladders in Western blot analysis (Figure 163 C). Long pili were visible when antisera against M5_orf80 was used in immunogold staining followed by electron microscopy (Figure 163L).

The M12 strain 20010296 contains genes for five predicted surface exposed proteins. (GAS AI-4) Antisera against three of the five products of the FCT region (GAS AI-4) of M12_20010296 (Cpa, EftLSL.A, Orf2) stained high molecular weight ladders in Westen blot analysis (Figure 163 D). Long pili were visible when antisera against EftLSL.A were used (Fig. 163 M).

The major pilus forming proteins identified in the four strains studied by applicants (T6, M1_128, M5_orf80 and EftLSL.A) share between 23% and 65% amino acid identity in any pairwise comparison, indicating that each pilus may represent a different Lancefield T-antigen. Each pilus is part of a trypsin resistant structure on the GAS bacteria surface, as is the case for the Lancefield T-antigens. See Figure 165, which provides a FACS analysis of bacteria harboring each of the FCT types that had or had not been treated with trypsin (*6*). Following treatment, surface expression of the

PCT/US2005/027239

pilus proteins was assayed by indirect immunofluorescence and flow cytometry using antibodies specific for the pilus proteins, the bacteria's respective M proteins, or surface proteins not associated with the pili (Figure 165). Staining the cells with sera specific for proteins associated with the pili was not effected by trypsin treatment, whereas trypsin treatment substantially reduced detection of M-proteins or surface proteins not associated with pili.

The pili structures identified on the surface of the GAS bacteria were confirmed to be Lancefield T antigens when commercially available T-serotyping sera detected the pili on the surface of bacteria. Western blot analysis was initially performed to determine if polyvalent serum pools (designated T, U, W, X, andY) could detect recombinant proteins for each of the major pilis components (T6, M1_128, M5_orf80 and EftLSL.A) identified in the strains of bacteria discussed above. Pool U, which contains the T6 serum, recognized the T6 protein specifically (a surface exposed pilus protein from GAS AI-1)(Fig. 166 B). Pool T specifically recognized M1_128 (a surface exposed pilus protein from GAS AI-2) (Fig. 166 A). Pool W recognized both M5_orf80 and EftLSL.A (Fig. 166 C). Using monovalent sera representative of each of the components of each polyvalent pool, applicants confirmed the specificity of the T6 antigen (corresponding to a surface exposed pilus protein from GAS AI-1)(Fig. 166 E) and identified M1_128 as antigen T1 (corresponding to a surface exposed pilus protein from GAS AI-2) (Fig. 166 D), EftLSL.A as antigen T12 (corresponding to a surface exposed pilus protein from GAS AI-4) (Fig. 166 G) and M5_orf80 as a common antigen recognized by the related sera T5, T27 and T44 (corresponding to a surface exposed pilus protein from GAS AI-3).

Confirming applicants observations, discussed above, that deleting the M1_128 gene from M1_SF370 abolishes pilus formation, the pool T sera stained whole M1_SF370 bacteria (Fig. 166 H) but failed to stain M1_SF370 bacteria lacking the M1_128 gene (Fig. 166 I).

As discussed above, Applicants have identified at least four different Group A Streptococcus Adhesin Islands. While these GAS AI sequences can be identified in numerous M types, Applicants have surprisingly discovered a correlation between the four main pilus subunits from the four different GAS AI types and specific T classifications. While other trypsin-resistant surface exposed proteins are likely also implicated in the T classification designations, the discovery of the role of the GAS adhesin islands (and the associated hyper-oligomeric pilus like structures) in T classification and GAS serotype variance has important implications for prevention and treatment of GAS infections. Applicants have identified protein components within each of the GAS adhesin islands which are associated with the pilus formation. These proteins are believed to be involved in the bacteria's initial adherence mechanisms. Immunological recognition of these proteins may allow the host immune response to slow or prevent the bacteria's transition into the more pathogenic later stages of infection. In addition, the GAS pili may be involved in formation of biofilms. Applicants have discovered that the GBS pili structures appear to be implicated in the formation of biofilms (populations of bacteria growing on a surface, often enclosed in an exopolysaccharide matrix). Biofilms are generally associated with bacterial resistance, as antibiotic treatments and host immune response are frequently

PCT/US2005/027239

unable to erradicate all of the bacteria components of the biofilm. Direction of a host immune response against surface proteins exposed during the first steps of bacterial attachment (i.e., before complete biofilm formation) is preferable.

The invention therefore provides for improved immunogenic compositions against GAS infection which may target GAS bacteria during their initial attachment efforts to the host epithelial cells and may provide protection against a wide range of GAS serotypes. The immunogenic compositions of the invention include GAS AI surface proteins which may be formulated in an oligomeric, or hyperoligomeric (pilus) form. The invention also includes combinations of GAS AI surface proteins. Combinations of GAS AI surface proteins may be selected from the same adhesin island or they may be selected from different GAS adhesin islands.

The invention comprises compositions comprising a first GAS AI protein and a second GAS AI protein wherein the first and second GAS AI proteins are derived from different GAS adhesin islands. For example, the invention includes a composition comprising at least two GAS AI proteins wherein the GAS AI proteins are encoded by the adhesin islands selected from the group consisting of GAS AI-1 and AI-2; GAS AI-1 and GAS AI-3; GAS AI-1 and GAS AI-4; GAS AI-2 and GAS AI-3; GAS AI-2 and GAS AI-4; and GAS AI-3 and GAS AI-4. Preferably the two GAS AI proteins are derived from different T-types.

A schematic arrangement of GAS Adhesin Island sequences is set forth in FIGURE 162. In all strains, the AI region is flanked by the highly conserved open reading frames M1_123 and M1-136. Between three and five genes in each locus code for surface proteins containing LPXTG motifs. These surface proteins also all belong to the family of genes coding for ECM binding adhesins.

Adhesin island sequences can be identified in numerous M types of Group A Streptococcus. Examples of AI sequences within M1, M6, M3, M5, M12, M18, and M49 serotypes are discussed below.

GAS Adhesin Islands generally include a series of open reading frames within a GAS genome that encode for a collection of surface proteins and sortases. A GAS Adhesin Island may encode for amino acid sequences comprising at least one surface protein. Alternatively, a GAS Adhesin Island may encode for at least two surface proteins and at least one sortase. Preferably, a GAS Adhesin Island encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. One or more GAS AI surface proteins may participate in the formation of a pilus structure on the surface of the Gram positive bacteria.

GAS Adhesin Islands of the invention preferably include a divergently transcribed transcriptional regulator. The transcriptional regulator may regulate the expression of the GAS AI operon. Examples of transcriptional regulators found in GAS AI sequences include *RofA* and *Nra*.

The GAS AI surface proteins may bind or otherwise adhere to fibrinogen, fibronectin, or collagen. One or more of the GAS AI surface proteins may comprise a fimbrial structural subunit.

PCT/US05/27239

One or more of the GAS AI surface proteins may include an LPXTG motif or other sortase substrate motif. The LPXTG motif may be followed by a hydrophobic region and a charged C terminus, which are thought to retard the protein in the cell membrane to facilitate recognition by the membrane-localized sortase. See Barnett, et al., J. Bacteriology (2004) 186 (17): 5865-5875.

GAS AI sequences may be generally categorized as Type 1, Type 2, Type 3, or Type 4, depending on the number and type of sortase sequences within the island and the percentage identity of other proteins (with the exception of RofA and cpa) within the island. Figure 167 provides a chart indicating the number and type of sortase sequences identified within the adhesin islands of various strains and serotypes of GAS. As can be seen in this figure, all GAS strains and serotypes thus far characterized as an AI-1 have a SrtB type sortase, all GAS strains and serotypes thus far characterized as an AI-2 have SrtB and SrtC1 type sortases, all GAS strains and serotypes thus far characterized as an AI-3 have a SrtC2 type sortase, and all GAS strains and serotypes thus far characterized as an AI-4 have SrtB and SrtC2 type sortases. A comparison of the percentage identity of sequences within the adhesin islands was presented in Table 45, see above.

(1) Adhesin Island sequence within M6: GAS Adhesin Island 1 ("GAS AI-1")

A GAS Adhesin Island within M6 serotype (MGAS10394) is outlined in Table 4 below. This GAS adhesin island 1 ("GAS AI-1") comprises surface proteins, a srtB sortase and a rofA divergently transcribed transcriptional regulator.

GAS AI-1 surface proteins include Spy0157 (a fibronectin binding protein), Spy0159 (a collagen adhesion protein) and Spy0160 (a fimbrial structural subunit). Preferably, each of these GAS AI-1 surface proteins includes an LPXTG sortase substrate motif, such as LPXTG (SEQ ID NO: 122) or LPXSG (SEQ ID NO: 134) (conservative replacement of threonine with serine).

GAS AI-1 includes a srtB type sortase. GAS srtB sortases may preferably anchor surface proteins with an LPSTG motif (SEQ ID NO: 166), particularly where the motif is followed by a serine.

Table 4: GAS AI-1 sequences from M6 isolate (MGAS10394)

AI-1 sequence identifier	Sortase substrate sequence or sortase type	functional description
M6_Spy0156		Transcriptional regulator (*rofA*)
M6_Spy0157	LPXTG	Fibronectin-binding protein
M6_Spy0158		Reverse transcriptase
M6_Spy0159	LPXSG	Collagen adhesion protein
M6_Spy0160	LPXTG	Fimbrial structural subunit
M6_Spy0161	srtB	Sortase

M6_Spy0160 appears to be present on the surface of GAS as part of oligomeric (pilus) structures. Figures 127-132 present electron micrographs of GAS serotype M6 strain 3650 immunogold stained for M6_Spy0160 using anti-M6_Spy0160 antiserum. Oligomeric or

PCT/US05/27239

hyperoligmeric structures labelled with gold particles can be seen extending from the surface of the GAS in each of these figures, indicating the presence of multiple M6_Spy0160 polypeptides in the oligomeric or hyperoligomeric structures. Figure 176 A-F present electron micrographs of GAS M6 strain 2724 immunogold stained for M6_Spy0160 using anti-M6_Spy0160 antiserum (Figures 176 A-E) or immunogold stained for M6_Spy0159 using anti-M6_Spy0159 antiserum (Figure 176 F). Oligomeric or hyperoligomeric structures labelled with gold particles can again be seen extending from the surface of the M6 strain 2724 GAS bacteria immunogold stained for M6_Spy0160. M6_Spy0159 is also detected on the surface of the M6 strain 2724 GAS.

FACS analysis has confirmed that the GAS AI-1 surface proteins spyM6_0159 and spyM6_0160 are indeed expressed on the surface of GAS. Figure 73 provides the results of FACS analysis for surface expression of spyM6_0159 on each of GAS serotypes M6 2724, M6 3650, and M6 2894. A shift in fluorescence is observed for each GAS serotype when anti-spyM6_0159 antiserum is present, demonstrating cell surface expression. Table 18, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-spyM6_0159 antiserum, and the difference in fluorescence value between the pre-immune and anti-spyM6_0159 antiserum.

Table 18; Summary of FACS values for surface expression of spyM6_0159

2724			3650			2894		
Pre-immune	Anti-spyM6_0159	Change	Pre-immune	Anti-spyM6_0159	Change	Pre-immune	Anti-spyM6_0159	Change
134.84	427.48	293	149.68	712.62	563	193.86	597.8	404

Figure 74 provides the results of FACS analysis for surface expression of spyM6_0160 on each of GAS serotypes M6 2724, M6 3650, and M6 2894. In the presence of of anti-spyM6_0160 antiserum, a shift in fluorescence is observed for each GAS serotype, which demonstrates its cell surface expression. Table 19, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-spyM6_0160 antiserum, and the change in fluorescence value between the pre-immune and anti-spyM6_0160 antiserum.

Table 19: Summary of FACS values for surface expression of spyM6_0160

2724			3650			2894		
Pre-immune	Anti-spyM6_0160	change	Pre-immune	Anti-spyM6_0160	change	Pre-immune	Anti-spyM6_0160	change
117.12	443.24	326	128.57	776.39	648	125.87	621.17	495

Surface expression of M6_Spy0159 and M6_Spy0160 on M6 serotype GAS has also been confirmed by Western blot analysis. Figure 98 shows that while pre-immune sera (P α-0159) does not detect expression of M6_Spy0159 in GAS serotype M6, anti-M6_Spy0159 immune sera (I α-0159) is able to detect M6_Spy0159 protein in both total GAS M6 extracts (M6 tot) and GAS M6 fractions enriched for cell surface proteins (M6 surf prot). The M6_Spy0159 proteins detected in the

PCT/US05/27239

total GAS M6 extracts or the GAS M6 extracts enriched for surface proteins are also present as high molecular weight structures, indicating that M6_Spy0159 may be in an oligomeric (pilus) form.

Figure 112 shows that while preimmune sera (Preimmune Anti 106) does not detect expression of M6_Spy0160 in GAS serotype M6 strain 2724, anti-M6_Spy0160 immune sera (Anti 160) does in both total GAS M6 strain 2724 extracts (M6 2724 tot) and GAS M6 strain 2724 fractions enriched for surface proteins. The M6_Spy0160 proteins detected in the total GAS M6 strain 2724 extracts or the GAS M6 strain 2724 extracts enriched for surface proteins are also present as high molecular weight structures, indicating that M6_Spy0160 may be in an oligomeric (pilus) form.

Figures 110 and 111 both further verify the presence of M6_Spy0159 and M6_Spy0160 in higher molecular weight structures on the surface of GAS. Figure 110 provides a Western blot performed to detect M6_Spy0159 and M6_Spy0160 in GAS M6 strain 2724 extracts enriched for surface proteins. Antiserum raised against either M6_Spy0159 (Anti-159) or M6_Spy0160 (Anti-160) cross-hybridizes with high molecular weight structures (pili) in these extracts. Figure 111 provides a similar Western blot that verifies the presence of M6_ Spy0159 and M6_Spy0160 in high molecular weight structures in GAS M6 strain 3650 extracts enriched for surface proteins.

SpyM6_0157 (a fibronectin-binding protein) may also be expressed on the surface of GAS serotype M6 bacteria. Figure 174 shows the results of FACS analysis for surface expression of spyM6_0157 on M6 strain 3650. A slight shift in fluorescence is observed, which demonstrates that some spyM6_0157 may be expressed on the GAS cell surface.

Adhesin Island sequence within M6: GAS Adhesin Island 2 ("GAS AI-2")

A GAS Adhesin Island within M1 serotype (SF370) is outlined in Table 5 below. This GAS adhesin island 2 ("GAS AI-2") comprises surface proteins, a SrtB sortase, a SrtC1 sortase and a RofA divergently transcribed transcriptional regulator.

GAS AI-2 surface proteins include GAS 15 (Cpa), Spy0128 (thought to be a fimbrial protein) and Spy0130 (a hypothetical protein). Preferably, each of these GAS AI-2 surface proteins includes an LPXTG sortase substrate motif, such as LPXTG (SEQ ID NO: 122), VVXTG (SEQ ID NO: 135), or EVXTG (SEQ ID NO: 136).

GAS AI-2 includes a srtB type sortase and a srtC1 sortase. As discussed above, GAS SrtB sortases may preferably anchor surface proteins with an LPSTG (SEQ ID NO: 166) motif, particularly where the motif is followed by a serine. GAS SrtC1 sortase may preferentially anchor surface proteins with a V(P/V)PTG (SEQ ID NO: 167) motif. GAS SrtC1 may be differentially regulated by RofA.

GAS AI-2 may also include a LepA putative signal peptidase I protein.

Table 5 : GAS AI-2 sequence from M1 isolate (SF370)

AI-2 sequence identifier	Sortase substrate sequence or sortase type	functional description
SPy0124		rofA regulatory protein
GAS15(not annotated in SF370)	VVXTG	cpa

PCT/US05/27239

SPy0127		LepA putative signal peptidase I
SPy0128 (GAS16)	EVXTG	hypothetical protein (fimbrial)
SPy0129 (GAS17)	srtC1	sortase
SPy0130 (GAS18)	LPXTG	hypothetical protein
SPy0131		conserved hypothetical protein
SPy0133		conserved hypothetical protein
SPy0135 (GAS20)	srtB	sortase (putative fimbrial-associated protein)

GAS 15, GAS 16, and GAS 18 appear to be present on the surface of GAS as part of oligomeric (pilus) structures. Figures 113-115 present electron micrographs of GAS serotype M1 strain SF370 immunogold stained for GAS 15 using anti-GAS 15 antiserum. Figures 116-121 provide electron micrographs of GAS serotype M1 strain SF370 immunogold stained for GAS 16 using anti-GAS 16 antiserum. Figures 122-125 present electron micrograph of GAS serotype M1 strain SF370 immunogold stained for GAS 18 using anti-GAS 18 antiserum. Oligomers of these proteins can be seen on the surface of SF370 bacteria in the immuno-gold stained micrographs.

Figure 126 reveals a hyperoligomer on the surface of a GAS serotype M1 strain SF370 bacterium immunogold stained for GAS 18. This long hyperoliogmeric structure comprising GAS 18 stretches far out into the supernatant from the surface of the bacteria.

FACS analysis has confirmed that the GAS AI-2 surface proteins GAS 15, GAS 16, and GAS 18 are expressed on the surface of GAS. Figure 75 provides the results of FACS analysis for surface expression of GAS 15 on each of GAS serotypes M1 2719, M1 2580, M1 3280, M1 SF370, M1 2913, and M1 3348. A shift in fluorescence is observed for each GAS serotype when anti-GAS 15 antiserum is present, demonstrating cell surface expression. Table 20, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-GAS 15 antiserum, and the difference in fluorescence value between the pre-immune and anti-GAS 15 antiserum.

Table 20: Summary of FACS values for surface expression of GAS 15

2719			2580			3280		
Pre-immune	Anti-GAS 15	Change	Pre-immune	Anti-GAS 15	Change	Pre-immune	Anti-GAS 15	Change
159.46	712.71	553	123.9	682.84	559	217.02	639.69	423
SF370			2913			3348		
Pre-immune	Anti-GAS 15	Change	Pre-immune	Anti-GAS 15	Change	Pre-immune	Anti-GAS 15	Change
201.93	722.68	521	121.41	600.45	479	152.09	446.41	294

Figures 76 and 79 provide the results of FACS analysis for surface expression of GAS 16 on each of GAS serotypes M1 2719, M1 2580, M1 3280, M1 SF370, M1 2913, and M1 3348. The FACS data in Figure 76 was obtained using antisera was raised against full length GAS 16. In the presence of this anti-GAS 16 antiserum, a shift in fluorescence is observed for each GAS serotype,

demonstrating its cell surface expression. Table 21, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-GAS 16 antiserum, and the change in fluorescence value between the pre-immune and anti-GAS 16 antiserum.

Table 21: Summary of FACS values for surface expression of GAS 16

2719			2580			3280		
Pre-immune	Anti-GAS 16	Change	Pre-immune	Anti-GAS 16	Change	Pre-immune	Anti-GAS 16	Change
233.27	690.09	457	133.82	732.29	598	264.47	649.43	385
SF370			2913			3348		
Pre-immune	Anti-GAS 16	Change	Pre-immune	Anti-GAS 16	Change	Pre-immune	Anti-GAS 16	Change
237.2	727.46	490	138.52	588.04	450	180.56	420.93	240

The FACS data in Figure 79 was obtained using antisera was raised against a truncated GAS 16, which is encoded by SEQ ID NO: 179, shown below.

SEQ ID NO: 179:

```
GCTACAACAGTTCACGGGGAGACTGTTGTAAACGGAGCCAAACTAACAGTTACAAAAAACCTTGATTTAGTTAAT
AGCAATGCATTAATTCCAAATACAGATTTTACATTTAAAATCGAACCTGATACTACTGTCAACGAAGACGGAAAT
AAGTTTAAAGGTGTAGCTTTGAACACACCGATGACTAAAGTCACTTACACCAATTCAGATAAAGGTGGATCAAAT
ACGAAAACTGCAGAATTTGATTTTTCAGAAGTTACTTTTGAAAAACCAGGTGTTTATTATTACAAAGTAACTGAG
GAGAAGATAGATAAAGTTCCTGGTGTTTCTTATGATACAACATCTTACACTGTTCAAGTTCATGTCTTGTGGAAT
GAAGAGCAACAAAAACCAGTAGCTACTTATATTGTTGGTTATAAAGAAGGTAGTAAGGTGCCAATTCAGTTCAAA
AATAGCTTAGATTCTACTACATTAACGGTGAAGAAAAAAGTTTCAGGTACCGGTGGAGATCGCTCTAAAGATTTT
AATTTTGGTCTGACTTTAAAAGCAAATCAGTATTATAAGGCGTCAGAAAAAGTCATGATTGAGAAGACAACTAAA
GGTGGTCAAGCTCCTGTTCAAACAGAGGCTAGTATAGATCAACTCTATCATTTTACCTTGAAAGATGGTGAATCA
ATCAAAGTCACAAATCTTCCAGTAGGTGTGGATTATGTTGTCACTGAAGACGATTACAAATCAGAAAAATATACA
ACCAACGTGGAAGTTAGTCCTCAAGATGGAGCTGTAAAAAATATCGCAGGTAATTCAACTGAACAAGAGACATCT
ACTGATAAAGATATGACCATTACTTTTACAAATAAAAAAGATTT
```

In the presence of this anti-GAS 16 antiserum, a shift in fluorescence is observed for each GAS serotype, demonstrating its cell surface expression. Table 22, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-GAS 16 antiserum, and the change in fluorescence value between the pre-immune and anti-GAS 16 antiserum.

Table 22: Summary of FACS values for surface expression of GAS 16 using a second antisera

2719			2580			3280		
Pre-immune	Anti-GAS 16	Change	Pre-immune	Anti-GAS 16	Change	Pre-immune	Anti-GAS 16	Change
141.55	650.22	509	119.57	672.35	553	209.18	666.71	458
SF370			2913			3348		
Pre-immune	Anti-GAS 16	Change	Pre-immune	Anti-GAS 16	Change	Pre-immune	Anti-GAS 16	Change
159.92	719.32	559	115.97	585.9	470	146.1	414.01	268

PCT/US05/27239

Figures 77 and 78 provide the results of FACS analysis for surface expression of GAS 18 on each of GAS serotypes M1 2719, M1 2580, M1 3280, M1 SF370, M1 2913, and M1 3348. The antiserum used to obtain the FACS data in each of Figures 77 and 78 was different, although each was raised against full length GAS 18. In the presence of each of the anti-GAS 18 antisera, a shift in fluorescence is observed for each GAS serotype, demonstrating its cell surface expression. Tables 23 and 24, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, first or second anti-GAS 18 antiserum, and the change in fluorescence value between the pre-immune and first or second anti-GAS 18 antiserum.

Table 23: Summary of FACS values for surface expression of GAS 18

2719			2580			3280		
Pre-immune	Anti-GAS 18	Change	Pre-immune	Anti-GAS 18	Change	Pre-immune	Anti-GAS 18	Change
135.68	327.98	192	116.32	379.41	263	208.12	380.84	173
SF370			2913			3348		
Pre-immune	Anti-GAS 18	Change	Pre-immune	Anti-GAS 18	Change	Pre-immune	Anti-GAS 18	Change
185.39	438.23	253	119.95	373.32	253	147.12	266.51	119

Table 24: Summary of FACS values for surface expression of GAS 18 using a second antisera

2719			2580			3280		
Pre-immune	Anti-GAS 18	Change	Pre-immune	Anti-GAS 18	Change	Pre-immune	Anti-GAS 18	Change
150.4	250.39	100	139.18	386.38	247	253.38	347.72	94
SF370			2913			3348		
Pre-immune	Anti-GAS 18	Change	Pre-immune	Anti-GAS 18	Change	Pre-immune	Anti-GAS 18	Change
188.64	373.11	184	124.94	384.82	260	168.8	213.65	45

Surface expression of GAS 15, GAS 16, and GAS 18 on M1 serotype GAS has also been confirmed by Western blot analysis. Figure 89 shows that while pre-immune sera does not detect GAS M1 expression of GAS 15, anti-GAS 15 immune sera is able to detect GAS 15 protein in both total GAS M1 extracts and GAS M1 proteins enriched for cell surface proteins. The GAS 15 proteins detected in the M1 extracts enriched for surface proteins are also present as high molecular weight structures, indicating that GAS 15 may be in an oligomeric (pilus) form. Figure 90 also shows the results of Western blot analysis of M1 serotype GAS using anti-GAS 15 antisera. Again, the lanes that contain GAS M1 extracts enriched for surface proteins (M1 prot sup) show the presence of high molecular weight structures that may be oligomers of GAS 15. Figure 91 provides an additional Western blot identical to that of Figure 90, but that was probed with pre-immune sera. As expected, no proteins were detected on this membrane.

Figure 92 provides a Western blot that was probed for GAS 16 protein. While pre-immune sera does not detect GAS M1 expression of GAS 16, anti-GAS 16 immune sera is able to detect GAS

PCT/US2005/027239

16 protein in GAS M1 extracts enriched for cell surface proteins. The GAS 16 proteins detected in the M1 extracts enriched for surface proteins are present as high molecular weight structures, indicating that GAS 16 may be in an oligomeric (pilus) form. Figure 93 also shows the results of Western blot analysis of M1 serotype GAS using anti-GAS 16 antisera. The lanes that contain total GAS M1 protein (M1 tot new and M1 tot old) and the lane that contains GAS M1 extracts enriched for surface proteins (M1 prot sup) show the presence of high molecular weight structures that may be oligomers of GAS 16. Figure 94 provides an additional Western blot identical to that of Figure 93, but that was probed with pre-immune sera. As expected, no proteins were detected on this membrane.

Figure 95 provides a Western blot that was probed for GAS 18 protein. While pre-immune sera does not detect GAS M1 expression of GAS 18, anti-GAS 18 immune sera is able to detect GAS 18 protein in GAS M1 extracts enriched for cell surface proteins. The GAS 18 proteins detected in the M1 extracts enriched for surface proteins are present as high molecular weight structures, indicating that GAS 18 may be in an oligomeric (pilus) form. Figure 96 also shows the results of Western blot analysis of M1 serotype GAS using anti-GAS 18 antisera. The lane that contains GAS M1 extracts enriched for surface proteins (M1 prot sup) show the presence of high molecular weight structures that may be oligomers of GAS 18. Figure 97 provides an additional Western blot identical to that of Figure 96, but that was probed with pre-immune sera. As expected, no proteins were detected on this membrane.

Figures 102-106 provide additional Western blots to verify the presence of GAS 15, GAS 16, and GAS 18 in high molecular weight structures in GAS. Each Western blot was performed using proteins from a different GAS M1 strain, 2580, 2913, 3280, 3348, and 2719. Each Western blot was probed with antisera raised against each of GAS 15, GAS 16, and GAS 18. As can be seen in Figures 102-106, none of the Western blots shows detection of proteins using pre-immune serum (Pα-158, Pα-15, Pα-16, or Pα-18), while each Western blot shows cross-hybridization of the GAS 15 (Iα-15), GAS 16 (Iα-16), and GAS 18 (Iα-18) antisera to high molecular weight structures. Thus, these Western blots confirm that GAS 15, GAS 16, and GAS 18 can be present in pili in GAS M1.

Figure 107 provides a similar Western blot performed to detect GAS 15, GAS 16, and GAS 18 proteins in a GAS serotype M1 strain SF370 protein fraction enriched for surface proteins. This Western blot also shows detection of GAS 15 (Anti-15), GAS 16 (Anti-16), and GAS 18 (Anti-18) as high molecular weight structures.

(3) Adhesin Island sequence within M3, M5, and M18: GAS Adhesin Island 3 ("GAS AI-3")

GAS Adhesin Island sequences within M3, M5, and M18 serotypes are outlined in Tables 6 – 8 and 10 below. This GAS adhesin island 3 ("GAS AI-3") comprises surface proteins, a SrtC2 sortase, and a Negative transcriptional regulator (Nra) divergently transcribed transcriptional regulator.

GAS AI-3 surface proteins within include a collagen binding protein, a fimbrial protein, a F2 like fibronectin-binding protein. GAS AI-3 surface proteins may also include a hypothetical surface

PCT/US2005/027239

protein. Preferably, each of these GAS AI-3 surface proteins include an LPXTG sortase substrate motif, such as LPXTG (SEQ ID NO: 122), VPXTG (SEQ ID NO: 137), QVXTG (SEQ ID NO: 138) or LPXAG (SEQ ID NO: 139).

GAS AI-3 includes a SrtC2 type sortase. GAS SrtC2 type sortases may preferably anchor surface proteins with a QVPTG (SEQ ID NO: 140) motif, particularly when the motif is followed by a hydrophobic region and a charged C terminus tail. GAS SrtC2 may be differentially regulated by Nra.

GAS AI-3 may also include a LepA putative signal peptidase I protein.

GAS AI-3 may also include a putative multiple sugar metabolism regulator.

Table 6: GAS AI-3 sequences from M3 isolate (MGAS315)

AI-3 sequence identifier	Sortase substrate sequence or sortase type	Functional description
SpyM3_0097		Negative transcriptional regulator (Nra)
SpyM3_0098	VPXTG	putative collagen binding protein (Cpb)
SpyM3_0099		LepA putative signal peptidase I
SpyM3_0100	QVXTG	conserved hypothetical protein (fimbrial)
SpyM3_0101	SrtC2	sortase
SpyM3_0102	LPXAG	hypothetical protein
SpyM3_0103		putative multiple sugar metabolism regulator
SpyM3_0104	LPXTG	protein F2 like fibronectin-binding protein

Table 7: GAS AI-3 sequence from M3 isolate (SSI-1)

AI-3 sequence identifier	Sortase Substrate sequence or sortase type	Functional description
SPs0099		Negative transcriptional regulator (Nra)
SPs0100	VPXTG	putative collagen binding protein (Cpb)
SPs0101		LepA putative signal peptidase I
SPs0102	QVXTG	conserved hypothetical protein (fimbrial)
SPs0103	SrtC2	sortase
SPs0104	LPXAG	hypothetical protein
SPs0105		putative multiple sugar metabolism regulator
SPs0106	LPXTG	protein F2 like fibronectin-binding protein

Table 10: GAS AI-3 sequences from M5 isolate (Manfredo)

AI-3 sequence identifier	Sortase substrate sequence or sortase type	Functional description
orf77		Negative transcriptional regulator (Nra)
orf78	VPXTG	putative collagen binding protein (Cpb)
orf79		LepA putative signal peptidase I
orf80	QVXTG	conserved hypothetical protein (fimbrial)
orf81	SrtC2	sortase

orf82	LPXAG	hypothetical protein
orf83		putative multiple sugar metabolism regulator
orf84	LPXTG	protein F2 like fibronectin-binding protein

Table 8: GAS AI-3 sequences from M18 isolate (MGAS8232)

AI-3 sequence identifier	Sortase substrate sequence or sortase type	Functional description
spyM18_0125		Negative transcriptional regulator (Nra) (N-terminal fragment)
spyM18_0126	VPXTG	putative collagen binding protein (Cpb)
spyM18_0127		LepA putative signal peptidase I
spyM18_0128	QVXTG	conserved hypothetical protein (fimbrial)
spyM18_0129	SrtC2	sortase
spyM18_0130	LPXAG	hypothetical protein
spyM18_0131		putative multiple sugar metabolism regulator
spyM18_0132	LPXTG	protein F2 like fibronectin-binding protein

Table 44: GAS AI-3 sequences from M49 isolate (591)

AI-3 sequence identifier	Sortase substrate sequence or sortase type	Functional description
SpyoM01000156		Negative transcriptional regulator (Nra)
SpyoM01000155	VPXTG	collagen binding protein (Cpa)
SpyoM01000154		LepA putative signal peptidase I
SpyoM01000153	QVXTG	conserved hypothetical protein (fimbrial)
SpyoM01000152	SrtC2	sortase
SpyoM01000151	LPXAG	hypothetical protein
SpyoM01000150		MsmRL
SpyoM01000149	LPXTG	protein F2 like fibronectin-binding protein

A schematic of AI-3 serotypes M3, M5, M18, and M49 is shown in Figure 51A. Each contains an open reading frame encoding a SrtC2-type sortase of nearly identical amino acid sequence. See Figure 52B for an amino acid sequence alignment for each of the SrtC2 amino acid sequences.

The protein F2-like fibronectin-binding protein of each these type 3 adhesin islands contains a pilin motif and an E-box. Figure 60 indicates the amino acid sequence of the pilin motif and E-box of each of GAS AI-3 serotype M3 MGAS315 (SpyM3_0104/21909640), GAS AI-3 serotype M3 SSI (Sps0106/28895018), GAS AI-3 serotype M18 (SpyM18_0132/19745307), and GASAI-3 serotype M5 (orf84).

FACS analysis has confirmed that the GAS AI-3 surface proteins SpyM3_0098, SpyM3_0100, SpyM3_0102, and SpyM3_0104 are expressed on the surface of GAS. Figure 80 provides the results of FACS analysis for surface expression of SpyM3_0098 on each of GAS

serotypes M3 2721 and M3 3135. A shift in fluorescence is observed for each GAS serotype when anti-SpyM3_0098 antiserum is present, demonstrating cell surface expression. Table 25, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-SpyM3_0098 antiserum, and the difference in fluorescence value between the pre-immune and anti-SpyM3_0098 antiserum.

Table 25: Summary of FACS values for surface expression of SpyM3_0098

2721			3135		
Pre-immune	Anti-spyM3_0098	Change	Pre-immune	Anti-spyM3_0098	Change
117.85	249.51	132	99.17	277.21	178

Figure 81 provides the results of FACS analysis for surface expression of SpyM3_0100 on each of GAS serotypes M3 2721 and M3 3135. A shift in fluorescence is observed for each GAS serotype when anti-SpyM3_0100 antiserum is present, demonstrating cell surface expression. Table 26, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-SpyM3_0100 antiserum, and the difference in fluorescence value between the pre-immune and anti-SpyM3_0100 antiserum.

Table 26: Summary of FACS values for surface expression of SpyM3_0100

2721			3135		
Pre-immune	Anti-spyM3_0100	Change	Pre-immune	Anti-spyM3_0100	Change
110.31	181.91	72	97.87	250.01	152

Figure 82 provides the results of FACS analysis for surface expression of SpyM3_0102 on each of GAS serotypes M3 2721 and M3 3135. A shift in fluorescence is observed for each GAS serotype when anti-SpyM3_0102 antiserum is present, demonstrating cell surface expression. Table 27, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-SpyM3_0102 antiserum, and the difference in fluorescence value between the pre-immune and anti-SpyM3_0102 antiserum.

Table 27: Summary of FACS values for surface expression of SpyM3_0102 in M3 serotypes

2721			3135		
Pre-immune	Anti-spyM3_0102	Change	Pre-immune	Anti-spyM3_0102	Change
109.86	155.26	45	100.02	112.58	13

Figure 82 also provides the results of FACS analysis for surface expression of a pilin antigen that has homology to SpyM3_0102 identified in a different GAS serotype, M6. FACS analysis conducted with the SpyM3_0102 antisera was able to detect surface expression of the homologous SpyM3_0102 antigen on each of GAS serotypes M6 2724, M6 3650, and M6 2894. Table 28, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-SpyM3_0102 antiserum, and the difference in fluorescence value between the pre-immune and anti-SpyM3_0102 antiserum.

Table 28: Summary of FACS values for surface expression of SpyM3_0102 in M6 serotypes

2724	3650	2894

Pre-immune	Anti-spyM3_0102	Change	Pre-immune	Anti-spyM3_0102	Change	Pre-immune	Anti-spyM3_0102	Change
146.59	254.03	107	162.56	294.03	131	175.49	313.69	138

SpyM3_0102 is also homologous to pilin antigen 19224139 of GAS serotype M12. Antisera raised against SpyM3_0102 is able to detect high molecular weight structures in GAS serotype M12 strain 2728 protein fractions enriched for surface proteins, which would contain the 19224139 antigen. See Figure 109 at the lane labelled M12 2728 surf prot.

Figure 83 provides the results of FACS analysis for surface expression of SpyM3_0104 on each of GAS serotypes M3 2721 and M3 3135. A shift in fluorescence is observed for each GAS serotype when anti-SpyM3_0104 antiserum is present, demonstrating cell surface expression. Table 29, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-SpyM3_0104 antiserum, and the difference in fluorescence value between the pre-immune and anti-SpyM3_0104 antiserum.

Table 29: Summary of FACS values for surface expression of SpyM3_0104 in M3 serotypes

2721			3135		
Pre-immune	Anti-spyM3_0104	Change	Pre-immune	Anti-spyM3_0104	Change
128.45	351.65	223	105.1	339.88	235

Figure 83 also provides the results of FACS analysis for surface expression of a pilin antigen that has homology to SpyM3_0104 identified in a different GAS serotype, M12. FACS analysis conducted with the SpyM3_0104 antisera was able to detect surface expression of the homologous SpyM3_0104 antigen on GAS serotype M12 2728. Table 30, below, quantitatively summarizes the FACS fluorescence values obtained for this GAS serotype in the presence of pre-immune antiserum, anti-SpyM3_0104 antiserum, and the difference in fluorescence value between the pre-immune and anti-SpyM3_0104 antiserum.

Table 30: Summary of FACS values for surface expression of SpyM3_0104 in an M12 serotype

2728		
Pre-immune	Anti-spyM3_0104	Change
198.57	288.75	90

Figure 84 provides the results of FACS analysis for surface expression of SPs_0106 on each of GAS serotypes M3 2721 and M3 3135. A shift in fluorescence is observed for each GAS serotype when anti-SPs_0106 antiserum is present, demonstrating cell surface expression. Table 31, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-SPs_0106 antiserum, and the difference in fluorescence value between the pre-immune and anti-SPs_0106 antiserum.

Table 31: Summary of FACS values for surface expression of SPs_0106 in M3 serotypes

2721			3135		
Pre-immune	Anti-SPs_0106	Change	Pre-immune	Anti-SPs_0106	Change
116	463.28	347	103.02	494.27	391

Figure 84 also provides the results of FACS analysis for surface expression of a pilin antigen that has homology to SPs_0106 identified in a different GAS serotype, M12. FACS analysis

conducted with the SPs_0106 antisera was able to detect surface expression of the homologous SPs_0106 antigen on GAS serotype M12 2728. Table 32, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-SPs_0106 antiserum, and the difference in fluorescence value between the pre-immune and anti-SPs_0106 antiserum.

Table 32: Summary of FACS values for surface expression of SPs_0106 in an M12 serotype

2728		
Pre-immune	Anti-SPs_0106	Change
304.01	254.64	-49

(4) Adhesin Island sequence within M12: GAS Adhesin Island 4 ("GAS AI-4")

GAS Adhesin Island sequences within M12 serotype are outlined in Table 11 below. This GAS adhesin island 4 ("GAS AI-4") comprises surface proteins, a SrtC2 sortase, and a RofA regulatory protein.

GAS AI-4 surface proteins within may include a fimbrial protein, an F or F2 like fibronectin-binding protein, and a capsular polysaccharide adhesion protein (Cpa). GAS AI-4 surface proteins may also include a hypothetical surface protein in an open reading frame (orf). Preferably, each of these GAS AI-4 surface proteins include an LPXTG sortase substrate motif, such as LPXTG (SEQ ID NO: 122), VPXTG (SEQ ID NO: 137), QVXTG (SEQ ID NO: 138) or LPXAG (SEQ ID NO: 139).

GAS AI-4 includes a SrtC2 type sortase. GAS SrtC2 type sortases may preferably anchor surface proteins with a QVPTG (SEQ ID NO: 140) motif, particularly when the motif is followed by a hydrophobic region and a charged C terminus tail.

GAS AI-4 may also include a LepA putative signal peptidase I protein and a MsmRL protein.

Table 11: GAS AI-4 sequences from M12 isolate (A735)

AI-4 sequence identifier	Sortase substrate sequence or sortase type	Functional description
19224133		RofA regulatory protein
19224134	LPXTG	protein F
	SrtB	SrtB (**stop codon***)
19224135	VPXTG	Cpa
19224136		LepA
19224137	QVXTG	EftLSL.A (fimbrial)
19224138	SrtC2	EftLSL.B
19224139	LPXAG	Orf2
19224140		MsmRL
19224141	LPXTG	protein F2

A schematic of AI-4 serotype M12 is shown in Figure 51A.

One of the open reading frames encodes a SrtC2-type sortase having an amino acid sequence nearly identical to the amino acid sequence of the SrtC2-type sortase of the AI-3 serotypes described

above. See Figure 52B for an amino acid sequence alignment for each of the SrtC2 amino acid sequences.

Other proteins encoded by the open reading frames of the AI-4 serotype M12 are homologous to proteins encoded by other known adhesin islands in *S. pyogenes*, as well as the GAS AI-3 serotype M5 (Manfredo). Figure 52 is an amino acid alignment of the capsular polysaccharide adhesion protein (cpa) of AI-4 serotype M12 (19224135), GAS AI-3 serotype M5 (ORF78), *S. pyogenes* strain MGAS315 serotype M3 (21909634), *S. pyogenes* SSI-1 serotype M3 (28810257), *S. pyogenes* MGAS8232 serotype M3 (19745301), and GAS AI-2 serotype M1 (GAS15). The amino acid sequence of the AI-4 serotype M12 cpa shares a high degree of homology with other cpa proteins.

Figure 53 shows that the F-like fibronectin-binding protein encoded by the AI-4 serotype M12 open reading frame (19224134) shares homology with a F-like fibronectin-binding protein found in *S. pyogenes* strain MGAS10394 serotype M6 (50913503).

Figure 54 is an amino acid sequence alignment that illustrates that the F2-like fibronectin-binding protein of AI-4 serotype M12 (19224141) shares homology with the F2-like fibronectin-binding protein of *S. pyogenes* strain MGAS8232 serotype M3 (19745307), GAS AI-3 serotype M5 (ORF84), *S. pyogenes* strain SSI serotype M3 (28810263), and *S. pyogenes* strain MGAS315 serotype M3 (21909640).

Figure 55 is an amino acid sequence alignment that illustrates that the fimbrial protein of AI-4 serotype M12 (19224137) shares homology with the fimbrial protein of GAS AI-3 serotype M5 (ORF80), and the hypothetical protein of *S. pyogenes* strain MGAS315 serotype M3 (21909636), *S. pyogenes* strain SSI serotype M3 (28810259), *S. pyogenes* strain MGAS8732 serotype M3 (19745303), and *S. pyogenes* strain M1 GAS serotype M1 (13621428).

Figure 56 is an amino acid sequence alignment that illustrates that the hypothetical protein of GAS AI-4 serotype M12 (19224139) shares homology with the hypothetical protein of *S. pyogenes* strain MGAS315 serotype M3 (21909638), *S. pyogenes* strain SSI-1 serotype M3 (28810261), GAS AI-3 serotype M5 (ORF82), and *S. pyogenes* strain MGAS8232 serotype M3 (19745305).

The protein F2-like fibronectin-binding protein of the type 4 adhesin island also contains a highly conserved pilin motif and an E-box. Figure 60 indicates the amino acid sequence of the pilin motif and E-box in AI-4 serotype M12.

FACS analysis has confirmed that the GAS AI-4 surface proteins 19224134, 19224135, 19224137, and 19224141 are expressed on the surface of GAS. Figure 85 provides the results of FACS analysis for surface expression of 19224134 on GAS serotype M12 2728. A shift in fluorescence is observed when anti-19224134 antiserum is present, demonstrating cell surface expression. Table 33, below, quantitatively summarizes the FACS fluorescence values obtained for GAS serotype M12 2728 in the presence of pre-immune antiserum, anti-19224134 antiserum, and the difference in fluorescence value between the pre-immune and anti-19224134 antiserum.

Table 33: Summary of FACS values for surface expression of 19224134 in an M12 serotype

2728

Pre-Immune	Anti-19224134	Change
137.8	485.32	348

Figure 85 also provides the results of FACS analysis for surface expression of a pilin antigen that has homology to 19224134 identified in a different GAS serotype, M6. FACS analysis conducted with the 19224134 antisera was able to detect surface expression of the homologous 19224134 antigen on each of GAS serotypes M6 2724, M6 3650, and M6 2894. Table 34, below, quantitatively summarizes the FACS fluorescence values obtained for each GAS serotype in the presence of pre-immune antiserum, anti-19224134 antiserum, and the difference in fluorescence value between the pre-immune and anti-19224134 antiserum.

Table 34: Summary of FACS values for surface expression of 19224134 in M6 serotypes

2724			3650			2894		
Pre-immune	Anti-19224134	Change	Pre-immune	Anti-19224134	Change	Pre-immune	Anti-19224134	Change
123.58	264.59	141	140.82	262.64	122	135.4	307.25	172

Figure 86 provides the results of FACS analysis for surface expression of 19224135 on GAS serotype M12 2728. A shift in fluorescence is observed when anti-19224135 antiserum is present, demonstrating cell surface expression. Table 35, below, quantitatively summarizes the FACS fluorescence values obtained for GAS serotype M12 2728 in the presence of pre-immune antiserum, anti-19224135 antiserum, and the difference in fluorescence value between the pre-immune and anti-19224135 antiserum.

Table 35: Summary of FACS values for surface expression of 19224135 in an M12 serotype

2728		
Pre-immune	Anti-19224135	Change
151.38	471.95	321

Figure 87 provides the results of FACS analysis for surface expression of 19224137 on GAS serotype M12 2728. A shift in fluorescence is observed when anti-19224137 antiserum is present, demonstrating cell surface expression. Table 36, below, quantitatively summarizes the FACS fluorescence values obtained for GAS serotype M12 2728 in the presence of pre-immune antiserum, anti-19224137 antiserum, and the difference in fluorescence value between the pre-immune and anti-19224137 antiserum.

Table 36: Summary of FACS values for surface expression of 19224137 in an M12 serotype

2728		
Pre-immune	Anti-19224137	Change
140.44	433.25	293

Figure 88 provides the results of FACS analysis for surface expression of 19224141 on GAS serotype M12 2728. A shift in fluorescence is observed when anti-19224141 antiserum is present, demonstrating cell surface expression. Table 37, below, quantitatively summarizes the FACS fluorescence values obtained for GAS serotype M12 2728 in the presence of pre-immune antiserum, anti-19224141 antiserum, and the difference in fluorescence value between the pre-immune and anti-19224141 antiserum.

Table 37. Summary of FACS values for surface expression of 19224141 in an M12 serotype

2728		
Pre-immune	Anti-19224141	Change
147.02	498	351

19224139 (designated as orf2) may also be expressed on the surface of GAS serotype M12 bacteria. Figure 175 shows the results of FACS analysis for surface expression of 19224139 on M12 strain 2728. A slight shift in fluorescence is observed, which demonstrates that some 19224139 may be expressed on the GAS cell surface.

Surface expression of 19224135 on M12 serotype GAS has also been confirmed by Western blot analysis. Figure 99 shows that while pre-immune sera (P α-4135) does not detect GAS M12 expression of 19224135, anti-19224135 immune sera (I α-4135) is able to detect 19224135 protein in both total GAS M12 extracts (M12 tot) and GAS M12 fractions enriched for cell surface proteins (M12 surf prot). The 19224135 proteins detected in the total GAS M12 extracts or the GAS M12 extracts enriched for surface proteins are also present as high molecular weight structures, indicating that 19224135 may be in an oligomeric (pilus) form. See also Figure 108, which provides a further Western blot showing that anti-19224135 antiserum (Anti-19224135) immunoreacts with high molecular weight structures in GAS M12 strain 2728 protein extracts enriched for surface proteins.

Surface expression of 19224137 on M12 serotype GAS has also been confirmed by Western blot analysis. Figure 100 shows that while pre-immune sera (P α-4137) does not detect GAS M12 expression of 19224137, anti-19224137 immune sera (I α-4137) is able to detect 19224137 protein in both total GAS M12 extracts (M12 tot) and GAS M12 fractions enriched for cell surface proteins (M12 surf prot). The 19224137 proteins detected in the total GAS M12 extracts or the GAS M12 extracts enriched for surface proteins are also present as high molecular weight structures, indicating that 19224137 may be in an oligomeric (pilus) form. See also Figure 108, which provides a further Western blot showing that anti-19224137 antiserum (Anti-19224137) immunoreacts with high molecular weight structures in GAS M12 strain 2728 protein extracts enriched for surface proteins.

Streptococcus pneumoniae

Adhesin island sequences can be identified in *Streptococcus pneumoniae* genomes. Several of these genomes include the publicly available *Streptococcus pneumoniae* TIGR4 genome or *Streptococcus pneumoniae* strain 670 genome. Examples of these *S. pneumoniae* AI sequence are discussed below.

S. pneumoniae Adhesin Islands generally include a series of open reading frames within a *S. pneumoniae* genome that encode for a collection of surface proteins and sortases. A *S. pneumoniae* Adhesin Island may encode for amino acid sequences comprising at least one surface protein. Alternatively, an *S. pneumoniae* Adhesin Island may encode for at least two surface proteins and at least one sortase. Preferably, a *S. pneumoniae* Adhesin Island encodes for at least three surface proteins and at least two sortases. One or more of the surface proteins may include an LPXTG motif (such as LPXTG (SEQ ID NO: 122)) or other sortase substrate motif. One or more *S. pneumoniae* AI

surface proteins may participate in the formation of a pilus structure on the surface of the *S. pneumoniae* bacteria.

S. pneumoniae Adhesin Islands of the invention preferably include a divergently transcribed transcriptional regulator. The transcriptional regulator may regulate the expression of the *S. pneumoniae* AI operon.

The *S. pneumoniae* AI surface proteins may bind or otherwise adhere to fibrinogen, fibronectin, or collagen.

A schematic of the organization of a *S. pneumoniae* AI locus is provided in Figure 137. The locus comprises open reading frames encoding a transcriptional regulator (rlrA), cell wall surface proteins (rrgA, rrgB, rrgC), and sortases (srtB, srtC, srtD). Figure 137 also indicates the *S. pneumoniae* strain TIGR4 gene name corresponding to each of these open reading reading frames.

Tables 9 and 38 identify the genomic location of each of these open reading frames in *S. pneumoniae* strains TIGR4 and 670, respectively.

Table 9: S. pneumoniae AI sequences from TIGR4

Genomic Location	Strand	Length	PID	Synonym (AI Sequence Identifier)	Functional description
436302..437831	-	509	15900377	SP0461	transcriptional regulator
438326..441007	+	893	15900378	SP0462	cell wall surface anchor family protein
441231..443228	+	665	15900379	SP0463	cell wall surface anchor family protein
443275..444456	+	393	15900380	SP0464	cell wall surface anchor family protein
444675..444806	-	43	15900381	SP0465	hypothetical protein
444857..445696	+	279	15900382	SP0466	sortase
445791..446576	+	261	15900383	SP0467	sortase
446563..447414	+	283	15900384	SP0468	sortase

Table 38: S. pneumoniae strain 670 AI sequences

Genomic Location	Strand	AI Sequence Identifier\	Functional description
4383-5645	-	Orf1_670	IS1167, transposase
5910-7439	-	Orf2_670	transcriptional regulator, putative
7934-10606	+	Orf3_670	cell wall surface anchor family protein
10839-12773	+	Orf4_670	cell wall surface anchor family protein
12796-14001	+	Orf5_670	cell wall surface anchor family protein
14327-15241	+	Orf6_670	sortase, putative
15336-16121	+	Orf7_670	sortase, putative
16108-16959	+	Orf8_670	sortase, putative

The full-length nucleotide sequence of the *S. pneumoniae* strain 670 AI is also shown in Figure 101, as is its translated amino acid sequence.

At least eight other *S. pneumoniae* strains contain an adhesin island locus described by the locus depicted in Figure 137. These strains were identified by an amplification analysis. The genomes of different *S. pneumoniae* strains were amplified with eleven separate sets of primers. The sequence of each of these primers is provided below in Table 41.

Table 41: Sequences of primers used to amplify AI locus

Primer	Forward Primer Sequence	Reverse Primer Sequence

Pair		
1	ACTTTCTAATGAGTTGTTTAGGCG	AGCGACAAGCCACTGTATCATATT
2	CTGGTCGATAACTCCTTCAATCTT	GTACGACAAAAGTGTGGCTTGTT
3	GAATGCGATATTCAGGACCAACTA	ATCTCACTGAGTTAATCCGTTCAC
4	TGTATACAAGTGTGTCATTGCCAG	CATCTTCACCTGTTCTCACATTTT
5	GCGGTCTTTAGTCTTCAAAAACA	CAAGAGAAAAACACAGAGCCATAA
6	TTGCTTAAGTAAGAGAGAAAGGAGC	CAGGAGTATAGTGTCCGCTTTCTT
7	GGCAATGTTGACTTTATGAAGGTG	TATCAGCATCCCTTTATCTTCAAAC
8	TGAGATTTTCTCGTTTCTCTTAGC	AATAGACGATGGGTATTGATCATGT
9	CCGACGAACTTTGATGATTTATTG	ACCAACAGACGATGACTGTTAATC
10	AATGACTTTGAGCCTGTCTTGAT	TTCTACAATTTCCTGGCCATTATC
11	GCCATTTGGATCAGCTAAAAGTT	TTTTTCAACCCACTACAGTTGACA

These primers hybridized along the entire length of the AI locus to generate amplification products representative of sequences throughout the locus. See Figure 138, which is a schematic of the location where each of these primers hybridizes to the *S. pneumoniae* AI locus. Figure 139A provides the set of amplicons obtained from amplification of the AI locus in *S. pneumoniae* strain TIGR4. Figure 139B provides the length, in base pairs, of each amplicon in *S. pneumoniae* strain TIGR4. Amplification of the genome of *S. pneumoniae* strains 19A Hungary 6, 6B Finland 12, 6B Spain 2, 9V Spain 3, 14 CSR 10, 19F Taiwan 14, 23F Taiwan 15, and 23F Poland 16 produced a set of eleven amplicons for the eleven primer pairs, indicating that each of these strains also contained the *S. pneumoniae* AI locus.

The *S. pneumoniae* strains were also identified as containing the AI locus by comparative genome hybridization (CGH) analysis. The genomes of sixteen *S. pneumoniae* strains were interrogated for the presence of the AI locus by comparison to unique open reading frames of strain TIGR4. The AI locus was detected by this method in strains 19A Hungary 6 (19AHUN), 6B Finland 12 (6BFIN12), 6B Spain 2 (6BSP2), 14CSR10 (14 CSR10), 9V Spain 3 (9VSP3), 19F Taiwan 14 (19FTW14), 23F Taiwan 15 (19FTW15), and 23F Poland 16 (23FP16). See Figure 140.

The AI locus has been sequenced for each of these strains and the nucleotide and encoded amino acid seqeunce for each orf has been determined. An alignment of the complete nucleotide sequence of the adhesin island present in each of the ten strains is provided in Figure 196. Aligning the amino acid sequences encoded by the orfs reveals conservation of many of the AI polypeptide amino acid sequences. For example, Table 39 provides a comparison of the percent identities of the polypeptides encoded within the *S. pneumoniae* strain 670 and TIGR4 adhesin islands.

Table 39: Pecent identity comparison of S. pneumoniae strains AI sequences

S. pneumoniae strain 670 polypeptide	*S. pneumoniae* from TIGR4 polypeptide	Shared identity of polypeptides
Orf1_670	SP0460	99.3% identity in 422 aa overlap
Orf2_670	SP0461	100.0% identity in 509 aa overlap
Orf3_670	SP0462	83.2% identity in 895 aa overlap
Orf4_670	SP0463	47.9% identity in 678 aa overlap
Orf5_670	SP0464	99.7% identity in 393 aa overlap
Orf6_670	SP0466	100.0% identity in 279 aa overlap
Orf7_670	SP0467	94.2% identity in 260 aa overlap
Orf8_670	SP0468	91.5% identity in 283 aa overlap

Figures 141-147 each provide a multiple sequence alignment for the polypeptides encoded by one of the open reading frames in all ten AI-positive *S. pneumoniae* strains. In each of the sequence alignments, light shading indicates an LPXTG motif and dark shading indicates the presence of an E-box motif with the conserved glutamic acid residue of the E-box motif in bold.

The sequence alignments also revealed that the polypeptides encoded by most of the open reading frames may be divided into two groups of homology, *S. pneumoniae* AI-a and AI-b. *S. pneumoniae* strains that comprise AI-a include 14 CSR 10, 19A Hungary 6, 23F Poland 15, 670, 6B Finland 12, and 6B Spain 2. *S. pneumoniae* strains that comprise AI-b include 19F Taiwan 14, 9V Spain 3, 23F Taiwan 15, and TIGR4. An immunogenic composition of the invention may comprise one or more polypeptides from within each of *S. pneumoniae* AI-a and AI-b. For example, polypeptide RrgB, encoded by open reading frame 4, may be divided within two such groups of homology. One group contains the RrgB sequences of six *S. pneumoniae* strains and a second group contains the RrgB sequences of four *S. pneumoniae* strains. While the amino acid sequence of the strains within each individual group is 99-100 percent identical, the amino acid sequence identity of the strains in the first relative to the second group is only 48%. Table 41 provides the identity comparisons of the amino acid sequences encoded by each open reading frame for the ten *S. pneumoniae* strains.

Table 42: Conservation of amino acid sequences encoded by the *S. pneumoniae* AI locus

Putative Role of Polypeptide	Encoded by Orf	Groups of Homology	% Identity in Group	% Identity Between Groups
RlrA, transcriptional regulator	2	1 group (10 strains)	100	-
RrgA, cell wall surface protein	3	2 groups (6 + 4)	98-100	83
RrgB, cell wall surface protein	4	2 groups (6 + 4)	99-100	48
RrgC, cell wall surface protein	5	2 groups (6 + 4)	99-100	97
SrtB, putative sortase	6	2 groups (7 + 3)	99-100	97
SrtC, putative sortase	7	2 groups (6 + 4)	95-100	93
SrtD, putative sortase	8	2 groups (6 + 4)	99-100	92

The division of homology between the RrgB polypeptide in the *S. pneumoniae* strains is due a lack of amino acid sequence identity in the central amino acid residues. Amino acid residues 1-30 and 617-665 are identical for each of the ten *S. pneumoniae* strains. However, amino acid residues 31-616 share between 42 and 100 percent identity between strains. See Figure 149. The shared N- and C-terminal regions of identity in the RrgB polypeptides may be preferred portions of the RrgB polypeptide for use in an immunogenic composition. Similarly, shared regions of identity in any of the polypeptides encoded by the *S. pneumoniae* AI locus may be preferable for use in immunogenic compositions. One of skill in the art, using the amino acid alignments provided in Figures 141-147, would readily be able to determine these regions of identity.

The *S. pneumoniae* comprising these AI loci do, in fact, express high molecular weight polymers on their surface, indicating the presence of pili. See Figure 182, which shows detection of high molecular weight structures expressed by *S. pneumoniae* strains that comprise the adhesin island

locus depicted in Figure 137, these strains are indicated as rlrA+. Confirming these findings, electron microscopy and negative staining detects the presence of pili extending from the surface of *S. pneumoniae*. See Figure 185. To demonstrate that the adhesin island locus was responsible for the pili, the rrgA-srtD region of TIGR 4 were deleted. Deletion of this region of the adhesin island resulted in a loss of pili expression. See Figure 186. See also Figure 235, which provides an electron micrograph of *S. pneumoniae* lacking the rrgA-srtD region immunogold stained using anti-RrgB and anti-RrgC antibodies. No pili can be seen. Similarly to that described above, a *S. pneumoniae* bacteria that lacks a transcriptional repressor, *mgrA*, of genes in the adhesin island expresses pili. See Figure 187. However, and as expected, a *S. pneumoniae* bacteria that lacks both the mgrA and adhesin island genes in the rrgA-srtD region does not express pili. See Figure 188.

These high molecular weight pili structures appear to play a role in adherence of *S. pneumoniae* to cells. *S. pneumoniae* TIGR4 that lack the pilus operon have significantly diminished ability to adhere to A549 alveolar cells in vitro. See Figure 184.

The Sp0463 (*S. pneumoniae* TIGR4 rrgB) adhesion island polypeptide is expressed in oligomeric form. Whole cell extracts were analyzed by Western blot using a Sp0463 antiserum. The antiserum cross-hybridized with high molecular weight Sp0463 polymers. See Figure 156. The antiserum did not cross-hybridize with polypeptides from D39 or R6 strains of *S. pneumoniae*, which do not contain the AI locus depicted in Figure 137. Immunogold labelling of *S. pneumoniae* TIGR 4 using RrgB antiserum confirms the presence of RrgB in pili. Figure 189 shows double-labeling of S. pneumoniae TIGR 4 bacteria with immunolabeling for RrgB (5 nm gold particles) and RrgC (10 nm gold particles) protein. The RrgB protein is detected as present at intervals along the pilus structure. The RrgC protein is detected at the tips of the pili. See Figure 234 at arrows; Figure 234 is a close up of a pilus in Figure 189 at the location indicated by *.

The RrgA protein appears to be present in and necessary for formation of high molecular weight structures on the surface of *S. pneumoniae* TIGR4. See Figure 181 which provides the results of Western blot analysis of TIGR4 *S. pneumoniae* lacking the gene encoding RrgA. No high molecular weight structures are detected in *S. pneumoniae* that do not express RrgA using antiserum raised against RrgB. See also Figure 183.

A detailed diagram of the amino acid sequence comparions of the RrgA protein in the ten *S. pneumoniae* strains is shown in Figure 148. The diagram reveals the division of the individual *S. pneumoniae* strains into the two different homology groups.

The cell surface polypeptides encoded by the *S. pneumoniae* TIGR4 AI, Sp0462 (rrgA), Sp0463 (rrgB), and Sp0464 (rrgC), have been cloned and expressed. See examples 15-17. A polyacrylamide gel showing successful recombinant expression of RrgA is provided in Figure 190A. Detection of the RrgA protein, which is expressed in pET21b with a histidine tag, is also shown by Western blot analysis in Figure 190B, using an anti-histidine tag antibody.

Antibodies that detect RrgB and RrgC antibodies have been produced in mice. See Figures 191 and 192, which show detection of RrgB and RrgC, respectively, using the raised antibodies.

PCT/US05/27239

In addition to the identification of these *S. pneumoniae* adhesion islands, coding sequences for SrtB type sortases have been identified in several *S. pneumoniae* clinical isolates, demonstrating conservation of a SrtB type sortase across these isolates.

Recombinantly Produced AI polypeptides

It is also an aspect of the invention to alter a non-AI polypeptide to be expressed as an AI polypeptide. The non-AI polypeptide may be genetically manipulated to additionally contain AI polypeptide sequences, *e.g.*, a sortase substrate, pilin, or E-box motif, which may cause expression of the non-AI polypeptide as an AI polypeptide. Alternatively the non-AI polypeptide may be genetically manipulated to replace an amino acid sequence within the non-AI polypeptide for AI polypeptide sequences, *e.g.*, a sortase substrate, pilin, or E-box motif, which may cause expression of the non-AI polypeptide as an AI polypeptide. Any number of amino acid residues may be added to the non-AI polypeptide or may be replaced within the non-AI polypeptide to cause its expression as an AI polypeptide. At least 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 50, 75, 100, 150, 200, or 250 amino acid residues may be replaced or added to the non-AI polypeptide amino acid sequence. GBS 322 may be one such non-AI polypeptide that may be expressed as an AI polypeptide.

GBS Adhesin Island Sequences

The GBS AI polypeptides of the invention can, of course, be prepared by various means (*e.g.* recombinant expression, purification from GBS, chemical synthesis *etc.*) and in various forms (*e.g.* native, fusions, glycosylated, non-glycosylated *etc.*). They are preferably prepared in substantially pure form (*i.e.* substantially free from other streptococcal or host cell proteins) or substantially isolated form.

The GBS AI proteins of the invention may include polypeptide sequences having sequence identity to the identified GBS proteins. The degree of sequence identity may vary depending on the amino acid sequence (a) in question, but is preferably greater than 50% (*e.g.* 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.5% or more). Polypeptides having sequence identity include homologs, orthologs, allelic variants and functional mutants of the identified GBS proteins. Typically, 50% identity or more between two proteins is considered to be an indication of functional equivalence. Identity between proteins is preferably determined by the Smith-Waterman homology search algorithm as implemented in the MPSRCH program (Oxford Molecular), using an affinity gap search with parameters *gap open penalty=12* and *gap extension penalty=1*.

The GBS adhesin island polynucleotide sequences may include polynucleotide sequences having sequence identity to the identified GBS adhesin island polynucleotide sequences. The degree of sequence identity may vary depending on the polynucleotide sequence in question, but is preferably greater than 50% (*e.g.* 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.5% or more).

The GBS adhesin island polynucleotide sequences of the invention may include polynucleotide fragments of the identified adhesin island sequences. The length of the fragment may

vary depending on the polynucleotide sequence of the specific adhesin island sequence, but the fragment is preferably at least 10 consecutive polynucleotides, (*e.g.* at least 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200 or more).

The GBS adhesin island amino acid sequences of the invention may include polypeptide fragments of the identified GBS proteins. The length of the fragment may vary depending on the amino acid sequence of the specific GBS antigen, but the fragment is preferably at least 7 consecutive amino acids, (*e.g.* 8, 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200 or more). Preferably the fragment comprises one or more epitopes from the sequence. Other preferred fragments include (1) the N-terminal signal peptides of each identified GBS protein, (2) the identified GBS protein without their N-terminal signal peptides, and (3) each identified GBS protein wherein up to 10 amino acid residues (e.g. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25 or more) are deleted from the N-terminus and/or the C-terminus *e.g.* the N-terminal amino acid residue may be deleted. Other fragments omit one or more domains of the protein (*e.g.* omission of a signal peptide, of a cytoplasmic domain, of a transmembrane domain, or of an extracellular domain).

GBS 80

Examples of preferred GBS 80 fragments are discussed below. Polynucleotide and polypeptide sequences of GBS 80 from a variety of GBS serotypes and strain isolates are set forth in Figures 18 and 22. The polynucleotide and polypeptide sequences for GBS 80 from GBS serotype V, strain isolate 2603 are also included below as SEQ ID NOS 1 and 2:

SEQ ID NO. 1

```
ATGAAATTATCGAAGAAGTTATTGTTTTCGGCTGCTGTTTTAACAATGGTGGCGGGGTCAACTGTTGAACCAGTA
GCTCAGTTTGCGACTGGAATGAGTATTGTAAGAGCTGCAGAAGTGTCACAAGAACGCCCAGCGAAAACAACAGTA
AATATCTATAAATTACAAGCTGATAGTTATAAATCGGAAATTACTTCTAATGGTGGTATCGAGAATAAAGACGGC
GAAGTAATATCTAACTATGCTAAACTTGGTGACAATGTAAAAGGTTTGCAAGGTGTACAGTTTAAACGTTATAAA
GTCAAGACGGATATTTCTGTTGATGAATTGAAAAAATTGACAACAGTTGAAGCAGCAGATGCAAAAGTTGGAACG
ATTCTTGAAGAAGGTGTCAGTCTACCTCAAAAAACTAATGCTCAAGGTTTGGTCGTCGATGCTCTGGATTCAAAA
AGTAATGTGAGATACTTGTATGTAGAAGATTTAAAGAATTCACCTTCAAACATTACCAAAGCTTATGCTGTACCG
TTTGTGTTGGAATTACCAGTTGCTAACTCTACAGGTACAGGTTTCCTTTCTGAAATTAATATTTACCCTAAAAAC
GTTGTAACTGATGAACCAAAAACAGATAAAGATGTTAAAAAATTAGGTCAGGACGATGCAGGTTATACGATTGGT
GAAGAATTCAAATGGTTCTTGAAATCTACAATCCCTGCCAATTTAGGTGACTATGAAAAATTTGAAATTACTGAT
AAATTTGCAGATGGCTTGACTTATAAATCTGTTGGAAAAATCAAGATTGGTTCGAAAACACTGAATAGAGATGAG
CACTACACTATTGATGAACCAACAGTTGATAACCAAAATACATTAAAAATTACGTTTAAACCAGAGAAATTTAAA
GAAATTGCTGAGCTACTTAAAGGAATGACCCTTGTTAAAAATCAAGATGCTCTTGATAAAGCTACTGCAAATACA
GATGATGCGGCATTTTTGGAAATTCCAGTTGCATCAACTATTAATGAAAAAGCAGTTTTAGGAAAAGCAATTGAA
AATACTTTTGAACTTCAATATGACCATACTCCTGATAAAGCTGACAATCCAAAACCATCTAATCCTCCAAGAAAA
CCAGAAGTTCATACTGGTGGGAAACGATTTGTAAAGAAAGACTCAACAGAAACACAAACACTAGGTGGTGCTGAG
TTTGATTTGTTGGCTTCTGATGGGACAGCAGTAAAATGGACAGATGCTCTTATTAAAGCGAATACTAATAAAAAC
TATATTGCTGGAGAAGCTGTTACTGGGCAACCAATCAAATTGAAATCACATACAGACGGTACGTTTGAGATTAAA
GGTTTGGCTTATGCAGTTGATGCGAATGCAGAGGGTACAGCAGTAACTTACAAATTAAAAGAAACAAAAGCACCA
GAAGGTTATGTAATCCCTGATAAAGAAATCGAGTTTACAGTATCACAAACATCTTATAATACAAAACCAACTGAC
ATCACGGTTGATAGTGCTGATGCAACACCTGATACAATTAAAAACAACAAACGTCCTTCAATCCCTAATACTGGT
GGTATTGGTACGGCTATCTTTGTCGCTATCGGTGCTGCGGTGATGGCTTTTGCTGTTAAGGGGATGAAGCGTCGT
ACAAAAGATAAC
```

SEQ ID NO: 2

MKLSKKLLFSAAVLTMVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTVNIYKLQADSYKSEITSNGGIENKDG
EVISNYAKLGDNVKGLQGVQFKRYKVKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK
SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKNVVTDEPKTDKDVKKLGQDDAGYTIG
EEFKWFLKSTIPANLGDYEKFEITDKFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK

ETAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIENTFELQYDHTPDKADNPKPSNPPRK
PEVHTGGKRFVKKDSTETQTLGGAEFDLLASDGTAVKWTDALIKANTNKNYIAGEAVTGQPIKLKSHTDGTFEIK
GLAYAVDANAEGTAVTYKLKETKAPEGYVIPDKEIEFTVSQTSYNTKPTDITVDSADATPDTIKNNKRPS*IPNTG*
GIGTAIFVAIGAAVMAFAVKGMKRRTKDN

As described above, the compositions of the invention may include fragments of AI proteins. In some instances, removal of one or more domains, such as a leader or signal sequence region, a transmembrane region, a cytoplasmic region or a cell wall anchoring motif, may facilitate cloning of the gene encoding the protein and/or recombinant expression of the GBS AI protein. In addition, fragments comprising immunogenic epitopes of the cited GBS AI proteins may be used in the compositions of the invention.

For example, GBS 80 contains an N-terminal leader or signal sequence region which is indicated by the underlined sequence at the beginning of SEQ ID NO: 2 above. In one embodiment, one or more amino acids from the leader or signal sequence region of GBS 80 are removed. An example of such a GBS 80 fragment is set forth below as SEQ ID NO: 3:

SEQ ID NO: 3
AEVSQERPAKTTVNIYKLQADSYKSEITSNGGIENKDGEVISNYAKLGDNVKGLQGVQFKRYKVKTDISVDELKK
LTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSKSNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTG
TGFLSEINIYPKNVVTDEPKTDKDVKKLGQDDAGYTIGEEFKWFLKSTIPANLGDYEKFEITDKFADGLTYKSVG
KIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFKEIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVAS
TINEKAVLGKAIENTFELQYDHTPDKADNPKPSNPPRKPEVHTGGKRFVKKDSTETQTLGGAEFDLLASDGTAVK
WTDALIKANTNKNYIAGEAVTGQPIKLKSHTDGTFEIKGLAYAVDANAEGTAVTYKLKETKAPEGYVIPDKEIEF
TVSQTSYNTKPTDITVDSADATPDTIKNNKRPSIPNTGGIGTAIFVAIGAAVMAFAVKGMKRRTKDN

GBS 80 contains a C-terminal transmembrane region which is indicated by the underlined sequence near the end of SEQ ID NO: 2 above. In one embodiment, one or more amino acids from the transmembrane region and/or a cytoplasmic region are removed. An example of such a GBS 80 fragment is set forth below as SEQ ID NO: 4:

SEQ ID NO: 4
MKLSKKLLFSAAVLTMVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTVNIYKLQADSYKSEITSNGGIENKDG
EVISNYAKLGDNVKGLQGVQFKRYKVKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK
SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKNVVTDEPKTDKDVKKLGQDDAGYTIG
EEFKWFLKSTIPANLGDYEKFEITDKFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK
EIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIENTFELQYDHTPDKADNPKPSNPPRK
PEVHTGGKRFVKKDSTETQTLGGAEFDLLASDGTAVKWTDALIKANTNKNYIAGEAVTGQPIKLKSHTDGTFEIK
GLAYAVDANAEGTAVTYKLKETKAPEGYVIPDKEIEFTVSQTSYNTKPTDITVDSADATPDTIKNNKRPS*IPNTG*

GBS 80 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 5** IPNTG (shown in italics in SEQ ID NO: 2 above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant GBS 80 protein from the host cell. Accordingly, in one preferred fragment of GBS 80 for use in the invention, the transmembrane and/or cytoplasmic regions and the cell wall anchor motif are removed from GBS 80. An example of such a GBS 80 fragment is set forth below as SEQ ID NO: 6.

SEQ ID NO: 6
MKLSKKLLFSAAVLTMVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTVNIYKLQADSYKSEITSNGGIENKDG
EVISNYAKLGDNVKGLQGVQFKRYKVKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK
SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKNVVTDEPKTDKDVKKLGQDDAGYTIG
EEFKWFLKSTIPANLGDYEKFEITDKFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK

EITAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIENTFELQYDHTPDKADNPKPSNPPRK
PEVHTGGKRFVKKDSTETQTLGGAEFDLLASDGTAVKWTDALIKANTNKNYIAGEAVTGQPIKLKSHTDGTFEIK
GLAYAVDANAEGTAVTYKLKETKAPEGYVIPDKEIEFTVSQTSYNTKPTDITVDSADATPDTIKNNKRPS

Alternatively, in some recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

In one embodiment, the leader or signal sequence region, the transmembrane and cytoplasmic regions and the cell wall anchor motif are removed from the GBS 80 sequence. An example of such a GBS 80 fragment is set forth below as SEQ ID NO: 7.

SEQ ID NO: 7
AEVSQERPAKTTVNIYKLQADSYKSEITSNGGIENKDGEVISNYAKLGDNVKGLQGVQFKRYKVKTDISVDELKK
LTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSKSNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTG
TGFLSEINIYPKNVVTDEPKTDKDVKKLGQDDAGYTIGEEFKWFLKSTIPANLGDYEKFEITDKFADGLTYKSVG
KIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFKEIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVAS
TINEKAVLGKAIENTFELQYDHTPDKADNPKPSNPPRKPEVHTGGKRFVKKDSTETQTLGGAEFDLLASDGTAVK
WTDALIKANTNKNYIAGEAVTGQPIKLKSHTDGTFEIKGLAYAVDANAEGTAVTYKLKETKAPEGYVIPDKEIEF
TVSQTSYNTKPTDITVDSADATPDTIKNNKRPS

Applicants have identified a particularly immunogenic fragment of the GBS 80 protein. This immunogenic fragment is located towards the N-terminus of the protein and is underlined in the GBS 80 SEQ ID NO: 2 sequence below. The underlined fragment is set forth below as SEQ ID NO: 8.

SEQ ID NO: 2
MKLSKKLLFSAAVLTMVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTVNIYKLQADSYKSEITSNGGIENKDG
EVISNYAKLGDNVKGLQGVQFKRYKVKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK
SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKNVVTDEPKTDKDVKKLGQDDAGYTIG
EEFKWFLKSTIPANLGDYEKFEITDKFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK
EIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIENTFELQYDHTPDKADNPKPSNPPRK
PEVHTGGKRFVKKDSTETQTLGGAEFDLLASDGTAVKWTDALIKANTNKNYIAGEAVTGQPIKLKSHTDGTFEIK
GLAYAVDANAEGTAVTYKLKETKAPEGYVIPDKEIEFTVSQTSYNTKPTDITVDSADATPDTIKNNKRPS*IPNTG*
GIGTAIFVAIGAAVMAFAVKGMKRRTKDN

SEQ ID NO: 8
AEVSQERPAKTTVNIYKLQADSYKSEITSNGGIENKDGEVISNYAKLGDNVKGLQGVQFKRYKVKTDISVDELKK
LTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSKSNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTG
TGFLSEINIYPKNVVTDEPKTDKDVKKLGQDDAGYTIGEEFKWFLKSTIPANLGDYEKFEITDKFADGLTYKSVG
KIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFKEIAELLKG

The immunogenicity of the protein encoded by SEQ ID NO: 7 was compared against PBS, GBS whole cell, GBS 80 (full length) and another fragment of GBS 80, located closer to the C-terminus of the peptide (SEQ ID NO: 9, below).

SEQ ID NO: 9
MTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIENTFELQYDHTPDKADNPKPSNPPRKPEVHTGGK
RFVKKDSTETQTLGGAEFDLLASDGTAVKWTDALIKANTNKNYIAGEAVTGQPIKLKSHTDGTFEIKGLAYAVDA
NAEGTAVTYKLKETKAPEGYVIPDKEIEFTVSQTSYNTKPTDITVDSADATPDTIKNNKRPS

Both an Active Maternal Immunization Assay and a Passive Maternal Immunization Assay were conducted on this collection of proteins.

PCT/US05/27239

As used herein, an Active Maternal Immunization assay refers to an *in vivo* protection assay where female mice are immunized with the test antigen composition. The female mice are then bred and their pups are challenged with a lethal dose of GBS. Serum titers of the female mice during the immunization schedule are measured as well as the survival time of the pups after challenge.

Specifically, the Active Maternal Immunization assays referred to herein used groups of four CD-1 female mice (Charles River Laboratories, Calco Italy). These mice were immunized intraperitoneally with the selected proteins in Freund's adjuvant at days 1, 21 and 35, prior to breeding. 6-8 weeks old mice received 20 μg protein/dose when immunized with a single antigen, 30-45 μg protein/dose (15 μg each antigen) when immunized with combination of antigens. The immune response of the dams was monitored by using serum samples taken on day 0 and 49. The female mice were bred 2-7 days after the last immunization (at approximately t= 36 – 37), and typically had a gestation period of 21 days. Within 48 hours of birth, the pups were challenged via I.P. with GBS in a dose approximately equal to a amount which would be sufficient to kill 70 – 90 % of unimmunized pups (as determined by empirical data gathered from PBS control groups). The GBS challenge dose is preferably administered in 50μl of THB medium. Preferably, the pup challenge takes place at 56 to 61 days after the first immunization. The challenge inocula were prepared starting from frozen cultures diluted to the appropriate concentration with THB prior to use. Survival of pups was monitored for 5 days after challenge.

As used herein, the Passive Maternal Immunization Assay refers to an *in vivo* protection assay where pregnant mice are passively immunized by injecting rabbit immune sera (or control sera) approximately 2 days before delivery. The pups are then challenged with a lethal dose of GBS.

Specifically, the Passive Maternal Immunization Assay referred to herein used groups of pregnant CD1 mice which were passively immunized by injecting 1 ml of rabbit immune sera or control sera via I.P., 2 days before delivery. Newborn mice (24-48 hrs after birth) are challenged via I.P. with a 70 - 90% lethal dose of GBS serotype III COH1. The challenge dose, obtained by diluting a frozen mid log phase culture, was administered in 50μl of THB medium.

For both assays, the number of pups surviving GBS infection was assessed every 12 hrs for 4 days. Statistical significance was estimated by Fisher's exact test.

The results of each assay for immunization with SEQ ID NO: 7, SEQ ID NO: 8, PBS and GBS whole cell are set forth in Tables 1 and 2 below.

TABLE 1: Immunization

Antigen	Alive/total	%Survival	Fisher's exact test
PBS (neg control)	13/80	16%	
GBS (whole cell)	54/65	83%	P<0.00000001
GBS80 (intact)	62/70	88%	P<0.00000001
GBS80 (fragment) SEQ ID 7	35/64	55%	P=0.0000013
GBS80 (fragment) SEQ ID 8	13/67	19%	P=0.66

Table 2: Passive Maternal Immunization

Antigen	Alive/total	%Survival	Fisher's exact test
PBS (neg control)	12/42	28%	
GBS (whole cell)	48/52	92%	P<0.00000001
GBS80 (intact)	48/55	87%	P<0.00000001
GBS80 (fragment) SEQ ID 7	45/57	79%	P=0.0000006
GBS80 (fragment) SEQ ID 8	13/54	24%	P=1

As shown in Tables 1 and 2, immunization with the SEQ ID NO: 7 GBS 80 fragment provided a substantially improved survival rate for the challenged pups than the comparison SEQ ID NO: 8 GBS 80 fragment. These results indicate that the SEQ ID NO: 7 GBS 80 fragment may comprise an important immunogenic epitope of GBS 80.

As discussed above, pilin motifs, containing conserved lysine (K) residues have been identified in GBS 80. The pilin motif sequences are underlined in SEQ ID NO: 2, below. Conserved lysine (K) residues are marked in bold, at amino acid residues 199 and 207 and at amino acid residues 368 and 375. The pilin sequences, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures of GBS 80. Preferred fragments of GBS 80 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 2

MKLSKKLLFSAAVLTMVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTVNIYKLQADSYKSEITSNGGIENKDG
EVISNYAKLGDNVKGLQGVQFKRYKVKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK
SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKNVVTDEPKTDKDVKKLGQDDAGYTIG
EEFKWFLKSTIPANLGDYEKFEITDKFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK
EIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIENTFELQYDHTPDKADNPKPSNPPRK
PEVHTGGKRFVKKDSTETQTLGGAEFDLLASDGTAVKWTDALIKANTNKNYIAGEAVTGQPIKLKSHTDGTFEIK
GLAYAVDANAEGTAVTYKLKETKAPEGYVIPDKEIEFTVSQTSYNTKPTDITVDSADATPDTIKNNKRPSIPNTG
GIGTAIFVAIGAAVMAFAVKGMKRRTKDN

E boxes containing conserved glutamic residues have also been identified in GBS 80. The E box motifs are underlined in SEQ ID NO: 2 below. The conserved glutamic acid (E) residues, at amino acid residues 392 and 471, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of GBS 80. Preferred fragments of GBS 80 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 2

MKLSKKLLFSAAVLTMVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTVNIYKLQADSYKSEITSNGGIENKDG
EVISNYAKLGDNVKGLQGVQFKRYKVKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK
SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKNVVTDEPKTDKDVKKLGQDDAGYTIG
EEFKWFLKSTIPANLGDYEKFEITDKFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK
EIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIENTFELQYDHTPDKADNPKPSNPPRK
PEVHTGGKRFVKKDSTETQTLGGAEFDLLASDGTAVKWTDALIKANTNKNYIAGEAVTGQPIKLKSHTDGTFEIK
GLAYAVDANAEGTAVTYKLKETKAPEGYVIPDKEIEFTVSQTSYNTKPTDITVDSADATPDTIKNNKRPSIPNTG
GIGTAIFVAIGAAVMAFAVKGMKRRTKDN

GBS 104

PCT/US05/27239

Similarly, the following offers examples of preferred GBS 104 fragments. Nucleotide and amino acid sequences of GBS 104 sequenced from serotype V isolated strain 2603 are set forth below as SEQ ID NOS 10 and 11:

SEQ ID NO. 10

ATGAAAAAGAGACAAAAAATATGGAGAGGGTTATCAGTTACTTTACTAATCCTGTCCCAAATTCCATTTGGTATA
TTGGTACAAGGTGAAACCCAAGATACCAATCAAGCACTTGGAAAAGTAATTGTTAAAAAAACGGGAGACAATGCT
ACACCATTAGGCAAAGCGACTTTTGTGTTAAAAAATGACAATGATAAGTCAGAAACAAGTCACGAAACGGTAGAG
GGTTCTGGAGAAGCAACCTTTGAAAACATAAAACCTGGAGACTACACATTAAGAGAAGAAACAGCACCAATTGGT
TATAAAAAAACTGATAAAACCTGGAAAGTTAAAGTTGCAGATAACGGAGCAACAATAATCGAGGGTATGGATGCA
GATAAAGCAGAGAAACGAAAAGAAGTTTTGAATGCCCAATATCCAAAATCAGCTATTTATGAGGATACAAAAGAA
AATTACCCATTAGTTAATGTAGAGGGTTCCAAAGTTGGTGAACAATACAAAGCATTGAATCCAATAAATGGAAAA
GATGGTCGAAGAGAGATTGCTGAAGGTTGGTTATCAAAAAAAATTACAGGGGTCAATGATCTCGATAAGAATAAA
TATAAAATTGAATTAACTGTTGAGGGTAAAACCACTGTTGAAACGAAAGAACTTAATCAACCACTAGATGTCGTT
GTGCTATTAGATAATTCAAATAGTATGAATAATGAAAGAGCCAATAATTCTCAAAGAGCATTAAAAGCTGGGGAA
GCAGTTGAAAAGCTGATTGATAAAATTACATCAAATAAAGACAATAGAGTAGCTCTTGTGACATATGCCTCAACC
ATTTTTGATGGTACTGAAGCGACCGTATCAAAGGGAGTTGCCGATCAAAATGGTAAAGCGCTGAATGATAGTGTA
TCATGGGATTATCATAAAACTACTTTTACAGCAACTACACATAATTACAGTTATTTAAATTTAACAAATGATGCT
AACGAAGTTAATATTCTAAAGTCAAGAATTCCAAAGGAAGCGGAGCATATAAATGGGGATCGCACGCTCTATCAA
TTTGGTGCGACATTTACTCAAAAAGCTCTAATGAAAGCAAATGAAATTTTAGAGACACAAAGTTCTAATGCTAGA
AAAAAACTTATTTTTCACGTAACTGATGGTGTCCCTACGATGTCTTATGCCATAAATTTTAATCCTTATATATCA
ACATCTTACCAAAACCAGTTTAATTCTTTTTTAAATAAAATACCAGATAGAAGTGGTATTCTCCAAGAGGATTTT
ATAATCAATGGTGATGATTATCAAATAGTAAAAGGAGATGGAGAGAGTTTTAAACTGTTTTCGGATAGAAAAGTT
CCTGTTACTGGAGGAACGACACAAGCAGCTTATCGAGTACCGCAAAATCAACTCTCTGTAATGAGTAATGAGGGA
TATGCAATTAATAGTGGATATATTTATCTCTATTGGAGAGATTACAACTGGGTCTATCCATTTGATCCTAAGACA
AAGAAAGTTTCTGCAACGAAACAAATCAAAACTCATGGTGAGCCAACAACATTATACTTTAATGGAAATATAAGA
CCTAAAGGTTATGACATTTTTACTGTTGGGATTGGTGTAAACGGAGATCCTGGTGCAACTCCTCTTGAAGCTGAG
AAATTTATGCAATCAATATCAAGTAAAACAGAAAATTATACTAATGTTGATGATACAAATAAAATTTATGATGAG
CTAAATAAATACTTTAAAACAATTGTTGAGGAAAAACATTCTATTGTTGATGGAAATGTGACTGATCCTATGGGA
GAGATGATTGAATTCCAATTAAAAAATGGTCAAAGTTTTACACATGATGATTACGTTTTGGTTGGAAATGATGGC
AGTCAATTAAAAAATGGTGTGGCTCTTGGTGGACCAAACAGTGATGGGGGAATTTTAAAAGATGTTACAGTGACT
TATGATAAGACATCTCAAACCATCAAAATCAATCATTTGAACTTAGGAAGTGGACAAAAAGTAGTTCTTACCTAT
GATGTACGTTTAAAAGATAACTATATAAGTAACAAATTTTACAATACAAATAATCGTACAACGCTAAGTCCGAAG
AGTGAAAAAGAACCAAATACTATTCGTGATTTCCCAATTCCCAAAATTCGTGATGTTCGTGAGTTTCCGGTACTA
ACCATCAGTAATCAGAAGAAAATGGGTGAGGTTGAATTTATTAAAGTTAATAAAGACAAACATTCAGAATCGCTT
TTGGGAGCTAAGTTTCAACTTCAGATAGAAAAAGATTTTTCTGGGTATAAGCAATTTGTTCCAGAGGGAAGTGAT
GTTACAACAAAGAATGATGGTAAAATTTATTTTAAAGCACTTCAAGATGGTAACTATAAATTATATGAAATTTCA
AGTCCAGATGGCTATATAGAGGTTAAAACGAAACCTGTTGTGACATTTACAATTCAAAATGGAGAAGTTACGAAC
CTGAAAGCAGATCCAAATGCTAATAAAAATCAAATCGGGTATCTTGAAGGAAATGGTAAACATCTTATTACCAAC
ACTCCCAAACGCCCACCAGGTGTTTTTCCTAAAACAGGGGGAATTGGTACAATTGTCTATATATTAGTTGGTTCT
ACTTTTATGATACTTACCATTTGTTCTTTCCGTCGTAAACAATTG

SEQ ID NO. 11

MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNATPLGKATFVLKNDNDKSETSHETVE
GSGEATFENIKPGDYTLREETAPIGYKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKE
NYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKKITGVNDLDKNKYKIELTVEGKTTVETKELNQPLDVV
VLLDNSNSMNNERANNSQRALKAGEAVEKLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSV
SWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQFGATFTQKALMKANEILETQSSNAR
KKLIFHVTDGVPTMSYAINFNPYISTSYQNQFNSFLNKIPDRSGILQEDFIINGDDYQIVKGDGESFKLFSDRKV
PVTGGTTQAAYRVPQNQLSVMSNEGYAINSGYIYLYWRDYNWVYPFDPKTKKVSATKQIKTHGEPTTLYFNGNIR
PKGYDIFTVGIGVNGDPGATPLEAEKFMQSISSKTENYTNVDDTNKIYDELNKYFKTIVEEKHSIVDGNVTDPMG
EMIEFQLKNGQSFTHDDYVLVGNDGSQLKNGVALGGPNSDGGILKDVTVTYDKTSQTIKINHLNLGSGQKVVLTY
DVRLKDNYISNKFYNTNNRTTLSPKSEKEPNTIRDFPIPKIRDVREFPVLTISNQKKMGEVEFIKVNKDKHSESL
LGAKFQLQIEKDFSGYKQFVPEGSDVTTKNDGKIYFKALQDGNYKLYEISSPDGYIEVKTKPVVTFTIQNGEVTN
LKADPNANKNQIGYLEGNGKHLITNTPKRPPGV*FPKTG*GIGTIVYILVGSTFMILTICSFRRKQL

GBS 104 contains an N-terminal leader or signal sequence region which is indicated by the underlined sequence at the beginning of SEQ ID NO 11 above. In one embodiment, one or more

amino acid sequences from the leader or signal sequence region of GBS 104 are removed. An example of such a GBS 104 fragment is set forth below as SEQ ID NO 12.

SEQ ID NO 12

GETQDTNQALGKVIVKKTGDNATPLGKATFVLKNDNDKSETSHETVEGSGEATFENIKPGDYTLREETAPIGYKK
TDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKENYPLVNVEGSKVGEQYKALNPINGKDGR
REIAEGWLSKKITGVNDLDKNKYKIELTVEGKTTVETKELNQPLDVVVLLDNSNSMNNERANNSQRALKAGEAVE
KLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSVSWDYHKTTFTATTHNYSYLNLTNDANEV
NILKSRIPKEAEHINGDRTLYQFGATFTQKALMKANEILETQSSNARKKLIFHVTDGVPTMSYAINFNPYISTSY
QNQFNSFLNKIPDRSGILQEDFIINGDDYQIVKGDGESFKLFSDRKVPVTGGTTQAAYRVPQNQLSVMSNEGYAI
NSGYIYLYWRDYNWVYPFDPKTKKVSATKQIKTHGEPTTLYFNGNIRPKGYDIFTVGIGVNGDPGATPLEAEKFM
QSISSKTENYTNVDDTNKIYDELNKYFKTIVEEKHSIVDGNVTDPMGEMIEFQLKNGQSFTHDDYVLVGNDGSQL
KNGVALGGPNSDGGILKDVTVTYDKTSQTIKINHLNLGSGQKVVLTYDVRLKDNYISNKFYNTNNRTTLSPKSEK
EPNTIRDFPIPKIRDVREFPVLTISNQKKMGEVEFIKVNKDKHSESLLGAKFQLQIEKDFSGYKQFVPEGSDVTT
KNDGKIYFKALQDGNYKLYEISSPDGYIEVKTKPVVTFTIQNGEVTNLKADPNANKNQIGYLEGNGKHLITNTPK
RPPGVFPKTGGIGTIVYILVGSTFMILTICSFRRKQL

GBS 104 contains a C-terminal transmembrane and/or cytoplasmic region which is indicated by the underlined region near the end of SEQ ID NO 11 above. In one embodiment, one or more amino acids from the transmembrane or cytomplasmic regions are removed. An example of such a GBS 104 fragment is set forth below as SEQ ID NO 13.

SEQ ID NO: 13

MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNATPLGKATFVLKNDNDKSETSHETVE
GSGEATFENIKPGDYTLREETAPIGYKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKE
NYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKKITGVNDLDKNKYKIELTVEGKTTVETKELNQPLDVV
VLLDNSNSMNNERANNSQRALKAGEAVEKLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSV
SWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQFGATFTQKALMKANEILETQSSNAR
KKLIFHVTDGVPTMSYAINFNPYISTSYQNQFNSFLNKIPDRSGILQEDFIINGDDYQIVKGDGESFKLFSDRKV
PVTGGTTQAAYRVPQNQLSVMSNEGYAINSGYIYLYWRDYNWVYPFDPKTKKVSATKQIKTHGEPTTLYFNGNIR
PKGYDIFTVGIGVNGDPGATPLEAEKFMQSISSKTENYTNVDDTNKIYDELNKYFKTIVEEKHSIVDGNVTDPMG
EMIEFQLKNGQSFTHDDYVLVGNDGSQLKNGVALGGPNSDGGILKDVTVTYDKTSQTIKINHLNLGSGQKVVLTY
DVRLKDNYISNKFYNTNNRTTLSPKSEKEPNTIRDFPIPKIRDVREFPVLTISNQKKMGEVEFIKVNKDKHSESL
LGAKFQLQIEKDFSGYKQFVPEGSDVTTKNDGKIYFKALQDGNYKLYEISSPDGYIEVKTKPVVTFTIQNGEVTN
LKADPNANKNQIGYLEGNGKHLITNT

In one embodiment, one or more amino acids from the leader or signal sequence region and one or more amino acids from the transmembrane or cytoplasmic regions are removed. An example of such a GBS 104 fragment is set forth below as SEQ ID NO 14.

SEQ ID NO: 14

GETQDTNQALGKVIVKKTGDNATPLGKATFVLKNDNDKSETSHETVEGSGEATFENIKPGDYTLREETAPIGYKK
TDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKENYPLVNVEGSKVGEQYKALNPINGKDGR
REIAEGWLSKKITGVNDLDKNKYKIELTVEGKTTVETKELNQPLDVVVLLDNSNSMNNERANNSQRALKAGEAVE
KLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSVSWDYHKTTFTATTHNYSYLNLTNDANEV
NILKSRIPKEAEHINGDRTLYQFGATFTQKALMKANEILETQSSNARKKLIFHVTDGVPTMSYAINFNPYISTSY
QNQFNSFLNKIPDRSGILQEDFIINGDDYQIVKGDGESFKLFSDRKVPVTGGTTQAAYRVPQNQLSVMSNEGYAI
NSGYIYLYWRDYNWVYPFDPKTKKVSATKQIKTHGEPTTLYFNGNIRPKGYDIFTVGIGVNGDPGATPLEAEKFM
QSISSKTENYTNVDDTNKIYDELNKYFKTIVEEKHSIVDGNVTDPMGEMIEFQLKNGQSFTHDDYVLVGNDGSQL
KNGVALGGPNSDGGILKDVTVTYDKTSQTIKINHLNLGSGQKVVLTYDVRLKDNYISNKFYNTNNRTTLSPKSEK
EPNTIRDFPIPKIRDVREFPVLTISNQKKMGEVEFIKVNKDKHSESLLGAKFQLQIEKDFSGYKQFVPEGSDVTT
KNDGKIYFKALQDGNYKLYEISSPDGYIEVKTKPVVTFTIQNGEVTNLKADPNANKNQIGYLEGNGKHLITNT

GBS 104, like GBS 80, contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 123** FPKTG (shown in italics in SEQ ID NO: 11 above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant GBS 104 protein from the host cell. Accordingly, in one preferred fragment of GBS 104 for use in the

PCT/US2005/027239

invention, only the transmembrane and/or cytoplasmic regions and the cell wall anchor motif are removed from GBS 104. Alternatively, in some recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Two pilin motifs, containing conserved lysine (K) residues, have been identified in GBS 104. The pilin motif sequences are underlined in SEQ ID NO: 11, below. Conserved lysine (K) residues are marked in bold, at amino acid residues 141 and 149 and at amino acid residues 499 and 507. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures of GBS 104. Preferred fragments of GBS 104 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO. 11

MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNATPLGKATFVLKNDNDKSETSHETVE
GSGEATFENIKPGDYTLREETAPIGYKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKE
NYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKKITGVNDLDKNKYKIELTVEGKTTVETKELNQPLDVV
VLLDNSNSMNNERANNSQRALKAGEAVEKLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSV
SWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQFGATFTQKALMKANEILETQSSNAR
KKLIFHVTDGVPTMSYAINFNPYISTSYQNQFNSFLNKIPDRSGILQEDFIINGDDYQIVKGDGESFKLFSDRKV
PVTGGTTQAAYRVPQNQLSVMSNEGYAINSGYIYLYWRDYNWVYPFDPKTKKVSATKQIKTHGEPTTLYFNGNIR
PKGYDIFTVGIGVNGDPGATPLEAEKFMQSISSKTENYTNVDDTNKIYDELNKYFKTIVEEKHSIVDGNVTDPMG
EMIEFQLKNGQSFTHDDYVLVGNDGSQLKNGVALGGPNSDGGILKDVTVTYDKTSQTIKINHLNLGSGQKVVLTY
DVRLKDNYISNKFYNTNNRTTLSPKSEKEPNTIRDFPIPKIRDVREFPVLTISNQKKMGEVEFIKVNKDKHSESL
LGAKFQLQIEKDFSGYKQFVPEGSDVTTKNDGKIYFKALQDGNYKLYEISSPDGYIEVKTKPVVTFTIQNGEVTN
LKADPNANKNQIGYLEGNGKHLITNTPKRPPGVFPKTGGIGTIVYILVGSTFMILTICSFRRKQL

Two E boxes containing a conserved glutamic residues have also been identified in GBS 104. The E box motifs are underlined in SEQ ID NO: 11 below. The conserved glutamic acid (E) residues, at amino acid residues 94 and 798, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of GBS 104. Preferred fragments of GBS 104 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO. 11

MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNATPLGKATFVLKNDNDKSETSHETVE
GSGEATFENIKPGDYTLREETAPIGYKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKE
NYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKKITGVNDLDKNKYKIELTVEGKTTVETKELNQPLDVV
VLLDNSNSMNNERANNSQRALKAGEAVEKLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSV
SWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQFGATFTQKALMKANEILETQSSNAR
KKLIFHVTDGVPTMSYAINFNPYISTSYQNQFNSFLNKIPDRSGILQEDFIINGDDYQIVKGDGESFKLFSDRKV
PVTGGTTQAAYRVPQNQLSVMSNEGYAINSGYIYLYWRDYNWVYPFDPKTKKVSATKQIKTHGEPTTLYFNGNIR
PKGYDIFTVGIGVNGDPGATPLEAEKFMQSISSKTENYTNVDDTNKIYDELNKYFKTIVEEKHSIVDGNVTDPMG
EMIEFQLKNGQSFTHDDYVLVGNDGSQLKNGVALGGPNSDGGILKDVTVTYDKTSQTIKINHLNLGSGQKVVLTY
DVRLKDNYISNKFYNTNNRTTLSPKSEKEPNTIRDFPIPKIRDVREFPVLTISNQKKMGEVEFIKVNKDKHSESL
LGAKFQLQIEKDFSGYKQFVPEGSDVTTKNDGKIYFKALQDGNYKLYEISSPDGYIEVKTKPVVTFTIQNGEVTN
LKADPNANKNQIGYLEGNGKHLITNTPKRPPGVFPKTGGIGTIVYILVGSTFMILTICSFRRKQL

GBS 067

The following offers examples of preferred GBS 067 fragments. Nucleotide and amino acid sequence of GBS 067 sequences from serotype V isolated strain 2603 are set forth below as SEQ ID NOS: 15 and 16.

SEQ ID NO: 15

ATGAGAAAATACCAAAAATTTTCTAAAATATTGACGTTAAGTCTTTTTTGTTTGTCGCAAATACCGCTTAATACC
AATGTTTTAGGGGAAAGTACCGTACCGGAAAATGGTGCTAAAGGAAAGTTAGTTGTTAAAAAGACAGATGACCAG
AACAAACCACTTTCAAAAGCTACCTTTGTTTTAAAAACTACTGCTCATCCAGAAAGTAAAATAGAAAAAGTAACT
GCTGAGCTAACAGGTGAAGCTACTTTTGATAATCTCATACCTGGAGATTATACTTTATCAGAAGAAACAGCGCCC
GAAGGTTATAAAAAGACTAACCAGACTTGGCAAGTTAAGGTTGAGAGTAATGGAAAAACTACGATACAAAATAGT
GGTGATAAAAATTCCACAATTGGACAAAATCAGGAAGAACTAGATAAGCAGTATCCCCCCACAGGAATTTATGAA
GATACAAAGGAATCTTATAAACTTGAGCATGTTAAAGGTTCAGTTCCAAATGGAAAGTCAGAGGCAAAAGCAGTT
AACCCATATTCAAGTGAAGGTGAGCATATAAGAGAAATTCCAGAGGGAACATTATCTAAACGTATTTCAGAAGTA
GGTGATTTAGCTCATAATAAATATAAAATTGAGTTAACTGTCAGTGGAAAAACCATAGTAAAACCAGTGGACAAA
CAAAAGCCGTTAGATGTTGTCTTCGTACTCGATAATTCTAACTCAATGAATAACGATGGCCCAAATTTTCAAAGG
CATAATAAAGCCAAGAAAGCTGCCGAAGCTCTTGGGACCGCAGTAAAAGATATTTTAGGAGCAAACAGTGATAAT
AGGGTTGCATTAGTTACCTATGGTTCAGATATTTTTGATGGTAGGAGTGTAGATGTCGTAAAAGGATTTAAAGAA
GATGATAAATATTATGGCCTTCAAACTAAGTTCACAATTCAGACAGAGAATTATAGTCATAAACAATTAACAAAT
AATGCTGAAGAGATTATAAAAAGGATTCCGACAGAAGCTCCTAAAGCTAAGTGGGGATCTACTACCAATGGATTA
ACTCCAGAGCAACAAAAGGAGTACTATCTTAGTAAAGTAGGAGAAACATTTACTATGAAAGCCTTCATGGAGGCA
GATGATATTTTGAGTCAAGTAAATCGAAATAGTCAAAAAATTATTGTTCATGTAACTGATGGTGTTCCTACGAGA
TCATATGCTATTAATAATTTTAAACTGGGTGCATCATATGAAAGCCAATTTGAACAAATGAAAAAAAATGGATAT
CTAAATAAAAGTAATTTTCTACTTACTGATAAGCCCGAGGATATAAAAGGAAATGGGGAGAGTTACTTTTTGTTT
CCCTTAGATAGTTATCAAACACAGATAATCTCTGGAAACTTACAAAAACTTCATTATTTAGATTTAAATCTTAAT
TACCCTAAAGGTACAATTTATCGAAATGGACCAGTGAAAGAACATGGAACACCAACCAAACTTTATATAAATAGT
TTAAAACAGAAAAATTATGACATTTTTAATTTTGGTATCGATATATCTGGTTTTAGACAAGTTTATAATGAGGAG
TATAAGAAAAATCAAGATGGTACTTTTCAAAAATTGAAAGAGGAAGCTTTTAAACTTTCAGATGGAGAAATCACA
GAACTAATGAGGTCGTTCTCTTCCAAACCTGAGTACTACACCCCTATCGTAACTTCAGCCGATACATCTAACAAT
GAAATTTTATCTAAAATTCAGCAACAATTTGAAACGATTTTAACAAAAGAAAACTCAATTGTTAATGGAACTATC
GAAGATCCTATGGGTGATAAAATCAATTTACAGCTTGGTAATGGACAAACATTACAGCCAAGTGATTATACTTTA
CAGGGAAATGATGGAAGTGTAATGAAGGATGGTATTGCAACTGGTGGGCCTAATAATGATGGTGGAATACTTAAG
GGGGTTAAATTAGAATACATCGGAAATAAACTCTATGTTAGAGGTTTGAATTTAGGAGAAGGTCAAAAAGTAACA
CTCACATATGATGTGAAACTAGATGACAGTTTTATAAGTAACAAATTCTATGACACTAATGGTAGAACAACATTG
AATCCTAAGTCAGAGGATCCTAATACACTTAGAGATTTTCCAATCCCTAAAATTCGTGATGTGAGAGAATATCCT
ACAATAACGATTAAAAACGAGAAGAAGTTAGGTGAAATTGAATTTATAAAAGTTGATAAAGATAATAATAAGTTG
CTTCTCAAAGGAGCTACGTTTGAACTTCAAGAATTTAATGAAGATTATAAACTTTATTTACCAATAAAAAATAAT
AATTCAAAAGTAGTGACGGGAGAAAACGGCAAAATTTCTTACAAAGATTTGAAAGATGGCAAATATCAGTTAATA
GAAGCAGTTTCGCCGGAGGATTATCAAAAAATTACTAATAAACCAATTTTAACTTTTGAAGTGGTTAAAGGATCG
ATAAAAAATATAATAGCTGTTAATAAACAGATTTCTGAATATCATGAGGAAGGTGACAAGCATTTAATTACCAAC
ACGCATATTCCACCAAAAGGAATTATTCCTATGACAGGTGGGAAAGGAATTCTATCTTTCATTTTAATAGGTGGA
GCTATGATGTCTATTGCAGGTGGAATTTATATTTGGAAAAGGTATAAGAAATCTAGTGATATGTCCATCAAAAAA
GAT

SEQ ID NO: 16

MRKYQKFSKILTLSLFCLSQIPLNTNVLGESTVPENGAKGKLVVKKTDDQNKPLSKATFVLKTTAHPESKIEKVT
AELTGEATFDNLIPGDYTLSEETAPEGYKKTNQTWQVKVESNGKTTIQNSGDKNSTIGQNQEELDKQYPPTGIYE
DTKESYKLEHVKGSVPNGKSEAKAVNPYSSEGEHIREIPEGTLSKRISEVGDLAHNKYKIELTVSGKTIVKPVDK
QKPLDVVFVLDNSNSMNNDGPNFQRHNKAKKAAEALGTAVKDILGANSDNRVALVTYGSDIFDGRSVDVVKGFKE
DDKYYGLQTKFTIQTENYSHKQLTNNAEEIIKRIPTEAPKAKWGSTTNGLTPEQQKEYYLSKVGETFTMKAFMEA
DDILSQVNRNSQKIIVHVTDGVPTRSYAINNFKLGASYESQFEQMKKNGYLNKSNFLLTDKPEDIKGNGESYFLF
PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVKEHGTPTKLYINSLKQKNYDIFNFGIDISGFRQVYNEE
YKKNQDGTFQKLKEEAFKLSDGEITELMRSFSSKPEYYTPIVTSADTSNNEILSKIQQQFETILTKENSIVNGTI
EDPMGDKINLQLGNGQTLQPSDYTLQGNDGSVMKDGIATGGPNNDGGILKGVKLEYIGNKLYVRGLNLGEGQKVT
LTYDVKLDDSFISNKFYDTNGRTTLNPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL
LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLIEAVSPEDYQKITNKPILTFEVVKGS
IKNIIAVNKQISEYHEEGDKHLITNTHIPPKGI*IPMTG*GKGILS<u>FILIGGAMMSIAGGIYI</u>WKRYKKSSDMSIKK
D

GBS 067 contains a C-terminus transmembrane region which is indicated by the underlined region closest to the C-terminus of SEQ ID NO: 16 above. In one embodiment, one or more amino acids from the transmembrane region is removed and or the amino acid is truncated before the

transmembrane region. An example of such a GBS 067 fragment is set forth below as SEQ ID NO: 17.

SEQ ID NO: 17

MRKYQKFSKILTLSLFCLSQIPLNTNVLGESTVPENGAKGKLVVKKTDDQNKPLSKATFVLKTTAHPESKIEKVT
AELTGEATFDNLIPGDYTLSEETAPEGYKKTNQTWQVKVESNGKTTIQNSGDKNSTIGQNQEELDKQYPPTGIYE
DTKESYKLEHVKGSVPNGKSEAKAVNPYSSEGEHIREIPEGTLSKRISEVGDLAHNKYKIELTVSGKTIVKPVDK
QKPLDVVFVLDNSNSMNNDGPNFQRHNKAKKAAEALGTAVKDILGANSDNRVALVTYGSDIFDGRSVDVVKGFKE
DDKYYGLQTKFTIQTENYSHKQLTNNAEEIIKRIPTEAPKAKWGSTTNGLTPEQQKEYYLSKVGETFTMKAFMEA
DDILSQVNRNSQKIIVHVTDGVPTRSYAINNFKLGASYESQFEQMKKNGYLNKSNFLLTDKPEDIKGNGESYFLF
PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVKEHGTPTKLYINSLKQKNYDIFNFGIDISGFRQVYNEE
YKKNQDGTFQKLKEEAFKLSDGEITELMRSFSSKPEYYTPIVTSADTSNNEILSKIQQQFETILTKENSIVNGTI
EDPMGDKINLQLGNGQTLQPSDYTLQGNDGSVMKDGIATGGPNNDGGILKGVKLEYIGNKLYVRGLNLGEGQKVT
LTYDVKLDDSFISNKFYDTNGRTTLNPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL
LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLIEAVSPEDYQKITNKPILTFEVVKGS
IKNIIAVNKQISEYHEEGDKHLITNTHIPPKGIIPMTGGKGILS

GBS 067 contains an amino acid motif indicative of a cell wall anchor (an LPXTG (SEQ ID NO: 122) motif): **SEQ ID NO: 18** IPMTG. (shown in italics in SEQ ID NO: 16 above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant GBS 067 protein from the host cell. Accordingly, in one preferred fragment of GBS 067 for use in the invention, the transmembrane and the cell wall anchor motif are removed from GBS 67. An example of such a GBS 067 fragment is set forth below as SEQ ID NO: 19.

SEQ ID NO: 19

MRKYQKFSKILTLSLFCLSQIPLNTNVLGESTVPENGAKGKLVVKKTDDQNKPLSKATFVLKTTAHPESKIEKVT
AELTGEATFDNLIPGDYTLSEETAPEGYKKTNQTWQVKVESNGKTTIQNSGDKNSTIGQNQEELDKQYPPTGIYE
DTKESYKLEHVKGSVPNGKSEAKAVNPYSSEGEHIREIPEGTLSKRISEVGDLAHNKYKIELTVSGKTIVKPVDK
QKPLDVVFVLDNSNSMNNDGPNFQRHNKAKKAAEALGTAVKDILGANSDNRVALVTYGSDIFDGRSVDVVKGFKE
DDKYYGLQTKFTIQTENYSHKQLTNNAEEIIKRIPTEAPKAKWGSTTNGLTPEQQKEYYLSKVGETFTMKAFMEA
DDILSQVNRNSQKIIVHVTDGVPTRSYAINNFKLGASYESQFEQMKKNGYLNKSNFLLTDKPEDIKGNGESYFLF
PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVKEHGTPTKLYINSLKQKNYDIFNFGIDISGFRQVYNEE
YKKNQDGTFQKLKEEAFKLSDGEITELMRSFSSKPEYYTPIVTSADTSNNEILSKIQQQFETILTKENSIVNGTI
EDPMGDKINLQLGNGQTLQPSDYTLQGNDGSVMKDGIATGGPNNDGGILKGVKLEYIGNKLYVRGLNLGEGQKVT
LTYDVKLDDSFISNKFYDTNGRTTLNPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL
LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLIEAVSPEDYQKITNKPILTFEVVKGS
IKNIIAVNKQISEYHEEGDKHLITNTHIPPKGI

Alternatively, in some recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Three pilin motifs, containing conserved lysine (K) residues have been identified in GBS 67. The pilin motif sequences are underlined in SEQ ID NO: 16, below. Conserved lysine (K) residues are marked in bold, at amino acid residues 478 and 488, at amino acid residues 340 and 342, and at amino acid residues 703 and 717. The pilin sequences, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures of GBS 67. Preferred fragments of GBS 67 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 16

MRKYQKFSKILTLSLFCLSQIPLNTNVLGESTVPENGAKGKLVVKKTDDQNKPLSKATFVLKTTAHPESKIEKVT
AELTGEATFDNLIPGDYTLSEETAPEGYKKTNQTWQVKVESNGKTTIQNSGDKNSTIGQNQEELDKQYPPTGIYE
DTKESYKLEHVKGSVPNGKSEAKAVNPYSSEGEHIREIPEGTLSKRISEVGDLAHNKYKIELTVSGKTIVKPVDK
QKPLDVVFVLDNSNSMNNDGPNFQRHNKAKKAAEALGTAVKDILGANSDNRVALVTYGSDIFDGRSVDVVKGFKE
DDKYYGLQTKFTIQTENYSHKQLTNNAEEIIKRIPTEAPKAKWGSTTNGLTPEQQKEYYLSKVGETFTMKAFMEA
DDILSQVNRNSQKIIVHVTDGVPTRSYAINNFKLGASYESQFEQMKKNGYLNKSNFLLTDKPEDIKGNGESYFLF
PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVKEHGTPTKLYINSLKQKNYDIFNFGIDISGFRQVYNEE
YKKNQDGTFQKLKEEAFKLSDGEITELMRSFSSKPEYYTPIVTSADTSNNEILSKIQQQFETILTKENSIVNGTI
EDPMGDKINLQLGNGQTLQPSDYTLQGNDGSVMKDGIATGGPNNDGGILKGVKLEYIGNKLYVRGLNLGEGQKVT
LTYDVKLDDSFISNKFYDTNGRTTLNPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL
LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLIEAVSPEDYQKITNKPILTFEVVKGS
IKNIIAVNKQISEYHEEGDKHLITNTHIPPKGIIPMTGGKGILSFILIGGAMMSIAGGIYIWKRYKKSSDMSIKK
D

Two E boxes containing conserved glutamic residues have also been identified in GBS 67. The E box motifs are underlined in SEQ ID NO: 16 below. The conserved glutamic acid (E) residues, at amino acid residues 96 and 801, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of GBS 67. Preferred fragments of GBS 67 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 16

MRKYQKFSKILTLSLFCLSQIPLNTNVLGESTVPENGAKGKLVVKKTDDQNKPLSKATFVLKTTAHPESKIEKVT
AELTGEATFDNLIPGDYTLS**E**ETAPEGYKKTNQTWQVKVESNGKTTIQNSGDKNSTIGQNQEELDKQYPPTGIYE
DTKESYKLEHVKGSVPNGKSEAKAVNPYSSEGEHIREIPEGTLSKRISEVGDLAHNKYKIELTVSGKTIVKPVDK
QKPLDVVFVLDNSNSMNNDGPNFQRHNKAKKAAEALGTAVKDILGANSDNRVALVTYGSDIFDGRSVDVVKGFKE
DDKYYGLQTKFTIQTENYSHKQLTNNAEEIIKRIPTEAPKAKWGSTTNGLTPEQQKEYYLSKVGETFTMKAFMEA
DDILSQVNRNSQKIIVHVTDGVPTRSYAINNFKLGASYESQFEQMKKNGYLNKSNFLLTDKPEDIKGNGESYFLF
PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVKEHGTPTKLYINSLKQKNYDIFNFGIDISGFRQVYNEE
YKKNQDGTFQKLKEEAFKLSDGEITELMRSFSSKPEYYTPIVTSADTSNNEILSKIQQQFETILTKENSIVNGTI
EDPMGDKINLQLGNGQTLQPSDYTLQGNDGSVMKDGIATGGPNNDGGILKGVKLEYIGNKLYVRGLNLGEGQKVT
LTYDVKLDDSFISNKFYDTNGRTTLNPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL
LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLI**E**AVSPEDYQKITNKPILTFEVVKGS
IKNIIAVNKQISEYHEEGDKHLITNTHIPPKGIIPMTGGKGILSFILIGGAMMSIAGGIYIWKRYKKSSDMSIKK
D

Predicted secondary structure for the GBS 067 amino acid sequence is set forth in FIGURE 33. As shown in this figure, GBS 067 contains several regions predicted to form alpha helical structures. Such alpha helical regions are likely to form coiled-coil structures and may be involved in oligomerization of GBS 067.

The amino acid sequence for GBS 067 also contains a region which is homologous to the Cna_B domain of the Staphylococcus aureus collagen-binding surface protein (pfam05738). Although the Cna_B region is not thought to mediate collagen binding, it is predicted to form a beta sandwich structure. In the Staph aureus protein, this beta sandwich structure is through to form a stalk that presents the ligand binding domain away from the bacterial cell surface. This same amino acid sequence region is also predicted to be an outer membrane protein involved in cell envelope biogenesis.

The amino acid sequence for GBS 067 contains a region which is homologous to a von Willebrand factor (vWF) type A domain. The vWF type A domain is present at amino acid residues 229-402 of GBS 067 as shown in SEQ ID NO: 16. This type of sequence is typically found in

extracellular proteins such as integrins and it thought to mediate adhesion, including adhesion to collagen, fibronectin, and fibrinogen, discussed above.

Because applicants have identified GBS 67 as a surface exposed protein on GBS and because GBS 67 may be involved in GBS adhesion, the immunogenicity of the GBS 67 protein was examined in mice. The results of an immunization assay with GBS 67 are set forth in Table 48, below.

Table 48: GBS 67 Protects Mice in an Immunization Assay

Challenge GBS strain (serotype)	GBS 67 immungen		PBS immunogen		FACS
	dead/treated	% survival	dead/treated	% survival	Δmean
3050 (II)	0/30	100	29/49	41	460
CJB111 (V)	76/185	59	143/189	24	481
7357 b (Ib)	34/56	39	65/74	12	316

As shown in Table 48, immunization with GBS 67 provides a substantially improved survival rate for challenged mice relative to negative control, PBS, immunized mice. These results indicate that GBS 67 may comprise an immunogenic composition of the invention.

GBS 59

The following offers examples of GBS 59 fragments. Nucleotide and amino acid sequences of GBS 59 sequenced from serotype V isolated strain 2603 are set forth below as SEQ ID NOS: 125 and 126. The GBS 59 polypeptide of SEQ ID NO: 126 is referred to as SAG1407.

SEQ ID NO: 125

```
ttaagcttcctttgattggcgtcttttcatgataactactgctccaagcataatgcttaaaccaataattgtgaa
aagaattgtaccaataccacctgtttgtgggattgttacctttttatttctacacgtgtcgcatcttttggtt
gctgttagcaacgtagtcaatgttaccacctgttatgtatgacccttgattaactacaaacttaatattacctgc
caacttagcaaatcctgctggagcaagtgtttcttcaaggttgtaagtaccgtctgcaagacctgtaacttcaaa
ttgaccttgatcgtttgaagtgtaggtaatggctctagccttatctgttatccactcataagctgtacgagcctc
aatgaaggctgcatcgtaatctgcttgtttagttttgataagttcttttgcagtaattccttttttcacctttttg
gtctgttgcagacaacttgttataagcagcgatagcttcatctaaagctattttcttagcagctaaagttttttg
accttctgattgatctgctttaagagcaaggtatttacctgctgagtttttcacaacgaattgtgcaccagccaa
acggtcaccttgttcattagttttgacaaatttcttaccatgagtttcaacttttggttcagttgggttcaatgg
tgttgggttatcagaatctttggtattggtaatggttactttaccattttctagatttattgcacttccgtaacc
agaaacacgttctgagatcatgtatgatttgttttctagaccagtgaatttacccgagaagttaccagatacttc
aaatttgataccatttccaaggtcgattgtacctttagatgttttgtcaatgatactgaagcaacagttttatc
tttatctttcaatgtgtaaacaacgtttacaccatcaggtgcaattccgtcagaccaagttttagcaactgttac
ttcacccttgaaggtgtaacaggaagttcagtcaagtctttacctggtttgttaccatacgacaatttgatatc
attggattctggattatcaataattgcttgaccattaacagtagcactataagtcaatgtaaattcaatatcagc
tgttttagctgctttttccaatttgcccaatccatcagctgtgaattttaatgtgaaaccacgggcatcaatgct
aagttcatagtctgtatccttagcaaaagtttctgtagttcctgaagctttaaggctaacagttgaacccattgt
caaaccatttgacattatatctgtccaaaccaagttttcgtatttagaacctttgtgaatttttgttttaacttc
ataaggaacaactttaccgatttcagcagtagcagttgctttgtcacgtgcataattaccataatttgcgccagc
tgtcaaaagtctattaacatctgtcaatgctgtcaaatcgtttgttttagcaaagtttttatcaatttctggttt
ttcttcagtgttctttggataaacatgggcatcagcaacaacaccatcttcatttaccaatggaagagtgatgtt
aactggaaccgcttttgaagcagccaggagggaaccattattgttgtaagtagattttgatttaacttcaacaat
tttaaactcgcctttcaatcctttggtgttgaaaacaagtccagtatctccctctggtgtcaatccagacacggc
ctcatcaatatttactgttatttcaggagtaccatctttattaattaaggctggtgttaatttgttaccttctttt
tgccttaacatattgcactttaccacttttatcttctttcaaagctaaagcaaagaacgcaccttcgatttcttt
agatccctcgccaaagtaaccagcaaggtcagaaatagctccacctttgtagtcttttccgttaagacctgtagt
tcctgggaagttacttttgttaagatttgattcggtttgcaaaatcttgtgcaaagtcactgtattagttgttgc
```

ttdatdcgcaaagdctggtgcgactgaghggtaatgacgttaaagtcagtaacaatgccgagaacattgcaaaata
tttgttgattcttttcat

SEQ ID NO: 126

MKRINKYFAMFSALLLTLTSLLSVAPAFADEATTNTVTLHKILQTESNLNKSNFPGTTGLNGKDYKGGAISDLAG
YFGEGSKEIEGAFFALALKEDKSGKVQYVKAKEGNKLTPALINKDGTPEITVNIDEAVSGLTPEGDTGLVFNTKG
LKGEFKIVEVKSKSTYNNNGSLLAASKAVPVNITLPLVNEDGVVADAHVYPKNTEEKPEIDKNFAKTNDLTALTD
VNRLLTAGANYGNYARDKATATAEIGKVVPYEVKTKIHKGSKYENLVWTDIMSNGLTMGSTVSLKASGTTETFAK
DTDYELSIDARGFTLKFTADGLGKLEKAAKTADIEFTLTYSATVNGQAIIDNPESNDIKLSYGNKPGKDLTELPV
TPSKGEVTVAKTWSDGIAPDGVNVVYTLKDKDKTVASVSLTKTSKGTIDLGNGIKFEVSGNFSGKFTGLENKSYM
ISERVSGYGSAINLENGKVTITNTKDSDNPTPLNPTEPKVETHGKKFVKTNEQGDRLAGAQFVVKNSAGKYLALK
ADQSEGQKTLAAKKIALDEAIAAYNKLSATDQKGEKGITAKELIKTKQADYDAAFIEARTAYEWITDKARAITYT
SNDQGQFEVTGLADGTYNLEETLAPAGFAKLAGNIKFVVNQGSYITGGNIDYVANSNQKDATRVENKKVT*IPQTG*
GIGTILFTIIGLSIMLGAVVIMKRRQSKEA

Nucleotide and amino acid sequences of GBS 59 sequenced from serotype V isolated strain CJB111 are set forth below as SEQ ID NOS: 127 and 128. The GBS 59 polypeptide of SEQ ID NO: 128 is referred to as BO1575.

SEQ ID NO: 127

ATGAAAAAAATCAACAAATGTCTTACAATGTTCTCGACACTGCTATTGATCTTAACGTCACTATTCTCAGTTGCA
CCAGCGTTTGCGGACGACGCAACAACTGATACTGTGACCTTGCACAAGATTGTCATGCCACAAGCTGCATTTGAT
AACTTTACTGAAGGTACAAAAGGTAAGAATGATAGCGATTATGTTGGTAAACAAATTAATGACCTTAAATCTTAT
TTTGGCTCAACCGATGCTAAAGAAATCAAGGGTGCTTTCTTTGTTTTCAAAAATGAAACTGGTACAAAATTCATT
ACTGAAAATGGTAAGGAAGTCGATACTTTGGAAGCTAAAGATGCTGAAGGTGGTGCTGTTCTTTCAGGGTTAACA
AAAGACAATGGTTTTGTTTTTAACACTGCTAAGTTAAAAGGAATTTACCAAATCGTTGAATTGAAAGAAAAATCA
AACTACGATAACAACGGTTCTATCTTGGCTGATTCAAAAGCAGTTCCAGTTAAAATCACTCTGCCATTGGTAAAC
AACCAAGGTGTTGTTAAAGATGCTCACATTTATCCAAAGAATACTGAAACAAAACCACAAGTAGATAAGAACTTT
GCAGATAAAGATCTTGATTATACTGACAACCGAAAAGACAAAGGTGTTGTCTCAGCGACAGTTGGTGACAAAAAA
GAATACATAGTTGGAACAAAAATTCTTAAAGGCTCAGACTATAAGAAACTGGTTTGGACTGATAGCATGACTAAA
GGTTTGACGTTCAACAACAACGTTAAAGTAACATTGGATGGTGAAGATTTTCCTGTTTTAAACTACAAACTCGTA
ACAGATGACCAAGGTTTCCGTCTTGCCTTGAATGCAACAGGTCTTGCAGCAGTAGCAGCAGCTGCAAAAGACAAA
GATGTTGAAATCAAGATCACTTACTCAGCTACGGTGAACGGCTCCACTACTGTTGAAATTCCAGAAACCAATGAT
GTTAAATTGGACTATGGTAATAACCCAACGGAAGAAAGTGAACCACAAGAAGGTACTCCAGCTAACCAAGAAATT
AAAGTCATTAAAGACTGGGCAGTAGATGGTACAATTACTGATGCTAATGTTGCAGTTAAAGCTATCTTTACCTTG
CAAGAAAAACAAACGGATGGTACATGGGTGAACGTTGCTTCACACGAAGCAACAAAACCATCACGCTTTGAACAT
ACTTTCACAGGTTTGGATAATGCTAAAACTTACCGCGTTGTCGAACGTGTTAGCGGCTACACTCCAGAATACGTA
TCATTTAAAAATGGTGTTGTGACTATCAAGAACAACAAAAACTCAAATGATCCAACTCCAATCAACCCATCAGAA
CCAAAAGTGGTGACTTATGGACGTAAATTTGTGAAAACAAATCAAGCTAACACTGAACGCTTGGCAGGAGCTACC
TTCCTCGTTAAGAAAGAAGGCAAATACTTGGCACGTAAAGCAGGTGCAGCAACTGCTGAAGCAAAGGCAGCTGTA
AAAACTGCTAAACTAGCATTGGATGAAGCTGTTAAAGCTTATAACGACTTGACTAAAGAAAAACAAGAAGGCCAA
GAAGGTAAAACAGCATTGGCTACTGTTGATCAAAAACAAAAAGCTTACAATGACGCTTTTGTTAAAGCTAACTAC
TCATATGAATGGGTTGCAGATAAAAAGGCTGATAATGTTGTTAAATTGATCTCTAACGCCGGTGGTCAATTTGAA
ATTACTGGTTTGGATAAAGGCACTTATGGCTTGGAAGAAACTCAAGCACCAGCAGGTTATGCGACATTGTCAGGT
GATGTAAACTTTGAAGTAACTGCCACATCATATAGCAAAGGGGCTACAACTGACATCGCATATGATAAAGGCTCT
GTAAAAAAAGATGCCCAACAAGTTCAAAACAAAAAAGTAACCATCCCACAAACAGGTGGTATTGGTACAATTCTT
TTCACAATTATTGGTTTAAGCATTATGCTTGGAGCAGTAGTTATCATGAAAAAACGTCAATCAGAGGAAGCTTAA

SEQ ID NO: 128

MKKINKCLTMFSTLLLILTSLFSVAPAFADDATTDTVTLHKIVMPQAAFDNFTEGTKGKNDSDYVGKQINDLKSY
FGSTDAKEIKGAFFVFKNETGTKFITENGKEVDTLEAKDAEGGAVLSGLTKDNGFVFNTAKLKGIYQIVELKEKS
NYDNNGSILADSKAVPVKITLPLVNNQGVVKDAHIYPKNTETKPQVDKNFADKDLDYTDNRKDKGVVSATVGDKK
EYIVGTKILKGSDYKKLVWTDSMTKGLTFNNNVKVTLDGEDFPVLNYKLVTDDQGFRLALNATGLAAVAAAAKDK
DVEIKITYSATVNGSTTVEIPETNDVKLDYGNNPTEESEPQEGTPANQEIKVIKDWAVDGTITDANVAVKAIFTL
QEKQTDGTWVNVASHEATKPSRFEHTFTGLDNAKTYRVVERVSGYTPEYVSFKNGVVTIKNNKNSNDPTPINPSE
PKVVTYGRKFVKTNQANTERLAGATFLVKKEGKYLARKAGAATAEAKAAVKTAKLALDEAVKAYNDLTKEKQEGQ
EGKTALATVDQKQKAYNDAFVKANYSYEWVADKKADNVVKLISNAGGQFEITGLDKGTYGLEETQAPAGYATLSG
DVNFEVTATSYSKGATTDIAYDKGSVKKDAQQVQNKKVT*IPQTG*GIGTILFTIIGLSIMLGAVVIMKKRQSEEA

The GBS 59 polypeptides contain an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 129** IPQTG (shown in italics in SEQ ID NOs: 126 and 128 above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant GBS 59 protein from the host cell. Alternatively, in some recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Pilin motifs, containing conserved lysine (K) residues have been identified in the GBS 59 polypeptides. The pilin motif sequences are underlined in each of SEQ ID NOs: 126 and 128, below. Conserved lysine (K) residues are marked in bold. The conserved lysine (K) residues are located at amino acid residues 202 and 212 and amino acid residues 489 and 495 of SEQ ID NO: 126 and at amino acid residues 188 and 198 of SEQ ID NO: 128. The pilin sequences, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures of GBS 59. Preferred fragments of GBS 59 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 126

MKRINKYFAMFSALLLTLTSLLSVAPAFADEATTNTVTLHKILQTESNLNKSNFPGTTGLNGKDYKGGAISDLAG
YFGEGSKEIEGAFFALALKEDKSGKVQYVKAKEGNKLTPALINKDGTPEITVNIDEAVSGLTPEGDTGLVFNTKG
LKGEFKIVEVKSKSTYNNNGSLLAASKAVPVNITLPLVNEDGVVADAHVYPKNTEEKPEIDKNFAKTNDLTALTD
VNRLLTAGANYGNYARDKATATAEIGKVVPYEVKTKIHKGSKYENLVWTDIMSNGLTMGSTVSLKASGTTETFAK
DTDYELSIDARGFTLKFTADGLGKLEKAAKTADIEFTLTYSATVNGQAIIDNPESNDIKLSYGNKPGKDLTELPV
TPSKGEVTVAKTWSDGIAPDGVNVVYTLKDKDKTVASVSLTKTSKGTIDLGNGIKFEVSGNFSGKFTGLENKSYM
ISERVSGYGSAINLENGKVTITNTKDSDNPTPLNPTEPKVETHGKKFVKTNEQGDRLAGAQFVVKNSAGKYLALK
ADQSEGQKTLAAKKIALDEAIAAYNKLSATDQKGEKGITAKELIKTKQADYDAAFIEARTAYEWITDKARAITYT
SNDQGQFEVTGLADGTYNLEETLAPAGFAKLAGNIKFVVNQGSYITGGNIDYVANSNQKDATRVENKKVTIPQTG
GIGTILFTIIGLSIMLGAVVIMKRRQSKEA

SEQ ID NO: 128

MKKINKCLTMFSTLLLILTSLFSVAPAFADDATTDTVTLHKIVMPQAAFDNFTEGTKGKNDSDYVGKQINDLKSY
FGSTDAKEIKGAFFVFKNETGTKFITENGKEVDTLEAKDAEGGAVLSGLTKDNGFVFNTAKLKGIYQIVELKEKS
NYDNNGSILADSKAVPVKITLPLVNNQGVVKDAHIYPKNTETKPQVDKNFADKDLDYTDNRKDKGVVSATVGDKK
EYIVGTKILKGSDYKKLVWTDSMTKGLTFNNNVKVTLDGEDFPVLNYKLVTDDQGFRLALNATGLAAVAAAAKDK
DVEIKITYSATVNGSTTVEIPETNDVKLDYGNNPTEESEPQEGTPANQEIKVIKDWAVDGTITDANVAVKAIFTL
QEKQTDGTWVNVASHEATKPSRFEHTFTGLDNAKTYRVVERVSGYTPEYVSFKNGVVTIKNNKNSNDPTPINPSE
PKVVTYGRKFVKTNQANTERLAGATFLVKKEGKYLARKAGAATAEAKAAVKTAKLALDEAVKAYNDLTKEKQEGQ
EGKTALATVDQKQKAYNDAFVKANYSYEWVADKKADNVVKLISNAGGQFEITGLDKGTYGLEETQAPAGYATLSG
DVNFEVTATSYSKGATTDIAYDKGSVKKDAQQVQNKKVTIPQTGGIGTILFTIIGLSIMLGAVVIMKKRQSEEA

An E box containing a conserved glutamic residue has also been identified in each of the GBS 59 polypeptides. The E box motif is underlined in each of SEQ ID NOs: 126 and 128 below. The conserved glutamic acid (E) is marked in bold at amino acid residue 621 in SEQ ID NO: 126 and at amino acid residue 588 in SEQ ID NO: 128. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of GBS 59. Preferred fragments of GBS 59 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 126

MKRINKYFAMFSALLLTLTSLLSVAPAFADEATTNTVTLHKILQTESNLNKSNFPGTTGLNGKDYKGGAISDLAG
YFGEGSKEIEGAFFALALKEDKSGKVQYVKAKEGNKLTPALINKDGTPEITVNIDEAVSGLTPEGDTGLVFNTKG
LKGEFKIVEVKSKSTYNNNGSLLAASKAVPVNITLPLVNEDGVVADAHVYPKNTEEKPEIDKNFAKTNDLTALTD
VNRLLTAGANYGNYARDKATATAEIGKVVPYEVKTKIHKGSKYENLVWTDIMSNGLTMGSTVSLKASGTTETFAK
DTDYELSIDARGFTLKFTADGLGKLEKAAKTADIEFTLTYSATVNGQAIIDNPESNDIKLSYGNKPGKDLTELPV
TPSKGEVTVAKTWSDGIAPDGVNVVYTLKDKDKTVASVSLTKTSKGTIDLGNGIKFEVSGNFSGKFTGLENKSYM
ISERVSGYGSAINLENGKVTITNTKDSDNPTPLNPTEPKVETHGKKFVKTNEQGDRLAGAQFVVKNSAGKYLALK
ADQSEGQKTLAAKKIALDEAIAAYNKLSATDQKGEKGITAKELIKTKQADYDAAFIEARTAYEWITDKARAITYT
SNDQGQFEVTGLADGTYNLEETLAPAGFAKLAGNIKFVVNQGSYITGGNIDYVANSNQKDATRVENKKVTIPQTG
GIGTILFTIIGLSIMLGAVVIMKRRQSKEA

SEQ ID NO: 128

MKKINKCLTMFSTLLLLILTSLFSVAPAFADDATTDTVTLHKIVMPQAAFDNFTEGTKGKNDSDYVGKQINDLKSY
FGSTDAKEIKGAFFVFKNETGTKFITENGKEVDTLEAKDAEGGAVLSGLTKDNGFVFNTAKLKGIYQIVELKEKS
NYDNNGSILADSKAVPVKITLPLVNNQGVVKDAHIYPKNTETKPQVDKNFADKDLDYTDNRKDKGVVSATVGDKK
EYIVGTKILKGSDYKKLVWTDSMTKGLTFNNNVKVTLDGEDFPVLNYKLVTDDQGFRLALNATGLAAVAAAAKDK
DVEIKITYSATVNGSTTVEIPETNDVKLDYGNNPTEESEPQEGTPANQEIKVIKDWAVDGTITDANVAVKAIFTL
QEKQTDGTWVNVASHEATKPSRFEHTFTGLDNAKTYRVVERVSGYTPEYVSFKNGVVTIKNNKNSNDPTPINPSE
PKVVTYGRKFVKTNQANTERLAGATFLVKKEGKYLARKAGAATAEAKAAVKTAKLALDEAVKAYNDLTKEKQEGQ
EGKTALATVDQKQKAYNDAFVKANYSYEWVADKKADNVVKLISNAGGQFEITGLDKGTYGLEETQAPAGYATLSG
DVNFEVTATSYSKGATTDIAYDKGSVKKDAQQVQNKKVTIPQTGGIGTILFTIIGLSIMLGAVVIMKKRQSEEA

Female mice were immunized with either SAG1407 (SEQ ID NO: 126) or BO1575 (SEQ ID NO: 128) in an active maternal immunization assay. Pups bred from the immunized female mice survived GBS challenge better than control (PBS) treated mice. Results of the active maternal immunization assay using the GBS 59 immunogenic compositions are shown in Table 17, below.

TABLE 17: Active maternal immunization assay for GBS 59

Challenge GBS strain (serotype)	GBS 59		PBS		
	Dead/treated	Survival (%)	Dead/treated	Survival (%)	FACS
CJB111 (V)*	7/20	65	41/49	16	493
18RS21 (II)**	18/30	40	39/40	2.5	380

* immunized with BO1575
**immunized with SAG1407

Opsonophagocytosis assays also demonstrated that antibodies against BO1575 are opsonic for GBS serotype V, strain CJB111. See Figure 67.

GBS 52

Examples of polynucleotide and amino acid sequences for GBS 52 are set forth below. SEQ ID NO: 20 and 21 represent GBS 52 sequences from GBS serotype V, strain isolate 2603.

SEQ ID NO: 20

ATGAAACAAACATTAAAACTTATGTTTTCTTTTCTGTTGATGTTAGGGACTATGTTTGGAATTAGCCAAACTGTT
TTAGCGCAAGAAACTCATCAGTTGACGATTGTTCATCTTGAAGCAAGGGATATTGATCGTCCAAATCCACAGTTG
GAGATTGCCCCTAAAGAAGGGACTCCAATTGAAGGAGTACTCTATCAGTTGTACCAATTAAAATCAACTGAAGAT
GGCGATTTGTTGGCACATTGGAATTCCCTAACTATCACAGAATTGAAAAAACAGGCGCAGCAGGTTTTTGAAGCC
ACTACTAATCAACAAGGAAAGGCTACATTTAACCAACTACCAGATGGAATTTATTATGGTCTGGCGGTTAAAGCC
GGTGAAAAAAATCGTAATGTCTCAGCTTTCTTGGTTGACTTGTCTGAGGATAAAGTGATTTATCCTAAAATCATC
TGGTCCACAGGTGAGTTGGACTTGCTTAAAGTTGGTGTGGATGGTGATACCAAAAAACCACTAGCAGGCGTTGTC
TTTGAACTTTATGAAAAGAATGGTAGGACTCCTATTCGTGTGAAAAATGGGGTGCATTCTCAAGATATTGACGCT
GCAAAACATTTAGAAACAGATTCATCAGGGCATATCAGAATTTCCGGGCTCATCCATGGGGACTATGTCTTAAAA
GAAATCGAGACACAGTCAGGATATCAGATCGGACAGGCAGAGACTGCTGTGACTATTGAAAAATCAAAAACAGTA

ACAGTAACGATTGAAATAAAAAAGTTCCGACACCTAAAGTGCCATCTCGAGGAGGTCTTATTCCCAAAACAGGT
GAGCAACAGGCAATGGCACTTGTAATTATTGGTGGTATTTTAATTGCTTTAGCCTTACGATTACTATCAAAACAT
CGGAAACATCAAAATAAGGAT

SEQ ID NO: 21

MKQTLKLMFSFLLMLGTMFGISQTVLAQETHQLTIVHLEARDIDRPNPQLEIAPKEGTPIEGVLYQLYQLKSTED
GDLLAHWNSLTITELKKQAQQVFEATTNQQGKATFNQLPDGIYYGLAVKAGEKNRNVSAFLVDLSEDKVIYPKII
WSTGELDLLKVGVDGDTKKPLAGVVFELYEKNGRTPIRVKNGVHSQDIDAAKHLETDSSGHIRISGLIHGDYVLK
EIETQSGYQIGQAETAVTIEKSKTVTVTIENKKVPTPKVPSRGGL*IPKTG*EQQAMALVIIGGILIALALRLLSKH
RKHQNKD

GBS 52 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 124** IPKTG (shown in italics in SEQ ID NO: 21, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant GBS 52 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in GBS 52. The pilin motif sequence is underlined in SEQ ID NO: 21, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 148 and 160. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of GBS 52 include at least one conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 21

MKQTLKLMFSFLLMLGTMFGISQTVLAQETHQLTIVHLEARDIDRPNPQLEIAPKEGTPIEGVLYQLYQLKSTED
GDLLAHWNSLTITELKKQAQQVFEATTNQQGKATFNQLPDGIYYGLAVKAGEKNRNVSAFLVDLSEDKVIYPKII
WSTGELDLLKVGVDGDTKKPLAGVVFELYEKNGRTPIRVKNGVHSQDIDAAKHLETDSSGHIRISGLIHGDYVLK
EIETQSGYQIGQAETAVTIEKSKTVTVTIENKKVPTPKVPSRGGLIPKTGEQQAMALVIIGGILIALALRLLSKH
RKHQNKD

An E box containing a conserved glutamic residue has been identified in GBS 52. The E-box motif is underlined in SEQ ID NO: 21, below. The conserved glutamic acid (E), at amino acid residue 226, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of GBS 52. Preferred fragments of GBS 52 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 21

MKQTLKLMFSFLLMLGTMFGISQTVLAQETHQLTIVHLEARDIDRPNPQLEIAPKEGTPIEGVLYQLYQLKSTED
GDLLAHWNSLTITELKKQAQQVFEATTNQQGKATFNQLPDGIYYGLAVKAGEKNRNVSAFLVDLSEDKVIYPKII
WSTGELDLLKVGVDGDTKKPLAGVVFELYEKNGRTPIRVKNGVHSQDIDAAKHLETDSSGHIRISGLIHGDYVLK
EIETQSGYQIGQAETAVTIEKSKTVTVTIENKKVPTPKVPSRGGLIPKTGEQQAMALVIIGGILIALALRLLSKH
RKHQNKD

SAG0647

Examples of polynucleotide and amino acid sequences for SAG0647 are set forth below.

SEQ ID NO: 22 and 23 represent SAG0647 sequences from GBS serotype V, strain isolate 2603.

SEQ ID NO: 22

ATGGGACAAAAATCAAAAATATCTCTAGCTACGAATATTCGTATATGGATTTTTCGTTTAATTTTCTTAGCGGGT
TTCCTTGTTTTGGCATTTCCCATCGTTAGTCAGGTCATGTACTTTCAAGCCTCTCACGCCAATATTAATGCTTTT
AAAGAAGCTGTTACCAAGATTGACCGGGTGGAGATTAATCGGCGTTTAGAACTTGCTTATGCTTATAACGCCAGT
ATAGCAGGTGCCAAAACTAATGGCGAATATCCAGCGCTTAAAGACCCCTACTCTGCTGAACAAAAGCAGGCAGGG
GTCGTTGAGTACGCCCGCATGCTTGAAGTCAAAGAACAAATAGGTCATGTGATTATTCCAAGAATTAATCAGGAT
ATCCCTATTTACGCTGGCTCTGCTGAAGAAAATCTTCAGAGGGGCGTTGGACATTTAGAGGGGACCAGTCTTCCA
GTCGGTGGTGAGTCAACTCATGCCGTTCTAACTGCCCATCGAGGGCTACCAACGGCCAAGCTATTTACCAATTTA
GACAAGGTAACAGTAGGTGACCGTTTTTACATTGAACACATCGGCGGAAAGATTGCTTATCAGGTAGACCAAATC
AAAGTTATCGCCCCTGATCAGTTAGAGGATTTGTACGTGATTCAAGGAGAAGATCACGTCACCCTATTAACTTGC
ACACCTTATATGATAAATAGTCATCGCCTCCTCGTTCGAGGCAAGCGAATTCCTTATGTGGAAAAAACAGTGCAG
AAAGATTCAAAGACCTTCAGGCAACAACAATACCTAACCTATGCTATGTGGGTAGTCGTTGGACTTATCTTGCTG
TCGCTTCTCATTTGGTTTAAAAAGACGAAACAGAAAAAGCGGAGAAAGAATGAAAAAGCGGCTAGTCAAAATAGT
CACAATAATTCGAAATAA

SEQ ID NO: 23

MGQKSKISLATNIRIWIFRLIFLAGFLVLAFPIVSQVMYFQASHANINAFKEAVTKIDRVEINRRLELAYAYNAS
IAGAKTNGEYPALKDPYSAEQKQAGVVEYARMLEVKEQIGHVIIPRINQDIPIYAGSAEENLQRGVGHLEGTSLP
VGGESTHAVLTAHRGLPTAKLFTNLDKVTVGDRFYIEHIGGKIAYQVDQIKVIAPDQLEDLYVIQGEDHVTLLTC
TPYMINSHRLLVRGKRIPYVEKTVQKDSKTFRQQQYLTYAMWVVVGLILLSLLIWFKKTKQKKRRKNEKAASQNS
HNNSK

SAG0648

Examples of polynucleotide and amino acid sequences for SAG0648 are set forth below.

SEQ ID NO: 24 and 25 represent SAG0648 sequences from GBS serotype V, strain isolate 2603.

SEQ ID NO: 24

ATGGGAAGTCTGATTCTCTTATTTCCGATTGTGAGCCAGGTAAGTTACTACCTTGCTTCGCATCAAAATATTAAT
CAATTTAAGCGGGAAGTCGCTAAGATTGATACTAATACGGTTGAACGACGCATCGCTTTAGCTAATGCTTACAAT
GAGACGTTATCAAGGAATCCCTTGCTTATAGACCCTTTTACCAGTAAGCAAAAAGAAGGTTTGAGAGAGTATGCT
CGTATGCTTGAAGTTCATGAGCAAATAGGTCATGTGGCAATCCCAAGTATTGGGGTTGATATTCCAATTTATGCT
GGAACATCCGAAACTGTGCTTCAGAAAGGTAGTGGGCATTTGGAGGGAACCAGTCTTCCAGTGGGAGGTTTGTCA
ACCCATTCAGTACTAACTGCCCACCGTGGCTTGCCAACAGCTAGGCTATTTACCGACTTAAATAAAGTTAAAAAA
GGCCAGATTTTCTATGTGACGAACATCAAGGAAACACTTGCCTACAAAGTCGTGTCTATCAAAGTTGTGGATCCA
ACAGCTTTAAGTGAGGTTAAGATTGTCAATGGTAAGGATTATATAACCTTGCTGACTTGCACACCTTACATGATC
AATAGTCATCGTCTCTTGGTAAAAGGAGAGCGTATTCCTTATGATTCTACCGAGGCGGAAAAGCACAAAGAACAA
ACCGTACAAGATTATCGTTTGTCACTAGTGTTGAAGATACTACTAGTATTATTAATTGGACTCTTCATCGTGATA
ATGATGAGAAGATGGATGCAACATCGTCAATAA

SEQ ID NO: 25

MGSLILLFPIVSQVSYYLASHQNINQFKREVAKIDTNTVERRIALANAYNETLSRNPLLIDPFTSKQKEGLREYA
RMLEVHEQIGHVAIPSIGVDIPIYAGTSETVLQKGSGHLEGTSLPVGGLSTHSVLTAHRGLPTARLFTDLNKVKK
GQIFYVTNIKETLAYKVVSIKVVDPTALSEVKIVNGKDYITLLTCTPYMINSHRLLVKGERIPYDSTEAEKHKEQ
TVQDYRLSLVLKILLVLLIGLFIVIMMRRWMQHRQ

GBS 150

Examples of polynucleotide and amino acid sequences for GBS 150 are set forth below. SEQ ID NO: 26 and 27 represent GBS 150 sequences from GBS serotype V, strain isolate 2603.

SEQ ID NO: 26

ATGAAAAAGATTAGAAAAAGTTTAGGACTTCTACTATGTTGCTTTTTAGGATTGGTACAATTAGCGTTTTTTTCG
GTAGCCAGTGTAAATGCTGATACCCCTAATCAACTAACAATCACACAGATAGGACTTCAGCCAAATACTACAGAG
GAGGGGATTTCTTATCGTTTATGGACTGTGACTGACAACTTAAAAGTTGATTTATTGAGCCAAATGACAGATAGC
GAATTGAACCAGAAGTATAAGAGTATCTTGACTTCTCCTACTGATACTAATGGTCAGACAAAGATAGCACTCCCA
AATGGTTCGTACTTTGGTCGTGCTTATAAAGCTGATCAAAGCGTTTCAACAATAGTACCTTTTTATATTGAATTA
CCAGATGATAAGTTATCAAATCAATTACAGATAAATCCTAAGCGAAAAGTTGAAACAGGCCGATTAAAACTTATT

AAATATACAAAAGAAGGAAAGATAAAGAAAAGGCTATCCGGAGTAATATTTGTATTATACGATAACCAGAATCAG
CCAGTTCGCTTTAAAAATGGACGATTTACGACCGATCAAGATGGGATTACTTCATTAGTAACTGATGATAAGGGA
GAAATTGAGGTTGAAGGTTTATTACCTGGTAAGTATATTTTTCGAGAAGCAAAAGCACTAACTGGTTACCGTATA
TCTATGAAGGATGCTGTAGTTGCTGTAGTTGCTAATAAAACACAGGAAGTAGAGGTAGAAAACGAAAAAGAAACT
CCTCCACCAACAAATCCTAAACCATCACAACCGCTTTTTCCACAATCATTTCTTCCTAAAACAGGAATGATTATT
GGTGGAGGACTGACAATTCTTGGTTGTATTATTTTGGGAATTTTGTTTATCTTTTTAAGAAAAACTAAAAATAGC
AAATCTGAAAGAAACGATACAGTA

SEQ ID NO: 27

MKKIRKSLGLLLCCFLGLVQLAFFSVASVNADTPNQLTITQIGLQPNTTEEGISYRLWTVTDNLKVDLLSQMTDS
ELNQKYKSILTSPTDTNGQTKIALPNGSYFGRAYKADQSVSTIVPFYIELPDDKLSNQLQINPKRKVETGRLKLI
KYTKEGKIKKRLSGVIFVLYDNQNQPVRFKNGRFTTDQDGITSLVTDDKGEIEVEGLLPGKYIFREAKALTGYRI
SMKDAVVAVVANKTQEVEVENEKETPPPTNPKPSQPLFPQS*FLPKTG*MIIGGGLTILGCIILGILFIFLRKTKNS
KSERNDTV

GBS 150 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 130** LPKTG (shown in italics in SEQ ID NO: 27 above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant GBS 150 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

As discussed above, a pilin motif, containing a conserved lysine (K) residue has been identified in GBS 150. The pilin motif sequence is underlined in SEQ ID NO: 27, below. Conserved lysine (K) residues are marked in bold, at amino acid residues 139 and 148. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures of GBS 150. Preferred fragments of GBS 150 include a conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 27

MKKIRKSLGLLLCCFLGLVQLAFFSVASVNADTPNQLTITQIGLQPNTTEEGISYRLWTVTDNLKVDLLSQMTDS
ELNQKYKSILTSPTDTNGQTKIALPNGSYFGRAYKADQSVSTIVPFYIELPDDKLSNQLQINP**K**RKVETGRL**K**LI
KYTKEGKIKKRLSGVIFVLYDNQNQPVRFKNGRFTTDQDGITSLVTDDKGEIEVEGLLPGKYIFREAKALTGYRI
SMKDAVVAVVANKTQEVEVENEKETPPPTNPKPSQPLFPQSFLPKTGMIIGGGLTILGCIILGILFIFLRKTKNS
KSERNDTV

An E box containing a conserved glutamic residue has also been identified in GBS 150. The E box motif is underlined in SEQ ID NO: 27 below. The conserved glutamic acid (E), at amino acid residue 216, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of GBS 150. Preferred fragments of GBS 150 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 27

MKKIRKSLGLLLCCFLGLVQLAFFSVASVNADTPNQLTITQIGLQPNTTEEGISYRLWTVTDNLKVDLLSQMTDS
ELNQKYKSILTSPTDTNGQTKIALPNGSYFGRAYKADQSVSTIVPFYIELPDDKLSNQLQINPKRKVETGRLKLI
KYTKEGKIKKRLSGVIFVLYDNQNQPVRFKNGRFTTDQDGITSLVTDDKGEIEVEGLLPGKYIFR**E**AKALTGYRI
SMKDAVVAVVANKTQEVEVENEKETPPPTNPKPSQPLFPQSFLPKTGMIIGGGLTILGCIILGILFIFLRKTKNS
KSERNDTV

SAG1405

Examples of polynucleotide and amino acid sequences for SAG1405 are set forth below.

SEQ ID NO: 28 and 29 represent SAG1405 sequences from GBS serotype V, strain isolate 2603.

SEQ ID NO: 28

ATGGGAGGAAAATTTCAGAAAAACCTTAAGAAATCGGTCGTTTTAAATCGATGGATGAATGTAGGCTTGATACTA
TTGTTCTTAGTTGGTCTTTTGATAACCTCATATCCTTTTATTTCAAATTGGTACTATAATATTAAAGCTAATAAT
CAAGTAACTAACTTTGATAATCAAACCCAAAAATTAAATACTAAAGAGATTAATAGACGATTTGAGTTAGCAAAA
GCTTATAATAGAACACTGGACCCAAGCCGCCTATCAGATCCCTATACTGAAAAAGAAAAAAAAGGTATTGCTGAA
TACGCCCACATGCTTGAGATTGCTGAAATGATTGGATATATTGATATACCGTCTATCAAGCAAAAATTACCTATC
TATGCGGGGACTACCAGTAGTGTTCTTGAAAAAGGAGCAGGACACCTTGAAGGAACCTCCTTGCCAATTGGTGGA
AAAAGTTCACATACTGTTATCACAGCTCATCGCGGCTTACCTAAAGCTAAGTTATTTACAGATTTAGATAAACTT
AAAAAAGGAAAAATTTTTTATATTCATAATATCAAAGAAGTTTTAGCCTATAAGGTTGATCAAATAAGTGTTGTA
AAGCCAGATAATTTTTCTAAATTATTGGTTGTTAAAGGTAAGGATTATGCGACTTTGCTAACATGTACACCTTAT
TCGATTAATTCACATCGTTTACTAGTTAGAGGGCATCGAATCAAGTATGTACCTCCTGTTAAAGAAAAGAACTAT
TTAATGAAAGAATTGCAAACACACTATAAACTTTATTTCCTCTTATCAATCCTAGTTATTCTTATATTAGTCGCT
TTACTATTATATTTAAAACGAAAATTTAAAGAGAGAAAGAGAAAGGGAAATCAAAAATGA

SEQ ID NO: 29

MGGKFQKNLKKSVVLNRWMNVGLILLFLVGLLITSYPFISNWYYNIKANNQVTNFDNQTQKLNTKEINRRFELAK
AYNRTLDPSRLSDPYTEKEKKGIAEYAHMLEIAEMIGYIDIPSIKQKLPIYAGTTSSVLEKGAGHLEGTSLPIGG
KSSHTVITAHRGLPKAKLFTDLDKLKKGKIFYIHNIKEVLAYKVDQISVVKPDNFSKLLVVKGKDYATLLTCTPY
SINSHRLLVRGHRIKYVPPVKEKNYLMKELQTHYKLYFLLSILVILILVALLLYLKRKFKERKRKGNQK

SAG1406

Examples of polynucleotide and amino acid sequences for SAG1405 are set forth below.

SEQ ID NO: 30 and 31 represent SAG1405 sequences from GBS serotype V, strain isolate 2603.

SEQ ID NO: 30

GTGAAGACTAAAAAAATCATCAAAAAAACAAAAAAAAAAGAAGAAGTCAAATCTTCCTTTTATCATTCTTTTTCTA
ATAGGTCTATCTATTTTATTGTATCCAGTGGTATCACGTTTTTACTATACGATAGAATCTAATAATCAAACACAG
GATTTTGAGAGAGCTGCTAAAAAACTTAGTCAGAAAGAAATCAATCGACGTATGGCTCTAGCACAAGCTTATAAT
GATTCTTTAAATAATGTCCATCTTGAAGATCCTTATGAGAAAAAACGAATTCAAAAGGGGGTAGCAGAGTACGCC
CGTATGTTAGAGGTAAGTGAAAAAATCGGAACAATTTCAGTTCCTAAGATAGGTCAAAAACTCCCTATATTTGCA
GGTTCAAGTCAAGAAGTTCTATCTAAAGGAGCAGGGCATTTAGAAGGTACCTCTCTTCCAATTGGGGGCAATAGT
ACACATACTGTTATAACAGCGCATTCAGGAATTCCAGATAAAGAACTCTTTTCTAACCTTAAAAAGTTAAAAAAA
GGAGATAAGTTTTATATTCAAAACATAAAAGAAACGATAGCATATCAAGTAGATCAGATAAAAGTCGTTACACCC
GATAACTTTTCAGATTTGTTGGTTGTTCCTGGACATGATTATGCAACCTTATTGACTTGCACCCCGATTATGATC
AATACACACAGACTTTTAGTAAGGGGACATCGTATCCCTTATAAAGGTCCTATTGATGAAAAATTAATAAAAGAC
GGTCATTTAAACACGATTTATAGATATCTATTCTATATATCTTTAGTTATTATTGCTTGGTTACTTTGGTTAATA
AAACGTCAACGTCAAAAAAATCGTTTAGCAAGTGTTAGAAAAGGAATTGAATCATAA

SEQ ID NO: 31

MKTKKIIKKTKKKKKSNLPFIILFLIGLSILLYPVVSRFYYTIESNNQTQDFERAAKKLSQKEINRRMALAQAYN
DSLNNVHLEDPYEKKRIQKGVAEYARMLEVSEKIGTISVPKIGQKLPIFAGSSQEVLSKGAGHLEGTSLPIGGNS
THTVITAHSGIPDKELFSNLKKLKKGDKFYIQNIKETIAYQVDQIKVVTPDNFSDLLVVPGHDYATLLTCTPIMI
NTHRLLVRGHRIPYKGPIDEKLIKDGHLNTIYRYLFYISLVIIAWLLWLIKRQRQKNRLASVRKGIES

01520

An example of an amino acid sequence for 01520 is set forth below. SEQ ID NO: 32 represents a 01520 sequence from GBS serotype III, strain isolate COH1.

SEQ ID NO: 32

MIRRYSANFLAILGIILVSSGIYWGWYNINQAHQADLTSQHIVKVLDKSITHQVKGSENGELPVKKLDKTDYLGT
LDIPNLKLHLPVAANYSFEQLSKTPTRYYGSYLTNNMVICAHNFPYHFDALKNVDMGTDVYFTTTTGQIYHYKIS
NREIIEPTAIEKVYKTATSDNDWDLSLFTCTKAGVARVLVRCQLIDVKN

01521

An example of an amino acid sequence for 01521 is set forth below. SEQ ID NO: 33 represents a 01521 sequence from GBS serotype III, strain isolate COH1.

SEQ ID NO: 33

MIYKKILKITLLLLFSLSTQLVSADTNDQMKTGSITIQNKYNNQGIAGGNLLVYQVAQAKDVDGNQVFTLTTPFQ
GIGIKDDDLTQVNLDSNQAKYVNLLTKAVHKTQPLQTFDNLPAEGIVANNLPQGIYLFIQTKTAQGYELMSPFIL
SIPKDGKYDITAFEKMSPLNAKPKKEETITPTVTHQTKGK*LPFTG*QVWWPIPILIMSGLLCLIIALKWRRRRD

01521 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 132** LPFTG (shown in italics in SEQ ID NO: 33 above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant 01521 protein from the host cell. Alternatively, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Two pilin motifs, containing conserved lysine (K) residues have been identified in 01521. The pilin motif sequences are underlined in SEQ ID NO: 33, below. Conserved lysine (K) residues are marked in bold, at amino acid residues 154 and 165 and at amino acid residues 174 and 188. The pilin sequences, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures of 01521. Preferred fragments of 01521 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 33

MIYKKILKITLLLLFSLSTQLVSADTNDQMKTGSITIQNKYNNQGIAGGNLLVYQVAQAKDVDGNQVFTLTTPFQ
GIGIKDDDLTQVNLDSNQAKYVNLLTKAVHKTQPLQTFDNLPAEGIVANNLPQGIYLFIQTKTAQGYELMSPFIL
SIP**K**DGKYDITAFE**K**MSPLNAKP**K**KEETITPTVTHQT**K**GKLPFTGQVWWPIPILIMSGLLCLIIALKWRRRRD

An E box containing a conserved glutamic residue has also been identified in 01521. The E box motif is underlined in SEQ ID NO: 33 below. The conserved glutamic acid (E), at amino acid residue 177, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of 01521. Preferred fragments of 01521 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 33

MIYKKILKITLLLLFSLSTQLVSADTNDQMKTGSITIQNKYNNQGIAGGNLLVYQVAQAKDVDGNQVFTLTTPFQ
GIGIKDDDLTQVNLDSNQAKYVNLLTKAVHKTQPLQTFDNLPAEGIVANNLPQGIYLFIQTKTAQGYELMSPFIL
SIPKDGKYDITAFEKMSPLNAKPKKE**E**TITPTVTHQTKGKLPFTGQVWWPIPILIMSGLLCLIIALKWRRRRD

01522

An example of an amino acid sequence for 01522 is set forth below. SEQ ID NO: 34 represents a 01522 sequence from GBS serotype III, strain isolate COH1.

SEQ ID NO: 34

MAYPSLANYWNSFHQSRAIMDYQDRVTHMDENDYKKIINRAKEYNKQFKTSGMKWHMTSQERLDYNSQLAIDKTG
NMGYISIPKINIKLPLYHGTSEKVLQTSIGHLEGSSLPIGGDSTHSILSGHRGLPSSRLFSDLDKLKVGDHWTVS
ILNETYTYQVDQIRTVKPDDLRDLQIVKGKDYQTLVTCTPYGVNTHRLLVRGHRVPNDNGNALVVAEAIQIEPIY
IAPFIAIFLTLILLLISLEVTRRARQRKKILKQAMRKEENNDL

01523

An example of an amino acid sequence for 01523 is set forth below. SEQ ID NO: 35 represents a 01523 sequence from GBS serotype III, strain isolate COH1.

SEQ ID NO: 35

MKKKMIQSLLVASLAFGMAVSPVTPIAFAAETGTITVQDTQKGATYKAYKVFDAEIDNANVSDSNKDGASYLIPQ
GKEAEYKASTDFNSLFTTTTNGGRTYVTKKDTASANEIATWAKSISANTTPVSTVTESNNDGTEVINVSQYGYYY
VSSTVNNGAVIMVTSVTPNATIHEKNTDATWGDGGGKTVDQKTYSVGDTVKYTITYKNAVNYHGTEKVYQYVIKD
TMPSASVVDLNEGSYEVTITDGSGNITTLTQGSEKATGKYNLLEENNNFTITIPWAATNTPTGNTQNGANDDFFY
KGINTITVTYTGVLKSGAKPGSADLPENTNIATINPNTSNDDPGQKVTVRDGQITIKKIDGSTKASLQGAIFVLK
NATGQFLNFNDTNNVEWGTEANATEYTTGADGIITITGLKEGTYYLVEKKAPLGYNLLDNSQKVILGDGATDTTN
SDNLLVNPTVENNKGTE*LPSTG*GIGTTIFYIIGAILVIGAGIVLVARRRLRS

01523 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 131** LPSTG (shown in italics in SEQ ID NO: 35 above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant 01523 protein from the host cell. Alternatively, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

An E box containing a conserved glutamic residue has also been identified in 01523. The E box motif is underlined in SEQ ID NO: 35 below. The conserved glutamic acid (E), at amino acid residue 423, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of 01523. Preferred fragments of 01523 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 35

MKKKMIQSLLVASLAFGMAVSPVTPIAFAAETGTITVQDTQKGATYKAYKVFDAEIDNANVSDSNKDGASYLIPQ
GKEAEYKASTDFNSLFTTTTNGGRTYVTKKDTASANEIATWAKSISANTTPVSTVTESNNDGTEVINVSQYGYYY
VSSTVNNGAVIMVTSVTPNATIHEKNTDATWGDGGGKTVDQKTYSVGDTVKYTITYKNAVNYHGTEKVYQYVIKD
TMPSASVVDLNEGSYEVTITDGSGNITTLTQGSEKATGKYNLLEENNNFTITIPWAATNTPTGNTQNGANDDFFY
KGINTITVTYTGVLKSGAKPGSADLPENTNIATINPNTSNDDPGQKVTVRDGQITIKKIDGSTKASLQGAIFVLK
NATGQFLNFNDTNNVEWGTEANATEYTTGADGIITITGLKEGTYYLV**E**KKAPLGYNLLDNSQKVILGDGATDTTN
SDNLLVNPTVENNKGTELPSTGGIGTTIFYIIGAILVIGAGIVLVARRRLRS

01524

An example of an amino acid sequence for 01524 is set forth below. SEQ ID NO: 36 represents a 01524 sequence from GBS serotype III, strain isolate COH1.

SEQ ID NO: 36

MLKKCQTFIIESLKKKKHPKEWKIIMWSLMILTTFLTTYFLILPAITVEETKTDDVGITLENKNSSQVTSSTSSS
QSSVEQSKPQTPASSVTETSSSEEAAYREEPLMFRGADYTVTVTLTKEAKIPKNADLKVTELKDNSATFKDYKKK
ALTEVAKQDSEIKNFKLYDITIESNGKEAEPQAPVKVEVNYDKPLEASDENLKVVHFKDDGQTEVLKSKDTAETK
NTSSDVAFKTDSFSIYAIVQEDNTEVPRLTYHFQNNDGTDYDFLTASGMQVHHQIIKDGESLGEVGIPTIKAGEH
FNGWYTYDPTTGKYGDPVKFGEPITVTETKEICVRPFMSKVATVTLYDDSAGKSILERYQVPLDSSGNGTADLSS
FKVSPPTSTLLFVGWSKTQNGAPLSESEIQALPVSSDISLYPVFKESYGVEFNTGDLSTGVTYIAPRRVLTGQPA
STIKPNDPTRPGYTFAGWYTAASGGAAFDFNQVLTKDTTLYAHWSPAQTTYTINYWQQSATDNKNATDAQKTYEY
AGQVTRSGLSLSNQTLTQQDINDKLPTGFKVNNTRTETSVMIKDDGSSVVNVYYDRKLITIKFAKYGGYSLPEYY
YSYNWSSDADTYTGLYGTTLAANGYQWKTGAWGYLANVGNNQVGTYGMSYLGEFILPNDTVDSDVIKLFPKGNIV
QTYRFFKQGLDGTYSLADTGGGAGADEFTFTEKYLGFNVKYYQRLYPDNYLFDQYASQTSAGVKVPISDEYYDRY
GAYHKDYLNLVVWYERNSYKIKYLDPLDNTELPNFPVKDVLYEQNLSSYAPDTTTVQPKPSRPGYVWDGKWYKDQ
AQTQVFDFNTTMPPHDVKVYAGWQKVTYRVNIDPNGGRLSKTDDTYLDLHYGDRIPDYTDITRDYIQDPSGTYYY
KYDSRDKDPDSTKDAYYTTDTSLSNVDTTTKYKYVKDAYKLVGWYYVNPDGSIRPYNFSGAVTQDINLRAIWRKA

GDYHIIYSNDAVGTDGKPALDASGQQLQTSNEPTDPDSYDDGSHSALLRRPTMPDGYRFRGWWYNGKIYNPYDSI
DIDAHLADANKNITIKPVIIPVGDIKLEDTSIKYNGNGGTRVENGNVVTQVETPRMELNSTTTIPENQYFTRTGY
NLIGWHHDKDLADTGRVEFTAGQSIGIDNNPDATNTLYAVWQPKEYTVRVSKTVVGLDEDKTKDFLFNPSETLQQ
ENFPLRDGQTKEFKVPYGTSISIDEQAYDEFKVSESITEKNLATGEADKTYDATGLQSLTVSGDVDISFTNTRIK
QKVRLQKVNVENDNNFLAGAVFDIYESDANGNKASHPMYSGLVTNDKGLLLVDANNYLSLPVGKYYLTETKAPPG
YLLPKNDISVLVISTGVTFEQNGNNATPIKENLVDGSTVYTFKITNSKGTE*LPSTG*GIGTHIYILVGLALALPSG
LILYYRKKI

01524 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 131** LPSTG (shown in italics in SEQ ID NO: 36 above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant 01524 protein from the host cell. Alternatively, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Three pilin motifs, containing conserved lysine (K) residues have been identified in 01524. The pilin motif sequences are underlined in SEQ ID NO: 36, below. Conserved lysine (K) residues are marked in bold, at amino acid residues 128 and 138, amino acid residues 671 and 682, and amino acid residues 809 and 820. The pilin sequences, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures of 01524. Preferred fragments of 01524 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 36

MLKKCQTFIIESLKKKKHPKEWKIIMWSLMILTTFLTTYFLILPAITVEETKTDDVGITLENKNSSQVTSSTSSS
QSSVEQSKPQTPASSVTETSSSEEAAYREEPLMFRGADYTVTVTLTKEAKIPKNADLKVTELKDNSATFKDYKKK
ALTEVAKQDSEIKNFKLYDITIESNGKEAEPQAPVKVEVNYDKPLEASDENLKVVHFKDDGQTEVLKSKDTAETK
NTSSDVAFKTDSFSIYAIVQEDNTEVPRLTYHFQNNDGTDYDFLTASGMQVHHQIIKDGESLGEVGIPTIKAGEH
FNGWYTYDPTTGKYGDPVKFGEPITVTETKEICVRPFMSKVATVTLYDDSAGKSILERYQVPLDSSGNGTADLSS
FKVSPPTSTLLFVGWSKTQNGAPLSESEIQALPVSSDISLYPVFKESYGVEFNTGDLSTGVTYIAPRRVLTGQPA
STIKPNDPTRPGYTFAGWYTAASGGAAFDFNQVLTKDTTLYAHWSPAQTTYTINYWQQSATDNKNATDAQKTYEY
AGQVTRSGLSLSNQTLTQQDINDKLPTGFKVNNTRTETSVMIKDDGSSVVNVYYDRKLITIKFAKYGGYSLPEYY
YSYNWSSDADTYTGLYGTTLAANGYQWKTGAWGYLANVGNNQVGTYGMSYLGEFILPNDTVDSDVIKLFPKGNIV
QTYRFFKQGLDGTYSLADTGGGAGADEFTFTEKYLGFNVKYYQRLYPDNYLFDQYASQTSAGVKVPISDEYYDRY
GAYHKDYLNLVVWYERNSYKIKYLDPLDNTELPNFPVKDVLYEQNLSSYAPDTTTVQPKPSRPGYVWDGKWYKDQ
AQTQVFDFNTTMPPHDVKVYAGWQKVTYRVNIDPNGGRLSKTDDTYLDLHYGDRIPDYTDITRDYIQDPSGTYYY
KYDSRDKDPDSTKDAYYTTDTSLSNVDTTTKYKYVKDAYKLVGWYYVNPDGSIRPYNFSGAVTQDINLRAIWRKA
GDYHIIYSNDAVGTDGKPALDASGQQLQTSNEPTDPDSYDDGSHSALLRRPTMPDGYRFRGWWYNGKIYNPYDSI
DIDAHLADANKNITIKPVIIPVGDIKLEDTSIKYNGNGGTRVENGNVVTQVETPRMELNSTTTIPENQYFTRTGY
NLIGWHHDKDLADTGRVEFTAGQSIGIDNNPDATNTLYAVWQPKEYTVRVSKTVVGLDEDKTKDFLFNPSETLQQ
ENFPLRDGQTKEFKVPYGTSISIDEQAYDEFKVSESITEKNLATGEADKTYDATGLQSLTVSGDVDISFTNTRIK
QKVRLQKVNVENDNNFLAGAVFDIYESDANGNKASHPMYSGLVTNDKGLLLVDANNYLSLPVGKYYLTETKAPPG
YLLPKNDISVLVISTGVTFEQNGNNATPIKENLVDGSTVYTFKITNSKGTELPSTGGIGTHIYILVGLALALPSG
LILYYRKKI

An E box containing a conserved glutamic residue has also been identified in 01524. The E box motif is underlined in SEQ ID NO: 36 below. The conserved glutamic acid (E), at amino acid residue 1344, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of 01524. Preferred

fragments of 01324 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 36

MLKKCQTFIIESLKKKKHPKEWKIIMWSLMILTTFLTTYFLILPAITVEETKTDDVGITLENKNSSQVTSSTSSS
QSSVEQSKPQTPASSVTETSSSEEAAYREEPLMFRGADYTVTVTLTKEAKIPKNADLKVTELKDNSATFKDYKKK
ALTEVAKQDSEIKNFKLYDITIESNGKEAEPQAPVKVEVNYDKPLEASDENLKVVHFKDDGQTEVLKSKDTAETK
NTSSDVAFKTDSFSIYAIVQEDNTEVPRLTYHFQNNDGTDYDFLTASGMQVHHQIIKDGESLGEVGIPTIKAGEH
FNGWYTYDPTTGKYGDPVKFGEPITVTETKEICVRPFMSKVATVTLYDDSAGKSILERYQVPLDSSGNGTADLSS
FKVSPPTSTLLFVGWSKTQNGAPLSESEIQALPVSSDISLYPVFKESYGVEFNTGDLSTGVTYIAPRRVLTGQPA
STIKPNDPTRPGYTFAGWYTAASGGAAFDFNQVLTKDTTLYAHWSPAQTTYTINYWQQSATDNKNATDAQKTYEY
AGQVTRSGLSLSNQTLTQQDINDKLPTGFKVNNTRTETSVMIKDDGSSVVNVYYDRKLITIKFAKYGGYSLPEYY
YSYNWSSDADTYTGLYGTTLAANGYQWKTGAWGYLANVGNNQVGTYGMSYLGEFILPNDTVDSDVIKLFPKGNIV
QTYRFFKQGLDGTYSLADTGGGAGADEFTFTEKYLGFNVKYYQRLYPDNYLFDQYASQTSAGVKVPISDEYYDRY
GAYHKDYLNLVVWYERNSYKIKYLDPLDNTELPNFPVKDVLYEQNLSSYAPDTTTVQPKPSRPGYVWDGKWYKDQ
AQTQVFDFNTTMPPHDVKVYAGWQKVTYRVNIDPNGGRLSKTDDTYLDLHYGDRIPDYTDITRDYIQDPSGTYYY
KYDSRDKDPDSTKDAYYTTDTSLSNVDTTTKYKYVKDAYKLVGWYYVNPDGSIRPYNFSGAVTQDINLRAIWRKA
GDYHIIYSNDAVGTDGKPALDASGQQLQTSNEPTDPDSYDDGSHSALLRRPTMPDGYRFRGWWYNGKIYNPYDSI
DIDAHLADANKNITIKPVIIPVGDIKLEDTSIKYNGNGGTRVENGNVVTQVETPRMELNSTTTIPENQYFTRTGY
NLIGWHHDKDLADTGRVEFTAGQSIGIDNNPDATNTLYAVWQPKEYTVRVSKTVVGLDEDKTKDFLFNPSETLQQ
ENFPLRDGQTKEFKVPYGTSISIDEQAYDEFKVSESITEKNLATGEADKTYDATGLQSLTVSGDVDISFTNTRIK
QKVRLQKVNVENDNNFLAGAVFDIYESDANGNKASHPMYSGLVTNDKGLLLVDANNYLSLPVGKYYLTETKAPPG
YLLPKNDISVLVISTGVTFEQNGNNATPIKENLVDGSTVYTFKITNSKGTELPSTGGIGTHIYILVGLALALPSG
LILYYRKKI

01525

An example of an amino acid sequence for 01525 is set forth below. SEQ ID NO: 37 represents a 01525 sequence from GBS serotype III, strain isolate COH1.

SEQ ID NO: 37

MKRQISSDKLSQELDRVTYQKRFWSVIKNTIYILMAVASIAILIAVLWLPVLRIYGHSMNKTLSAGDVVFTVKGS
NFKTGDVVAFYYNNKVLVKRVIAESGDWVNIDSQGDVYVNQHKLKEPYVIHKALGNSNIKYPYQVPDKKIFVLGD
NRKTSIDSRSTSVGDVSEEQIVGKISFRIWPLGKISSIN

GBS 322

GBS 322 refers to a surface immunogenic protein, also referred to as "sip". Nucleotide and amino acid sequences of GBS 322 sequenced from serotype V isolated strain 2603 V/R are set forth in Ref. 3 as SEQ ID 8539 and SEQ ID 8540. These sequences are set forth below as SEQ ID NOS 38 and 39:

SEQ ID NO. 38

ATGAATAAAAAGGTACTATTGACATCGACAATGGCAGCTTCGCTATTATCAGTCGCAAGTGTTCAAGCACAAGAA
ACAGATACGACGTGGACAGCACGTACTGTTTCAGAGGTAAAGGCTGATTTGGTAAAGCAAGACAATAAATCATCA
TATACTGTGAAATATGGTGATACACTAAGCGTTATTTCAGAAGCAATGTCAATTGATATGAATGTCTTAGCAAAA
ATAAATAACATTGCAGATATCAATCTTATTTATCCTGAGACAACACTGACAGTAACTTACGATCAGAAGAGTCAT
ACTGCCACTTCAATGAAAATAGAAACACCAGCAACAAATGCTGCTGGTCAAACAACAGCTACTGTGGATTTGAAA
ACCAATCAAGTTTCTGTTGCAGACCAAAAAGTTTCTCTCAATACAATTTCGGAAGGTATGACACCAGAAGCAGCA
ACAACGATTGTTTCGCCAATGAAGACATATTCTTCTGCGCCAGCTTTGAAATCAAAAGAAGTATTAGCACAAGAG
CAAGCTGTTAGTCAAGCAGCAGCTAATGAACAGGTATCACCAGCTCCTGTGAAGTCGATTACTTCAGAAGTTCCA
GCAGCTAAAGAGGAAGTTAAACCAACTCAGACGTCAGTCAGTCAGTCAACAACAGTATCACCAGCTTCTGTTGCC
GCTGAAACACCAGCTCCAGTAGCTAAAGTAGCACCGGTAAGAACTGTAGCAGCCCCTAGAGTGGCAAGTGTTAAA
GTAGTCACTCCTAAAGTAGAAACTGGTGCATCACCAGAGCATGTATCAGCTCCAGCAGTTCCTGTGACTACGACT
TCACCAGCTACAGACAGTAAGTTACAAGCGACTGAAGTTAAGAGCGTTCCGGTAGCACAAAAAGCTCCAACAGCA
ACACCGGTAGCACAACCAGCTTCAACAACAAATGCAGTAGCTGCACATCCTGAAAATGCAGGGCTCCAACCTCAT
GTTGCAGCTTATAAAGAAAAAGTAGCGTCAACTTATGGAGTTAATGAATTCAGTACATACCGTGCGGGAGATCCA
GGTGATCATGGTAAAGGTTTAGCAGTTGACTTTATTGTAGGTACTAATCAAGCACTTGGTAATAAAGTTGCACAG
TACTCTACACAAAATATGGCAGCAAATAACATTTCATATGTTATCTGGCAACAAAAGTTTTACTCAAATACAAAC
AGTATTTATGGACCTGCTAATACTTGGAATGCAATGCCAGATCGTGGTGGCGTTACTGCCAACCACTATGACCAC

GTTCAGGTATCATTAACAAATAATATAAAAAAGGAAGCTATTTGGCTTCTTTTTTATATGCCTTGAATAGACTT
TCAAGGTTCTTATATAATTTTTATTA

SEQ ID NO. 39

MNKKVLLTSTMAASLLSVASVQAQETDTTWTARTVSEVKADLVKQDNKSSYTVKYGDTLSVISEAMSIDMNVLAK
INNIADINLIYPETTLTVTYDQKSHTATSMKIETPATNAAGQTTATVDLKTNQVSVADQKVSLNTISEGMTPEAA
TTIVSPMKTYSSAPALKSKEVLAQEQAVSQAAANEQVSPAPVKSITSEVPAAKEEVKPTQTSVSQSTTVSPASVA
AETPAPVAKVAPVRTVAAPRVASVKVVTPKVETGASPEHVSAPAVPVTTTSPATDSKLQATEVKSVPVAQKAPTA
TPVAQPASTTNAVAAHPENAGLQPHVAAYKEKVASTYGVNEFSTYRAGDPGDHGKGLAVDFIVGTNQALGNKVAQ
YSTQNMAANNISYVIWQQKFYSNTNSIYGPANTWNAMPDRGGVTANHYDHVHVSFNK

GBS 322 contains an N-terminal leader or signal sequence region which is indicated by the underlined sequence near the beginning of SEQ ID NO: 39. In one embodiment, one or more amino acids from the leader or signal sequence region of GBS 322 are removed. An example of such a GBS 322 fragment is set forth below as SEQ ID NO: 40.

SEQ ID NO: 40

DLVKQDNKSSYTVKYGDTLSVISEAMSIDMNVLAKINNIADINLIYPETTLTVTYDQKSHTATSMKIETPATNAA
GQTTATVDLKTNQVSVADQKVSLNTISEGMTPEAATTIVSPMKTYSSAPALKSKEVLAQEQAVSQAAANEQVSPA
PVKSITSEVPAAKEEVKPTQTSVSQSTTVSPASVAAETPAPVAKVAPVRTVAAPRVASVKVVTPKVETGASPEHV
SAPAVPVTTTSPATDSKLQATEVKSVPVAQKAPTATPVAQPASTTNAVAAHPENAGLQPHVAAYKEKVASTYGVN
EFSTYRAGDPGDHGKGLAVDFIVGTNQALGNKVAQYSTQNMAANNISYVIWQQKFYSNTNSIYGPANTWNAMPDR
GGVTANHYDHVHVSFNK

Additional preferred fragments of GBS 322 comprise the immunogenic epitopes identified in WO 03/068813, each of which are specifically incorporated by reference herein.

There may be an upper limit to the number of GBS proteins which will be in the compositions of the invention. Preferably, the number of GBS proteins in a composition of the invention is less than 20, less than 19, less than 18, less than 17, less than 16, less than 15, less than 14, less than 13, less than 12, less than 11, less than 10, less than 9, less than 8, less than 7, less than 6, less than 5, less than 4, or less than 3. Still more preferably, the number of GBS proteins in a composition of the invention is less than 6, less than 5, or less than 4. Still more preferably, the number of GBS proteins in a composition of the invention is 3.

The GBS proteins and polynucleotides used in the invention are preferably isolated, *i.e.*, separate and discrete, from the whole organism with which the molecule is found in nature or, when the polynucleotide or polypeptide is not found in nature, is sufficiently free of other biological macromolecules so that the polynucleotide or polypeptide can be used for its intended purpose.

Group A Streptococcus Adhesin Island Sequences

The GAS AI polypeptides of the invention can, of course, be prepared by various means (*e.g.* recombinant expression, purification from GAS, chemical synthesis *etc.*) and in various forms (*e.g.* native, fusions, glycosylated, non-glycosylated *etc.*). They are preferably prepared in substantially pure form (*i.e.* substantially free from other streptococcal or host cell proteins) or substantially isolated form.

The GAS AI proteins of the invention may include polypeptide sequences having sequence identity to the identified GAS proteins. The degree of sequence identity may vary depending on the amino acid sequence (a) in question, but is preferably greater than 50% (*e.g.* 60%, 65%, 70%, 75%,

80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.5% or more). Polypeptides having sequence identity include homologs, orthologs, allelic variants and functional mutants of the identified GBS proteins. Typically, 50% identity or more between two proteins is considered to be an indication of functional equivalence. Identity between proteins is preferably determined by the Smith-Waterman homology search algorithm as implemented in the MPSRCH program (Oxford Molecular), using an affinity gap search with parameters *gap open penalty=12* and *gap extension penalty=1*.

The GAS adhesin island polynucleotide sequences may include polynucleotide sequences having sequence identity to the identified GAS adhesin island polynucleotide sequences. The degree of sequence identity may vary depending on the polynucleotide sequence in question, but is preferably greater than 50% (*e.g.* 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.5% or more).

The GAS adhesin island polynucleotide sequences of the invention may include polynucleotide fragments of the identified adhesin island sequences. The length of the fragment may vary depending on the polynucleotide sequence of the specific adhesin island sequence, but the fragment is preferably at least 10 consecutive polynucleotides, (*e.g.* at least 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200 or more).

The GAS adhesin island amino acid sequences of the invention may include polypeptide fragments of the identified GAS proteins. The length of the fragment may vary depending on the amino acid sequence of the specific GAS antigen, but the fragment is preferably at least 7 consecutive amino acids, (*e.g.* 8, 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200 or more). Preferably the fragment comprises one or more epitopes from the sequence. Other preferred fragments include (1) the N-terminal signal peptides of each identified GAS protein, (2) the identified GAS protein without their N-terminal signal peptides, and (3) each identified GAS protein wherein up to 10 amino acid residues (e.g. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25 or more) are deleted from the N-terminus and/or the C-terminus *e.g.* the N-terminal amino acid residue may be deleted. Other fragments omit one or more domains of the protein (*e.g.* omission of a signal peptide, of a cytoplasmic domain, of a transmembrane domain, or of an extracellular domain).

GAS AI-1 sequences

As discussed above, a GAS AI-1 sequence is present in an M6 strain isolate (MGAS10394). Examples of GAS AI-1 sequences from M6 strain isolate MGAS10394 are set forth below.

M6_Spy0156: Spy0156 is a rofA transcriptional regulator. An example of an amino acid sequence for M6_Spy0156 is set forth in SEQ ID NO: 41.

SEQ ID NO: 41

MIEKYLESSIESKCQLVVLFFKTSYLPITEVAEKTGLTFLQLNHYCEELNAFFPDSLSMTIQKRMISCQFTHPFK
ETYLYQLYASSNVLQLLAFLIKNGSHSRPLTDFARSHFLSNSSAYRMREALIPLLRNFELKLSKNKIVGEEYRIR
YLIALLYSKFGIKVYDLTQQDKNTIHSFLSHSSTHLKTSPWLSESFSFYDILLALSWKRHQFSVTIPQTRIFQQL
KKLFIYDSLKKSSRDIIETYCQLNFSAGDLDYLYLIYITANNSFASLQWTPEHIRQCCQLFEENDTFRLLLKPII
TLLPNLKEQKPSLVKALMFFSKSFLFNLQHFIPETNLFVSPYYKGNQKLYTSLKLIVEEWLAKLPGKRYLNHKHF
HLFCHYVEQILRNIQPPLVVVFVASNFINAHLLTDSFPRYFSDKSIDFHSYIAR

PCT/US2005/027239

M6_Spy0157: M6_Spy0157 is a fibronectin binding protein. It contains a sortase substrate motif LPXTG (SEQ ID NO: 122), shown in italics in the amino acid sequence SEQ ID NO: 42 .

SEQ ID NO: 42

MVSSYMFVRGEKMNNKIFLNKEASFLAHTKRKRRFAVTLVGVFFMLLACAGAIGFGQVAYAADEKTVPSHSSPNP
EFPWYGYDAYGKEYPGYNIWTRYHDLRVNLNGSRSYQVYCFNIQSNYPSQKNSFIKNWFKKIEGNGKSFVDYAHT
TKLGKEELEQRLLSLLYNAYPNDANGYMKGLEHLNAITVTQYAVWHYSDNSQYQFETLWESEAKEGKISRSQVTL
MREALKKLIDPNLEATAVNKIPSGYRLNIFESENEAYQNLLSAEYVPDDPPKPGETSEHNPKTPELDGTPIPEDP
KHPDDNLEPTLPPVMLDGEEVPEVPSESLEPALPPLMPELDGQEVPEKPSIDLPIEVPRYEFNNKDQSPLAGESG
ETEYITEVYGNQQNPVDIDKKLPNETGFSGNMVETEDTKEPEVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDT
KEPEVLMGGQSESVEFTKDTQTGMSGQTTPQIETEDTKEPEVLMGGQSESVEFTKDTQTGMSGQTTPQIETEDTK
EPEVLMGGQSESVEFTKDTQTGMSGFSETATVVEDTRPKLVFHFDNNEPKVEENREKPTKNITPI*LPATG*DIENV
LAFLGILILSVLSIFSLLKNKQSNKKV

M6_Spy0157 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 180** LPATG (shown in italics in SEQ ID NO: 42, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant M6_Spy0157 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in M6_Spy0157. The pilin motif sequence is underlined in SEQ ID NO: 42, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 277, 287, and 301. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of M6_Spy0157 include at least one conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 42

MVSSYMFVRGEKMNNKIFLNKEASFLAHTKRKRRFAVTLVGVFFMLLACAGAIGFGQVAYAADEKTVPSHSSPNP
EFPWYGYDAYGKEYPGYNIWTRYHDLRVNLNGSRSYQVYCFNIQSNYPSQKNSFIKNWFKKIEGNGKSFVDYAHT
TKLGKEELEQRLLSLLYNAYPNDANGYMKGLEHLNAITVTQYAVWHYSDNSQYQFETLWESEAKEGKISRSQVTL
MREALKKLIDPNLEATAVNKIPSGYRLNIFESENEAYQNLLSAE<u>YVPDDPP**K**PGETSEHNP**K**TPELDGTPIPEDP</u>
<u>**K**HPDDNLEPTLPPVMLDGEEVPEVPSESLEPALPPLMPELDGQEVPEKPSIDLPIEVPRYEFNNKDQSPLAGESG</u>
ETEYITEVYGNQQNPVDIDKKLPNETGFSGNMVETEDTKEPEVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDT
KEPEVLMGGQSESVEFTKDTQTGMSGQTTPQIETEDTKEPEVLMGGQSESVEFTKDTQTGMSGQTTPQIETEDTK
EPEVLMGGQSESVEFTKDTQTGMSGFSETATVVEDTRPKLVFHFDNNEPKVEENREKPTKNITPILPATGDIENV
LAFLGILILSVLSIFSLLKNKQSNKKV

A repeated series of four E boxes containing a conserved glutamic residue have been identified in M6_Spy0157. The E-box motifs are underlined in SEQ ID NO: 42, below. The conserved glutamic acid (E) residues, at amino acid residues 415, 452, 489, and 526 are marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of M6_Spy0157. Preferred fragments of M6_Spy0157 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 42

MVSSYMFVRGEKMNNKIFLNKEASFLAHTKRKRRFAVTLVGVFFMLLACAGAIGFGQVAYAADEKTVPSHSSPNP
EFPWYGYDAYGKEYPGYNIWTRYHDLRVNLNGSRSYQVYCFNIQSNYPSQKNSFIKNWFKKIEGNGKSFVDYAHT

TKLGKEELEQRLLSLLYNAYPNDANGYMKGLEHLNAITVTQYAVWHYSDNSQYQFETLWESEAKEGKISRSQVTL
MREALKKLIDPNLEATAVNKIPSGYRLNIFESENEAYQNLLSAEYVPDDPPKPGETSEHNPKTPELDGTPIPEDP
KHPDDNLEPTLPPVMLDGEEVPEVPSESLEPALPPLMPELDGQEVPEKPSIDLPIEVPRYEFNNKDQSPLAGESG
ETEYITEVYGNQQNPVDIDKKLPNETGFSGNMVETEDTKEPEVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDT
KEPEVLMGGQSESVEFTKDTQTGMSGQTTPQIETEDTKEPEVLMGGQSESVEFTKDTQTGMSGQTTPQIETEDTK
EPEVLMGGQSESVEFTKDTQTGMSGFSETATVVEDTRPKLVFHFDNNEPKVEENREKPTKNITPILPATGDIENV
LAFLGILILSVLSIFSLLKNKQSNKKV

M6_Spy0158: M6_Spy0158 is a reverse transcriptase. An example of Spy0158 is shown in the amino acid sequence SEQ ID NO 43.

SEQ ID NO: 43
MSLRHQNKKGIRKEGWKSRPQSRWSDHCQLVAQKSVLKQAISKTVLAERGLFSCLDDYLERHALKVN

M6_Spy0159: M6_Spy0159 is a collagen adhesion protein. It contains a sortase substrate motif LPXSG, shown in italics in the amino acid sequence SEQ ID NO: 44.

SEQ ID NO: 44
MYSRLKRELVIVINRKKKYKLIRLMVTVGLIFSQLVLPIRRLGLQMISTQTKVIPQEIVTQTETQGTQVVATKQK
LESENSSLKVALKRESGFEHNATIDASLDTESQGDNSQRSVTQAIVTMALELRKQGLSIVDTKIVRIQSSTNQRN
DITTTLTFKNGLSLEGASTEANDPNVRVGIVNPNDTVQTITPTIKQDADGKVKNLVFTGRLGKQVIIVSTTRLKE
EQTISLDSYGELVIDGAVGLSQKDRPPYSKPITVNILKPKLSSIESSLDSKDFEIVKTIDNLYTWDDQFYLLDFI
SKQYEVLKTDYQSAKDSTPQTRDILFGEYTVEPLVMNKGHNNTINIYIRSTRPLGLKPIGAAPALIQPRSFRSLT
PRSTRMKRSAPVEKFEGELEHHKRIDYLGDNQNNPDTTIDDKEDEHDTSDLYRLYLDMTGKKNPLDILVVVDKSG
SMQEGIGSVQRYRYYAQRWDDYYSQWVYHGTFDYSSYQGESFNRGQIHYRYRGIVSVSDGIRRDDAVKNSLLGVN
GLLQRFVNINPENKLSVIGFQGSADYHAGKWYPDQSPRGGFYQPNLNNSRDAELLKGWSTNSLLDPNTLTALHNN
GTNYHAALLKAKEILNEVKDDGRRKIMIFISDGVPTFYFGEDGYRSGNGSSNDRNNVTRSQEGSKLAIDEFKARY
PNLSIYSLGVSKDINSDTASSSPVVLKYLSGEEHYYGITDTAELEKTLNKIVEDSKLSQLGISDSLSQYVDYYDK
QPDVLVTRKSKVNDETEILYQKDQVQEAGKDIIDKVVFTPKTTSQPKGKVTLTFKSDYKVDDEYTYTLSFNVKAS
DEAYEKYKDNEGRYSEMGDSDTDYGTNQTSSGKGGLPSNSDASVNYMADGREQKLPYKHPVIQVKTVPITFTKVD
ADNNQKKLAGVEFELRKEDKKIVWEKGTTGSNGQLNFKYLQKGKTYYLYETKAKLGYTLPENPWEVAVANNGDIK
VKHPIEGELKSKDGSYMIKNYKIYQ*LPSSG*GRGSQIFIIVGSMTATVALLFYRRQHRKKQY

M6_Spy0159 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 181** LPSSG (shown in italics in SEQ ID NO: 44, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant M6_Spy0159 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in M6_Spy0159. The pilin motif sequence is underlined in SEQ ID NO: 44, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 265 and 276. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of M6_Spy0159 include at least one conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 44
MYSRLKRELVIVINRKKKYKLIRLMVTVGLIFSQLVLPIRRLGLQMISTQTKVIPQEIVTQTETQGTQVVATKQK
LESENSSLKVALKRESGFEHNATIDASLDTESQGDNSQRSVTQAIVTMALELRKQGLSIVDTKIVRIQSSTNQRN
DITTTLTFKNGLSLEGASTEANDPNVRVGIVNPNDTVQTITPTIKQDADGKVKNLVFTGRLGKQVIIVSTTRLKE
EQTISLDSYGELVIDGAVGLSQKDRPPYSKPITVNILKPKLSSIESSLDSKDFEIVKTIDNLYTWDDQFYLLDFI
SKQYEVLKTDYQSAKDSTPQTRDILFGEYTVEPLVMNKGHNNTINIYIRSTRPLGLKPIGAAPALIQPRSFRSLT
PRSTRMKRSAPVEKFEGELEHHKRIDYLGDNQNNPDTTIDDKEDEHDTSDLYRLYLDMTGKKNPLDILVVVDKSG

SMQEGIGSVQRYRYYAQRWDDYYSQWVYHGTFDYSSYQGESFNRGQIHYRYRGIVSVSDGIRRDDAVKNSLLGVN
GLLQRFVNINPENKLSVIGFQGSADYHAGKWYPDQSPRGGFYQPNLNNSRDAELLKGWSTNSLLDPNTLTALHNN
GTNYHAALLKAKEILNEVKDDGRRKIMIFISDGVPTFYFGEDGYRSGNGSSNDRNNVTRSQEGSKLAIDEFKARY
PNLSIYSLGVSKDINSDTASSSPVVLKYLSGEEHYYGITDTAELEKTLNKIVEDSKLSQLGISDSLSQYVDYYDK
QPDVLVTRKSKVNDETEILYQKDQVQEAGKDIIDKVVFTPKTTSQPKGKVTLTFKSDYKVDDEYTYTLSFNVKAS
DEAYEKYKDNEGRYSEMGDSDTDYGTNQTSSGKGGLPSNSDASVNYMADGREQKLPYKHPVIQVKTVPITFTKVD
ADNNQKKLAGVEFELRKEDKKIVWEKGTTGSNGQLNFKYLQKGKTYYLYETKAKLGYTLPENPWEVAVANNGDIK
VKHPIEGELKSKDGSYMIKNYKIYQLPSSGGRGSQIFIIVGSMTATVALLFYRRQHRKKQY

An E box containing a conserved glutamic residue has been identified in M6_Spy0159. The E-box motif is underlined in SEQ ID NO: 44, below. The conserved glutamic acid (E), at amino acid residue 950, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of M6_Spy0159. Preferred fragments of M6_Spy0159 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 44

MYSRLKRELVIVINRKKKYKLIRLMVTVGLIFSQLVLPIRRLGLQMISTQTKVIPQEIVTQTETQGTQVVATKQK
LESENSSLKVALKRESGFEHNATIDASLDTESQGDNSQRSVTQAIVTMALELRKQGLSIVDTKIVRIQSSTNQRN
DITTTLTFKNGLSLEGASTEANDPNVRVGIVNPNDTVQTITPTIKQDADGKVKNLVFTGRLGKQVIIVSTTRLKE
EQTISLDSYGELVIDGAVGLSQKDRPPYSKPITVNILKPKLSSIESSLDSKDFEIVKTIDNLYTWDDQFYLLDFI
SKQYEVLKTDYQSAKDSTPQTRDILFGEYTVEPLVMNKGHNNTINIYIRSTRPLGLKPIGAAPALIQPRSFRSLT
PRSTRMKRSAPVEKFEGELEHHKRIDYLGDNQNNPDTTIDDKEDEHDTSDLYRLYLDMTGKKNPLDILVVVDKSG
SMQEGIGSVQRYRYYAQRWDDYYSQWVYHGTFDYSSYQGESFNRGQIHYRYRGIVSVSDGIRRDDAVKNSLLGVN
GLLQRFVNINPENKLSVIGFQGSADYHAGKWYPDQSPRGGFYQPNLNNSRDAELLKGWSTNSLLDPNTLTALHNN
GTNYHAALLKAKEILNEVKDDGRRKIMIFISDGVPTFYFGEDGYRSGNGSSNDRNNVTRSQEGSKLAIDEFKARY
PNLSIYSLGVSKDINSDTASSSPVVLKYLSGEEHYYGITDTAELEKTLNKIVEDSKLSQLGISDSLSQYVDYYDK
QPDVLVTRKSKVNDETEILYQKDQVQEAGKDIIDKVVFTPKTTSQPKGKVTLTFKSDYKVDDEYTYTLSFNVKAS
DEAYEKYKDNEGRYSEMGDSDTDYGTNQTSSGKGGLPSNSDASVNYMADGREQKLPYKHPVIQVKTVPITFTKVD
ADNNQKKLAGVEFELRKEDKKIVWEKGTTGSNGQLNFKYLQKGKTYYLY**E**TKAKLGYTLPENPWEVAVANNGDIK
VKHPIEGELKSKDGSYMIKNYKIYQLPSSGGRGSQIFIIVGSMTATVALLFYRRQHRKKQY

M6_Spy0160: M6_Spy0160 is a fimbrial structural subunit. It contains a sortase substrate motif LPXTG (SEQ ID NO: 122), shown in italics in amino acid sequence SEQ ID NO: 45.

SEQ ID NO: 45

MTNRRETVREKILITAKKLMLACLAILAVVGLGMTRVSALSKDDTAQLKITNIEGGPTVTLYKIGEGVYNTNGDS
FINFKYAEGVSLTETGPTSQEITTIANGINTGKIKPFSTENVSISNGTATYNARGASVYIALLTGATDGRTYNPI
LLAASYNGEGNLVTKNIDSKSNYLYGQTSVAKSSLPSITKKVTGTIDDVNKKTTSLGSVLSYSLTFELPSYTKEA
VNKTVYVSDNMSEGLTFNFNSLTVEWKGKMANITEDGSVMVENTKIGIAKEVNNGFNLSFIYDSLESISPNISYK
AVVNNKAIVGEEGNPNKAEFFYSNNPTKGNTYDNLDKKPDKGNGITSKEDSKIVYTYQIAFRKVDSVSKTPLIGA
IFGVYDTSNKLIDIVTTNKNGYAISTQVSSGKYKIKELKAPKGYSLNTETYEITANWVTATVKTSANSKSTTYTS
DKNKATDNSEQVGWLKNGIFYSIDSRPTGNDVKEAYIESTKALTDGTTFSKSNEGSGTVLLETDIPNTKLGE*LPS*
*TG*SIGTYLFKAIGSAAMIGAIGIYIVKRRKA

M6_Spy0160 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 131** LPSTG (shown in italics in SEQ ID NO: 45, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant M6_Spy0160 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

PCT Published Text

An E box containing a conserved glutamic residue has been identified in M6_Spy0160. The E-box motif is underlined in SEQ ID NO: 45, below. The conserved glutamic acid (E), at amino acid residue 412, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of M6_Spy0160. Preferred fragments of M6_Spy0160 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 45

MTNRRETVREKILITAKKLMLACLAILAVVGLGMTRVSALSKDDTAQLKITNIEGGPTVTLYKIGEGVYNTNGDS
FINFKYAEGVSLTETGPTSQEITTIANGINTGKIKPFSTENVSISNGTATYNARGASVYIALLTGATDGRTYNPI
LLAASYNGEGNLVTKNIDSKSNYLYGQTSVAKSSLPSITKKVTGTIDDVNKKTTSLGSVLSYSLTFELPSYTKEA
VNKTVYVSDNMSEGLTFNFNSLTVEWKGKMANITEDGSVMVENTKIGIAKEVNNGFNLSFIYDSLESISPNISYK
AVVNNKAIVGEEGNPNKAEFFYSNNPTKGNTYDNLDKKPDKGNGITSKEDSKIVYTYQIAFRKVDSVSKTPLIGA
IFGVYDTSNKLIDIVTTNKNGYAISTQVSSGKYKIKELKAPKGYSLNTETYEITANWVTATVKTSANSKSTTYTS
DKNKATDNSEQVGWLKNGIFYSIDSRPTGNDVKEAYIESTKALTDGTTFSKSNEGSGTVLLETDIPNTKLGELPS
TGSIGTYLFKAIGSAAMIGAIGIYIVKRRKA

M6_Spy0161 is a srtB type sortase. An example of an amino acid sequence of M6_Spy-161 is shown in SEQ ID NO: 46.

SEQ ID NO: 46

MTERLKNLGILLLFLLGTAIFLYPTLSSQWNAYRDRQLLSTYHKQVIQKKPSEMEEVWQKAKAYNARLGIQPVPD
AFSFRDGIHDKNYESLLQIENNDIMGYVEVPSIKVTLPIYHYTTDEVLTKGAGHLFGSALPVGGDGTHTVISAHR
GLPSAEMFTNLNLVKKGDTFYFRVLNKVLAYKVDQILIVEPDQATSLSGVMGKDYATLVTCTPYGVNTKRLLVRG
HRIAYHYKKYQQAKKAMKLVDKSRMWAEVVCAAFGVVIAIILVFMYSRVSAKKSK

As discussed above, applicants have also determined the nucleotide and encoded amino acid sequence of fimbrial structural subunits in several other GAS AI-1 strains of bacteria. Examples of sequences of these fimbrial structural subunits are set forth below.

M6 strain isolate CDC SS 410 is a GAS AI-1 strain of bacteria. CDC SS 410_fimbrial is thought to be a fimbrial structural subunit of M6 strain isolate CDC SS 410. An example of a nucleotide sequence encoding the CDC SS 410_fimbrial protein (SEQ ID NO: 267) and a CDC SS 410_fimbrial protein amino acid sequence (SEQ ID NO: 268) are set forth below.

SEQ ID NO: 267

aaagatgatactgcacaactaaagataacaaatattgaaggtgggccaacagtaacactt
tataaaataggagaaggtgtttacaacactaatggtgattcttttattaactttaaatat
gctgaggggtttctttaactgaaacaggacctacatcacaagaaattactactattgca
aatggtattaatacgggtaaaataaagccttttagtactgaaaacgttagtatttctaat
ggaacagcaacttataatgcgagaggtgcatctgtttatattgcattattaacaggtgcg
acagatggccgtacctacaatcctatttttattagctgcatcttataatggtgagggaaat
ttagttactaaaaatattgattccaaatctaattatttatatggacaaacaagtgttgca
aaatcatcattaccatctattacaaagaaagtaaccgggacaatagatgacgtgaataaa
aagactacctcgttaggaagtgtattgtcttattcgctgacatttgaattaccaagttat
accaaagaagcagtcaataaaacagtatatgtttctgataatatgtcggaaggtcttact
tttaactttaatagtcttacagtagaatggaaaggtaagatggctaatattactgaagat
ggttcagtaatggtagaaaatacaaaaatcggaatagctaaggaggttaataacggtttt
aatttaagttttatttatgatagtttagaatctatatcaccaaatataagttataaagct
gttgtaaacaataaagctattgttggtgaagagggtaatcctaataaagctgaattcttc
tattcaaataatccaacaaaaggtaatacatacgataatttagataagaagcctgataaa
gggaatggtattacatccaaagaagattctaaaattgtttatacttatcaaatagcgttt
agaaaagttgatagtgttagtaagaccccacttattggtgcaatttttggagtttatgat
actagtaataaattaattgatattgttacaaccaataaaaatggatatgctatttcaaca

caagtatcttcaggaaaatataaaattaaggaattaaaagctcctaaaggttattcattg
aatacagaaacttatgaaattacggcaaattgggtaactgctacagtcaagacaagtgct
aattcaaaaagtactacttatacatctgataaaaataaggcgacagataattcagagcaa
gtaggatggttaaaaaatggtatattctattctatagatagtagacctacaggaaatgat
gttaaagaggcttatattgaatctactaaggctttaactgatggaacaactttctcaaaa
tcgaatgaaggttcaggtacagtattattagaaactgacatccctaacaccaaagctaggt
gaactc

SEQ ID NO: 268

KDDTAQLKITNIEGGPTVTLYKIGEGVYNTNGDSFINFKYAEGV
SLTETGPTSQEITTIANGINTGKIKPFSTENVSISNGTATYNARGASVYIALLTGATD
GRTYNPILLAASYNGEGNLVTKNIDSKSNYLYGQTSVAKSSLPSITKKVTGTIDDVNK
KTTSLGSVLSYSLTFELPSYTKEAVNKTVYVSDNMSEGLTFNFNSLTVEWKGKMANIT
EDGSVMVENTKIGIAKEVNNGFNLSFIYDSLESISPNISYKAVVNNKAIVGEEGNPNK
AEFFYSNNPTKGNTYDNLDKKPDKGNGITSKEDSKIVYTYQIAFRKVDSVSKTPLIGA
IFGVYDTSNKLIDIVTTNKNGYAISTQVSSGKYKIKELKAPKGYSLNTETYEITANWV
TATVKTSANSKSTTYTSDKNKATDNSEQVGWLKNGIFYSIDSRPTGNDVKEAYIESTK
ALTDGTTFSKSNEGSGTVLLETDIPNTKLGEL

M6 strain isolate ISS 3650 is a GAS AI-1 strain of bacteria. ISS3650_fimbrial is thought to be a fimbrial structural subunit of M6 strain isolate ISS 3650. An example of a nucleotide sequence encoding the ISS3650_fimbrial protein (SEQ ID NO: 269) and an ISS3650_fimbrial protein amino acid sequence (SEQ ID NO: 270) are set forth below.

SEQ ID NO: 269

gaatggaaaggtaagatggctaatattactgaagatggttcagtaatggtagaaaataca
aaaatcggaatagctaaggaggttaataacggttttaatttaagttttatttatgatagt
ttagaatctatatcaccaaatataagttataaagctgttgtaaacaataaagctattgtt
ggtgaagagggtaatcctaataaagctgaattcttctattcaaataatccaacaaaaggt
aatacatacgataatttagataagaagcctgataaagggaatggtattacatccaaagaa
gattctaaaattgtttatacttatcaaatagcgtttagaaaagttgatagtgttagtaag
accccacttattggtgcaatttttggagtttatgatactagtaataaattaattgatatt
gttacaaccaataaaaatggatatgctatttcaacacaagtatcttcaggaaaatataaa
attaaggaattaaaagctcctaaaggttattcattgaatacagaaacttatgaaattacg
gcaaattgggtaactgctacagtcaagacaagtgctaattcaaaaagtactacttataca
tctgataaaaataaggcgacagataattcagagcaagtaggatggttaaaaaatggtata
ttctattctatagatagtagacctacaggaaatgatgttaaagaggcttatattgaatct
actaaggctttaactgatggaacaactttctcaaaatcgaatgaaggttcaggtacagta
ttattagaaactgacatcc

SEQ ID NO: 270

EWKGKMANITEDGSVMVENTKIGIAKEVNNGFNLSFIYDSLESI
SPNISYKAVVNNKAIVGEEGNPNKAEFFYSNNPTKGNTYDNLDKKPDKGNGITSKEDS
KIVYTYQIAFRKVDSVSKTPLIGAIFGVYDTSNKLIDIVTTNKNGYAISTQVSSGKYK
IKELKAPKGYSLNTETYEITANWVTATVKTSANSKSTTYTSDKNKATDNSEQVGWLKN
GIFYSIDSRPTGNDVKEAYIESTKALTDGTTFSKSNEGSGTVLLETDI

M23 strain isolate DSM2071 is a GAS AI-1 strain of bacteria. DSM2071_fimbrial is thought to be a fimbrial structural subunit of M23 strain DSM2071. An example of a nucleotide sequence encoding the DSM2071_fimbrial protein (SEQ ID NO: 251) and a DSM2071_fimbrial protein amino acid sequence (SEQ ID NO: 252) are set forth below.

SEQ ID NO: 251

atgagagagaaaatattaatagcagcaaaaaaactaatgctagcttgtttagctatctta
gctgtagtagggcttggaatgacaagagtatcagctttatcaaaagatgataaggcggag
ttgaagataacaaatatcgaaggtaaaccgaccgtgacactgtataaaattggtgatgga
aaatacagtgagcgaggggattcttttattggatttgagttaaagcaaggtgtggagcta
ataaggcaaaacctacatctcaagaaataaataaaatcgctaatggtattaataaaggt
agtgttaaggctgaagtagttaatataaaagaacatgctagtacaacttatagttataca

aaaactggtgcaggtatttacttggctaattgactggagctactgatggacgtgcctat
aatcctatcttactgacagcttcttacaatgaggaaaatccacttaagggagggcagatt
gacgcaactagtcattatctttttggagaagaagcagttgctaaatctagccaaccaaca
attagcaagtcaattacaaaatccacaaaagatggtgataaagatacagcatctgtaggt
gaaaaagttgattacaaattaactgttcagttaccaagttattcgaaagatgctatcaat
aaaacggtgtttatcactgacaaattgtctcagggacttactttccttccaaaaagttta
aagattatctggaatggtcaaacgttaacaaaggtgaatgaagaatttaaagctggagat
aaggtaattgctcaacttaaggttgaaaataatggatttaatctgaactttaattatgat
aaccttgataatcatgccccagaagttaactatagtgctctactaaatgaaaacgcagtt
gttggtaaaggtggtaatgacaataatgtagactattactattcaaataatccgaataaa
ggagagacccataaaacaactgagaagcctaaagagggtgaaggtactggtatcactaaa
aagacggataaaaaaaccgtctacacctatcgtgtagcctttaagaaaacaggcaaagat
catgccccactagctggtgctgttttcggtatctattcagataaggaagcgaaacaatta
gtcgatattgttgtgacaaatgcacagggttatgcagcatcaagcgaagttgggaaaggg
acttattacattaaagaaattaaatcccctaagggttactctttaaatacaaatatttat
gaagtggaaacttcatgggaaaaagctacaacgacttctacaactaatcgtttagagaca
atttatacaacagatgataatcaaaagtctccaggaactaatacagttggttggttggaa
gatggtgtcttttacaaagaaaatccaggtggtgatgctaaacttgcctatatcaaacaa
tcaacagaggagacttctacaactatagaagtcaaagaaaatcaagctgaaggttcaggt
acggtattattagaaactgaaattcctaacaccaaattaggtgaattaccttcgacaggt
agcattggtacttacctctttaaagctattggttcggctgctatgatcggtgcaattggt
atttatattgttaaacgtcgtaaagcttaa

SEQ ID NO: 252

MREKILIAAKKLMLACLAILAVVGLGMTRVSALSKDDKAELKIT
NIEGKPTVTLYKIGDGKYSERGDSFIGFELKQGVELNKAKPTSQEINKIANGINKGSV
KAEVVNIKEHASTTYSYTTTGAGIYLAILTGATDGRAYNPILLTASYNEENPLKGGQI
DATSHYLFGEEAVAKSSQPTISKSITKSTKDGDKDTASVGEKVDYKLTVQLPSYSKDA
INKTVFITDKLSQGLTFLPKSLKIIWNGQTLTKVNEEFKAGDKVIAQLKVENNGFNLN
FNYDNLDNHAPEVNYSALLNENAVVGKGGNDNNVDYYYSNNPNKGETHKTTEKPKEGE
GTGITKKTDKKTVYTYRVAFKKTGKDHAPLAGAVFGIYSDKEAKQLVDIVVTNAQGYA
ASSEVGKGTYYIKEIKSPKGYSLNTNIYEVETSWEKATTTSTTNRLETIYTTDDNQKS
PGTNTVGWLEDGVFYKENPGGDAKLAYIKQSTEETSTTIEVKENQAEGSGTVLLETEI
PNTKLGELPSTGSIGTYLFKAIGSAAMIGAIGIYIVKRRKA

GAS AI-2 sequences

As discussed above, a GAS AI-2 sequence is present in an M1 strain isolate (SF370). Examples of GAS AI-2 sequences from M1 strain isolate SF370 are set forth below.

Spy0124 is a rofA transcriptional regulator. An example of an amino acid sequence for Spy0124 is set forth in SEQ ID NO:47.

SEQ ID NO: 47

MIEKYLESSIESKCQLIVLFFKTSYLPITEVAEKTGLTFLQLNHYCEELNAFFPGSLSMTIQKRMISCQFTHPFK
ETYLYQLYASSNVLQLLAFLIKNGSHSRPLTDFARSHFLSNSSAYRMREALIPLLRNFELKLSKNKIVGEEYRIR
YLIALLYSKFGIKVYDLTQQDKNTIHSFLSHSSTHLKTSPWLSESFSFYDILLALSWKRHQFSVTIPQTRIFQQL
KKLFVYDSLKKSSHDIIETYCQLNFSAGDLDYLYLIYITANNSFASLQWTPEHIRQYCQLFEENDTFRLLLNPII
TLLPNLKEQKASLVKALMFFSKSFLFNLQHFIPETNLFVSPYYKGNQKLYTSLKLIVEEWMAKLPGKRDLNHKHF
HLFCHYVEQSLRNIQPPLVVVFVASNFINAHLLTDSFPRYFSDKSIDFHSYYLLQDNVYQIPDLKPDLVITHSQL
IPFVHHELTKGIAVAEISFDESILSIQELMYQVKEEKFQADLTKQLT

GAS 015 is also referred to as Cpa. It contains a sortase substrate motif VVXTG (SEQ ID NO: 135), shown in italics in SEQ ID NO: 48.

SEQ ID NO: 48

LRGEKMKKTRFPNKLNTLNTQRVLSKNSKRFTVTLVGVFLMIFALVTSMVGAKTVFGLVESSTPNAINPDSSSEY
RWYGYESYVRGHPYYKQFRVAHDLRVNLEGSRSYQVYCFNLKKAFPLGSDSSVKKWYKKHDGISTKFEDYAMSPR
ITGDELNQKLRAVMYNGHPQNANGIMEGLEPLNAIRVTQEAVWYYSDNAPISNPDESFKRESESNLVSTSQLSLM
RQALKQLIDPNLATKMPKQVPDDFQLSIFESEDKGDKYNKGYQNLLSGGLVPTKPPTPGDPPMPPNQPQTTSVLI
RKYAIGDYSKLLEGATLQLTGDNVNSFQARVFSSNDIGERIELSDGTYTLTELNSPAGYSIAEPITFKVEAGKVY

TIIDGKQIENPNKEIVEPYSVEAYNDFEEFSVLTTQNYAKFYYAKNKNGSSQVVYCFNADLKSPPDSEDGGKTMT
PDFTTGEVKYTHIAGRDLFKYTVKPRDTDPDTFLKHIKKVIEKGYREKGQAIEYSGLTETQLRAATQLAIYYFTD
SAELDKDKLKDYHGFGDMNDSTLAVAKILVEYAQDSNPPQLTDLDFFIPNNNKYQSLIGTQWHPEDLVDIIRMED
KKEVIPVTHNLTLRKTVTGLAGDRTKDFHFEIELKNNKQELLSQTVKTDKTNLEFKDGKATINLKHGESLTLQGL
PEGYSYLVKETDSEGYKVKVNSQEVANATVSKTGITSDETLAFENNKEP*VVPTG*VDQKINGYLALIVIAGISLGI
WGIHTIRIRKHD

GAS 015 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 182** VVPTG (shown in italics in SEQ ID NO: 48, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant GAS 015 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in GAS 015. The pilin motif sequence is underlined in SEQ ID NO: 48, below. Conserved lysine (K) residues are also marked in bold, at amino acid residue 243. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of GAS 015 include the conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 48
LRGEKMKKTRFPNKLNTLNTQRVLSKNSKRFTVTLVGVFLMIFALVTSMVGAKTVFGLVESSTPNAINPDSSSEY
RWYGYESYVRGHPYYKQFRVAHDLRVNLEGSRSYQVYCFNLKKAFPLGSDSSVKKWYKKHDGISTKFEDYAMSPR
ITGDELNQKLRAVMYNGHPQNANGIMEGLEPLNAIRVTQEAVWYYSDNAPISNPDESFKRESESNLVSTSQLSLM
RQALKQLIDPNLATKMPKQVPDDFQLSIFESEDKGDKYNKGYQNLLSGGLVPTKPPTPGDPPMPPNQPQTTSVLI
RKYAIGDYSKLLEGATLQLTGDNVNSFQARVFSSNDIGERIELSDGTYTLTELNSPAGYSIAEPITFKVEAGKVY
TIIDGKQIENPNKEIVEPYSVEAYNDFEEFSVLTTQNYAKFYYAKNKNGSSQVVYCFNADLKSPPDSEDGGKTMT
PDFTTGEVKYTHIAGRDLFKYTVKPRDTDPDTFLKHIKKVIEKGYREKGQAIEYSGLTETQLRAATQLAIYYFTD
SAELDKDKLKDYHGFGDMNDSTLAVAKILVEYAQDSNPPQLTDLDFFIPNNNKYQSLIGTQWHPEDLVDIIRMED
KKEVIPVTHNLTLRKTVTGLAGDRTKDFHFEIELKNNKQELLSQTVKTDKTNLEFKDGKATINLKHGESLTLQGL
PEGYSYLVKETDSEGYKVKVNSQEVANATVSKTGITSDETLAFENNKEPVVPTGVDQKINGYLALIVIAGISLGI
WGIHTIRIRKHD

An E box containing a conserved glutamic residue has been identified in GAS 015. The E-box motif is underlined in SEQ ID NO: 48, below. The conserved glutamic acid (E), at amino acid residue 352, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of GAS 015. Preferred fragments of GAS 015 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 48
LRGEKMKKTRFPNKLNTLNTQRVLSKNSKRFTVTLVGVFLMIFALVTSMVGAKTVFGLVESSTPNAINPDSSSEY
RWYGYESYVRGHPYYKQFRVAHDLRVNLEGSRSYQVYCFNLKKAFPLGSDSSVKKWYKKHDGISTKFEDYAMSPR
ITGDELNQKLRAVMYNGHPQNANGIMEGLEPLNAIRVTQEAVWYYSDNAPISNPDESFKRESESNLVSTSQLSLM
RQALKQLIDPNLATKMPKQVPDDFQLSIFESEDKGDKYNKGYQNLLSGGLVPTKPPTPGDPPMPPNQPQTTSVLI
RKYAIGDYSKLLEGATLQLTGDNVNSFQARVFSSNDIGERIELSDGTYTLT**E**LNSPAGYSIAEPITFKVEAGKVY
TIIDGKQIENPNKEIVEPYSVEAYNDFEEFSVLTTQNYAKFYYAKNKNGSSQVVYCFNADLKSPPDSEDGGKTMT
PDFTTGEVKYTHIAGRDLFKYTVKPRDTDPDTFLKHIKKVIEKGYREKGQAIEYSGLTETQLRAATQLAIYYFTD
SAELDKDKLKDYHGFGDMNDSTLAVAKILVEYAQDSNPPQLTDLDFFIPNNNKYQSLIGTQWHPEDLVDIIRMED
KKEVIPVTHNLTLRKTVTGLAGDRTKDFHFEIELKNNKQELLSQTVKTDKTNLEFKDGKATINLKHGESLTLQGL

PIGYSYIVKETDSEGYKVKVNGQEVANATVSKTGITSDETLAFENNKEPVVPTGVDQKINGYLALIVIAGISLGI
WGIHTIRIRKHD

Spy0127 is a LepA putative signal peptidase. An example of an amino acid sequence for Spy0127 is set forth in SEQ ID NO: 49.

SEQ ID NO: 49
MIIKRNDMAPSVKAGDAILFYRLSQTYKVEEAVVYEDSKTSITKVGRIIAQAGDEVDLTEQGELKINGHIQNEGL
TFIKSREANYPYRIADNSYLILNDYYSQESENYLQDAIAKDAIKGTINTLIRLRNH

Spy0128 is thought to be a fibrial protein. It contains a sortase substrate motif EVXTG (SEQ ID NO: 136) shown in italics in SEQ ID NO: 50.

SEQ ID NO: 50
MKLRHLLLTGAALTSFAATTVHGETVVNGAKLTVTKNLDLVNSNALIPNTDFTFKIEPDTTVNEDGNKFKGVALN
TPMTKVTYTNSDKGGSNTKTAEFDFSEVTFEKPGVYYYKVTEEKIDKVPGVSYDTTSYTVQVHVLWNEEQQKPVA
TYIVGYKEGSKVPIQFKNSLDSTTLTVKKKVSGTGGDRSKDFNFGLTLKANQYYKASEKVMIEKTTKGGQAPVQT
EASIDQLYHFTLKDGESIKVTNLPVGVDYVVTEDDYKSEKYTTNVEVSPQDGAVKNIAGNSTEQETSTDKDMTIT
FTNKKDF*EVPTG*VAMTVAPYIALGIVAVGGALYFVKKKNA

Spy0128 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 183** EVPTG (shown in italics in SEQ ID NO: 50, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant Spy0128 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Two E boxes containing a conserved glutamic residue have been identified in Spy0128. The E-box motifs are underlined in SEQ ID NO: 50, below. The conserved glutamic acid (E) residues, at amino acid residues 271 and 290, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of Spy0128. Preferred fragments of Spy0128 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 50
MKLRHLLLTGAALTSFAATTVHGETVVNGAKLTVTKNLDLVNSNALIPNTDFTFKIEPDTTVNEDGNKFKGVALN
TPMTKVTYTNSDKGGSNTKTAEFDFSEVTFEKPGVYYYKVTEEKIDKVPGVSYDTTSYTVQVHVLWNEEQQKPVA
TYIVGYKEGSKVPIQFKNSLDSTTLTVKKKVSGTGGDRSKDFNFGLTLKANQYYKASEKVMIEKTTKGGQAPVQT
EASIDQLYHFTLKDGESIKVTNLPVGVDYVVTEDDYKSEKYTTNV**E**VSPQDGAVKNIAGNSTEQ**E**TSTDKDMTIT
FTNKKDFEVPTGVAMTVAPYIALGIVAVGGALYFVKKKNA

Spy0129 is a srtC1 type sortase. An example of an amino acid sequence for Spy0129 is set forth in SEQ ID NO: 51.

SEQ ID NO: 51
MIVRLIKLLDKLINVIVLCFFFLCLLIAALGIYDALTVYQGANATNYQQYKKKGVQFDDLLAINSDVMAWLTVKG
THIDYPIVQGENNLEYINKSVEGEYSLSGSVFLDYRNKVTFEDKYSLIYAHHMAGNVMFGELPNFRKKSFFNKHK
EFSIETKTKQKLKINIFACIQTDAFDSLLFNPIDVDISSKNEFLNHIKQKSVQYREILTTNESRFVALSTCEDMT
TDGRIIVIGQIE"

Spy0130 is referred to as a hypothetical protein. It contains a sortase substrate motif LPXTG (SEQ ID NO: 122), shown in italics in SEQ ID NO: 52.

SEQ ID NO:52

MKKSILRILAIGYLLMSFCLLDSVEAENLTASINIEVINQVDVATNKQSSDIDETFMFVIEALDKESPLPNSVTT
SVKGNGKTSFEQLTFSEVGQYHYKIHQLLGKNSQYHYDETVYEVVIYVLYNEQSGALETNLVSNKLGETEKSELI
FKQEYSEKTPEPHQPDTTEKEKPQKKRNGI*LPSTG*EMVSYVSALGIVLVATITLYSIYKKLKTSK

Spy0130 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 131** LPSTG (shown in italics in SEQ ID NO: 52, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant Spy0130 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Two E boxes containing conserved glutamic residues have been identified in Spy0130. The E-box motifs are underlined in SEQ ID NO: 52, below. The conserved glutamic acid (E) residues, at amino acid residues 118 and 148, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of Spy0130. Preferred fragments of Spy0130 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 52

MKKSILRILAIGYLLMSFCLLDSVEAENLTASINIEVINQVDVATNKQSSDIDETFMFVIEALDKESPLPNSVTT
SVKGNGKTSFEQLTFSEVGQYHYKIHQLLGKNSQYHYDETVY**E**VVIYVLYNEQSGALETNLVSNKLGETEKS**E**LI
FKQEYSEKTPEPHQPDTTEKEKPQKKRNGILPSTGEMVSYVSALGIVLVATITLYSIYKKLKTSK

Spy0131 is referred to as a conserved hypothetical protein. An example of an amino acid sequence of Spy0131 is set forth in SEQ ID NO: 53

SEQ ID NO: 53

MTRTNYQKKRMTCPVETEDITYRRKKIKGRRQAILAQFEPELVHHELIGDSCTCPDCHGTLTEIGSVVQRQELVF
IPAQLKRINHVQHAYKCQTCSDNSLSDKIIKAPVPKAPLAHSLGSASIIAHTVHQKFTLKVPNYRQEEDWNKLGL
SISRKEIANWHIKSSQYYFEPLYDLLRDILLSQEVIHADETSYRVLESDTQLTYYWTFLSGKHEKKGITLYHHDK
RRSGLVTQEVLGDYSGYVHCDMHGAYRQLEHAKLVGCWAHVRRKFFEATPKQADKTSLGRKGLVYCDKLFALEAE
WCELPPQERLVKRKEILTPLMTTFFDWCREQVVLSGSKLGLAIAYSLKHERTFRTVLEDGHIVLSNNMAERAIKS
LVMGRKNWLFSQSFEGAKAAAIIMSLLETAKRHGLNSEKYISYLLDRLPNEETLAKREVLEAYLPWAKKVQTNCQ

Spy0133 is referred to as a conserved hypothetical protein. An example of an amino acid sequence of Spy0133 is set forth in SEQ ID NO: 54.

SEQ ID NO: 54

MTIRLNDLGQVYLVCGKTDMRQGIDSLAYLVKSQHELDLFSGAVYLFCGGRRDRFKALYWDGQGFWLLYKRFENG
KLAWPRNRDEVKCLTAVQVDWLMKGFFISPNIKISKSHDFY

Spy0135 is a SrtB type sortase. It is also referred to as a putative fibria-associated protein. An example of an amino acid sequence of Spy0135 is set forth in SEQ ID NO: 55.

SEQ ID NO: 55

MECYRDRQLLSTYHKQVTQKKPSEMEEVWQKAKAYNARLGIQPVPDAFSFRDGIHDKNYESLLQIENNDIMGYVE
VPSIKVTLPIYHYTTDEVLTKGAGHLFGSALPVGGDGTHTVISAHRGLPSAEMFTNLNLVKKGDTFYFRVLNKVL
AYKVDQILTVEPDQVTSLSGVMGKDYATLVTCTPYGVNTKRLLVRGHRIAYHYKKYQQAKKAMKLVDKSRMWAEV
VCAAFGVVIAIILVFMYSRVSAKKSK

GAS AI-3 sequences

PC As discussed above, a GAS AI-3 sequence is present in a M3, M18 and M5 strain isolates. Examples of GAS AI-3 sequences from M3 strain isolate MGAS315 are set forth below.

SpyM30097 is as a negative transcriptional regulator (Nra). An example of an amino acid sequence of SpyM30097 is set forth in SEQ ID NO: 56.

SEQ ID NO: 56

MPYVKKKKDSFLVETYLEQSIRDKSELVLLLFKSPTIIFSHVAKQTGLTAVQLKYYCKELDDFFGNNLDITIKKG
KIICCFVKPVKEFYLHQLYDTSTILKLLVFFIKNGTSSQPLIKFSKKYFLSSSSAYRLRESLIKLLREFGLRVSK
NTIVGEEYRIRYLIAMLYSKFGIVIYPLDHLDNQIIYRFLSQSATNLRTSPWLEEPFSFYNMLLALSWKRHQFAV
SIPQTRIFRQLKKLFIYDCLTRSSRQVIENAFSLTFSQGDLEYLFLIYITTNNSFASLQWTPQHIETCCHIFEKN
DTFRLLLEPILKRLPQLNHSKQDLIKALMYFSKSFLFNLQHFVIEIPSFSLPTYTGNSNLYKALKNIVNQWLAQL
PGKRHLNEKHLQLFCSHIEQILKNKQPALTVVLISSNFINAKLLTDTIPRYFSDKGIHFYSFYLLRDDIYQIPSL
KPDLVITHSRLIPFVKNDLVKGVTVAEFSFDNPDYSIASIQNLIYQLKDKKYQDFLNEQLQ

SpyM30098 is thought to be a collagen binding protein (Cpb). It contains a sortase substrate motif VPXTG (SEQ ID NO: 137) shown in italics in SEQ ID NO: 57.

SEQ ID NO: 57

MQKRDKTNYGSANNKRRQTTIGLLKVFLTFVALIGIVGFSIRAFGAEEQSVPNKQSSVQDYPWYGYDSYSKGYPD
YSPLKTYHNLKVNLDGSKEYQAYCFNLTKHFPSKSDSVRSQWYKKLEGTNENFIKLADKPRIEDGQLQQNILRIL
YNGYPNDRNGIMKGIDPLNAILVTQNAIWYYTDSSYISDTSKAFQQEETDLKLDSQQLQLMRNALKRLINPKEVE
SLPNQVPANYQLSIFQSSDKTFQNLLSAEYVPDTPPKPGEEPPAKTEKTSVIIRKYAEGDYSKLLEGATLKLAQI
EGSGFQEKIFDSNKSGEKVELPNGTYVLSELKPPQGYGVATPITFKVAAEKVLIKNKEGQFVENQNKEIAEPYSV
TAFNDFEEIGYLSDFNNYGKFYYAKNTNGTNQVVYCFNADLHSPPDSYDHGANIDPDVSESKEIKYTHVSGYDLY
KYAATPRDKDADFFLKHIKKILDKGYKKKGDTYKTLTEAQFRAATQLAIYYYTDSADLTTLKTYNDNKGYHGFDK
LDDATLAVVHELITYAEDVTLPMTQNLDFFVPNSSRYQALIGTQYHPNELIDVISMEDKQAPIIPITHKLTISKT
VTGTIADKKKEFNFEIHLKSSDGQAISGTYPTNSGELTVTDGKATFTLKDGESLIVEGLPSGYSYEITETGASDY
EVSVNGKNAPDGKATKASVKEDETVAFENRKDL*VPPTG*LTTDGAIYLWLLLLVPFGLLVWLFGRKGTKK

SpyM30098 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 184** VPPTG (shown in italics in SEQ ID NO: 57, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyM30098 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in SpyM30098. The pilin motif sequence is underlined in SEQ ID NO: 57, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 262 and 270. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyM30098 include at least one conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 57

MQKRDKTNYGSANNKRRQTTIGLLKVFLTFVALIGIVGFSIRAFGAEEQSVPNKQSSVQDYPWYGYDSYSKGYPD
YSPLKTYHNLKVNLDGSKEYQAYCFNLTKHFPSKSDSVRSQWYKKLEGTNENFIKLADKPRIEDGQLQQNILRIL
YNGYPNDRNGIMKGIDPLNAILVTQNAIWYYTDSSYISDTSKAFQQEETDLKLDSQQLQLMRNALKRLINPKEVE
SLPNQVPANYQLSIFQSSDKTFQNLLSAEYVPDTPP**K**PGEEPPA**K**TEKTSVIIRKYAEGDYSKLLEGATLKLAQI
EGSGFQEKIFDSNKSGEKVELPNGTYVLSELKPPQGYGVATPITFKVAAEKVLIKNKEGQFVENQNKEIAEPYSV
TAFNDFEEIGYLSDFNNYGKFYYAKNTNGTNQVVYCFNADLHSPPDSYDHGANIDPDVSESKEIKYTHVSGYDLY
KYAATPRDKDADFFLKHIKKILDKGYKKKGDTYKTLTEAQFRAATQLAIYYYTDSADLTTLKTYNDNKGYHGFDK
LDDATLAVVHELITYAEDVTLPMTQNLDFFVPNSSRYQALIGTQYHPNELIDVISMEDKQAPIIPITHKLTISKT

VTGTIADKKKEFNFEIHLKSSDGQAISGTYPTNSGELTVTDGKATFTLKDGESLIVEGLPSGYSYEITETGASDY
EVSVNGKNAPDGKATKASVKEDETVAFENRKDLVPPTGLTTDGAIYLWLLLLVPFGLLVWLFGRKGTKK

An E box containing a conserved glutamic residue has been identified in SpyM30098. The E-box motif is underlined in SEQ ID NO: 57, below. The conserved glutamic acid (E), at amino acid residue 330, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of SpyM30098. Preferred fragments of SpyM30098 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 57

MQKRDKTNYGSANNKRRQTTIGLLKVFLTFVALIGIVGFSIRAFGAEEQSVPNKQSSVQDYPWYGYDSYSKGYPD
YSPLKTYHNLKVNLDGSKEYQAYCFNLTKHFPSKSDSVRSQWYKKLEGTNENFIKLADKPRIEDGQLQQNILRIL
YNGYPNDRNGIMKGIDPLNAILVTQNAIWYYTDSSYISDTSKAFQQEETDLKLDSQQLQLMRNALKRLINPKEVE
SLPNQVPANYQLSIFQSSDKTFQNLLSAEYVPDTPPKPGEEPPAKTEKTSVIIRKYAEGDYSKLLEGATLKLAQI
EGSGFQEKIFDSNKSGEKVELPNGTYVLS**E**LKPPQGYGVATPITFKVAAEKVLIKNKEGQFVENQNKEIAEPYSV
TAFNDFEEIGYLSDFNNYGKFYYAKNTNGTNQVVYCFNADLHSPPDSYDHGANIDPDVSESKEIKYTHVSGYDLY
KYAATPRDKDADFFLKHIKKILDKGYKKKGDTYKTLTEAQFRAATQLAIYYYTDSADLTTLKTYNDNKGYHGFDK
LDDATLAVVHELITYAEDVTLPMTQNLDFFVPNSSRYQALIGTQYHPNELIDVISMEDKQAPIIPITHKLTISKT
VTGTIADKKKEFNFEIHLKSSDGQAISGTYPTNSGELTVTDGKATFTLKDGESLIVEGLPSGYSYEITETGASDY
EVSVNGKNAPDGKATKASVKEDETVAFENRKDLVPPTGLTTDGAIYLWLLLLVPFGLLVWLFGRKGTKK

SpyM30099 is referred to as LepA. An example of an amino acid sequence of SpyM30099 is set forth in SEQ ID NO: 58.

SEQ ID NO: 58

MTNYLNRLNENPLLKAFIRLVLKISIIGFLGYILFQYVFGVMIVNTNQMSPAVSAGDGVLYYRLTDRYHINDVVV
YEVDDTLKVGRIAAQAGDEVNFTQEGGLLINGHPPEKEVPYLTYPHSSGPNFPYKVPTGTYFILNDYREERLDSR
YYGALPINQIKGKISTLLRVRGI

SpyM30100 is thought to be a fimbrial protein. An example of an amino acid sequence of SpyM30100 is set forth in SEQ ID NO: 59.

SEQ ID NO: 59

MKKNKLLLATAILATALGTASLNQNVKAETAGVSENAKLIVKKTFDSYTDNEVLMPKADYTFKVEADSTASGKTK
DGLEIKPGIVNGLTEQIISYTNTDKPDSKVKSTEFDFSKVVFPGIGVYRYTVSEKQGDVEGITYDTKKWTVDVYV
GNKEGGGFEPKFIVSKEQGTDVKKPVNFNNSFATTSLKVKKNVSGNTGELQKEFDFTLTLNESTNFKKDQIVSLQ
KGNEKFEVKIGTPYKFKLKNGESIQLDKLPVGITYKVNEMEANKDGYKTTASLKEGDGQSKMYQLDMEQKTDESA
DEIVVTNKRDT*QVPTG*VVGTLAPFAVLSIVAIGGVIYITKRKKA

SpyM30100 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 140** QVPTG (shown in italics in SEQ ID NO: 59, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyM30100 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Two pilin motifs, discussed above, containing conserved lysine (K) residues have also been identified in SpyM30100. The pilin motif sequences are underlined in SEQ ID NO: 59, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 57 and 63 and at amino acid residues 161 and 166. The pilin sequences, in particular the conserved lysine residues, are

SUBSTITUTE SHEET (RULE 26)

thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyM30100 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 59

MKKNKLLLATAILATALGTASLNQNVKAETAGVSENAKLIVKKTFDSYTDNEVLMPKADYTFKVEADSTASGKTK
DGLEIKPGIVNGLTEQIISYTNTDKPDSKVKSTEFDFSKVVFPGIGVYRYTVSEKQGDVEGITYDTKKWTVDVYV
GNKEGGGFEPKFIVSKEQGTDVKKPVNFNNSFATTSLKVKKNVSGNTGELQKEFDFTLTLNESTNFKKDQIVSLQ
KGNEKFEVKIGTPYKFKLKNGESIQLDKLPVGITYKVNEMEANKDGYKTTASLKEGDGQSKMYQLDMEQKTDESA
DEIVVTNKRDTQVPTGVVGTLAPFAVLSIVAIGGVIYITKRKKA

Two E boxes, each containing a conserved glutamic residue, have been identified in SpyM30100. The E-box motifs are underlined in SEQ ID NO: 59, below. The conserved glutamic acid (E) residues, at amino acid residues 232 and 264, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of SpyM30100. Preferred fragments of SpyM30100 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 59

MKKNKLLLATAILATALGTASLNQNVKAETAGVSENAKLIVKKTFDSYTDNEVLMPKADYTFKVEADSTASGKTK
DGLEIKPGIVNGLTEQIISYTNTDKPDSKVKSTEFDFSKVVFPGIGVYRYTVSEKQGDVEGITYDTKKWTVDVYV
GNKEGGGFEPKFIVSKEQGTDVKKPVNFNNSFATTSLKVKKNVSGNTGELQKEFDFTLTLNESTNFKKDQIVSLQ
KGNEKFEVKIGTPYKFKLKNGESIQLDKLPVGITYKVNEMEANKDGYKTTASLKEGDGQSKMYQLDMEQKTDESA
DEIVVTNKRDTQVPTGVVGTLAPFAVLSIVAIGGVIYITKRKKA

SpyM30101 is a SrtC2 type sortase. An example of an amino acid sequence of SpyM30101 is set forth in SEQ ID NO: 60.

SEQ ID NO: 60

MTIVQVINKAIDTLILIFCLVVLFLAGFGLWDSYHLYQQADASNFKKFKTAQQQPKFEDLLALNEDVIGWLNIPG
THIDYPLVQGKTNLEYINKAVDGSVAMSGSLFLDTRNHNDFTDDYSLIYGHHMAGNAMFGEIPKFLKKDFFSKHN
KAIIETKERKKLTVTIFACLKTDAFNQLVFNPNAITNQDQQRQLVDYISKRSKQFKPVKLKHHTKFVAFSTCENF
STDNRVIVVGTIQE

SpyM30102 is referred to as a hypothetical protein. An example of an amino acid sequence of SpyM30102 is set forth in SEQ ID NO: 61.

SEQ ID NO: 61

MILTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTPFSIALESIDAMKTIEEITIAGSGKASFSPLTFTTVGQY
TYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKPKWLVKPIPPRQPNIPKTP*L*
*PLAG*EVKSLLGILSIVLLGLLVLLYVKKLKSRL

SpyM30102 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 185** LPLAG (shown in italics in SEQ ID NO: 61, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyM30102 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in SpyM30102. The pilin motif sequence is underlined in SEQ ID NO: 61, below. The conserved lysine (K) residue is also marked in bold, at amino acid residue 132. The pilin sequence, in

particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyM30102 include the conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 61

MILTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTPFSIALESIDAMKTIEEITIAGSGKASFSPLTFTTVGQY TYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKPKWLVKPIPPRQPNIPKTPL PLAGEVKSLLGILSIVLLGLLVLLYVKKLKSRL

Two E boxes containing conserved glutamic residues have been identified in SpyM30102. The E-box motifs are underlined in SEQ ID NO: 61, below. The conserved glutamic acid (E) residues, at amino acid residues 52 and 122, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of SpyM30102. Preferred fragments of SpyM30102 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 61

MILTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTPFSIALESIDAMKTIEEITIAGSGKASFSPLTFTTVGQY TYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKPKWLVKPIPPRQPNIPKTPL PLAGEVKSLLGILSIVLLGLLVLLYVKKLKSRL

SpyM30103 is referred to as a putative multiple sugar metabolism regulator. An example of an amino acid sequence for SpyM3103 is set forth in SEQ ID NO: 62.

SEQ ID NO: 62

MVRFDLKHVQTLHSLSQLPISVMSQDKALIQVYGNDDYLLCYYQFLKHLAIPQAAQDVIFYEGLFEESFMIFPLC HYIIAIGPFYPYSLNKDYQEQLANNCLKHSSHRSKEELLSYMALVPHFPINNVRNLLIAIDAFFDTQFETTCQQT IHQLLQHSKQMTADPDIIHRLKHISKASSQLPPVLEHLNHIMDLVKLGNPQLLKQEINRIPLSSITSSSISALRA EKNLTVIYLTRLLEFSFVENTDVAKHYSLVKYYMALNEEASDLLKVLRIRCAAIIHFSESLTNKSISDKRQMYNS VLHYVDSHLYSKLKVSDIAKRLYVSESHLRSVFKKYSNVSLQHYILSTKIKEAQLLLKRGIPVGEVAKSLYFYDT THFHKIFKKYTGISSKDYLAKYRDNI

SpyM30104 is thought to be a F2 like fibronectic binding protein. An example of an amino acid sequence for SpyM30104 is set forth in SEQ ID NO: 63.

SEQ ID NO: 63

MSSSDEETLKQYASKYTSNRRGDTSGNLKKQIAKVLTEGYPTNKSDWLNGLTENEKIEVTQDAIWYFTETTVPAD RSYTNRNVNSQKMKEVYQKLIDTTDIDKYEDVQFDLFVPQDTNLQAVISVEPVIESLPWTSLKPIAQKDITAKKI WVDAPKEKPIIYFKLYRQLPGEKEVAVDDAELKQINSEGQQEISVTWTNQLVTDEKGMAYIYSVKEVDKNGELLE PKDYIKKEDGLTVTNTYVKPTSGHYDIEVTFGNGHIDITEDTTPDIVSGENQMKQIEGEDSKPIDEVTENNLIEF GKNTMPGEEDGTNSNKYEEVEDSRPVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRDIDGKELAGATMELRDSS GKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEVATAITFTVNEQGQVTVNGKATKGDAHIVMVDAYKPTKGS GQVIDIEEKLPDEQGHSGSTTEIEDSKSSDVIIGGQGEVVDTTEDTQSGMTGHSGSTTEIEDSKSSDVIIGGQGE VVDTTEDTQSGMTGHSGSTTKIEDSKSSDVIVGGQGQIVETTEDTQTGMHGDSGRKTEVEDTKLVQSFHFDNKEP ESNSEIPKKDKSKSNT*SLPATG*EKQHNKFFWMVTSCSLISSVFVISLKSKKRLSSC

SpyM30104 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 180** LPATG (shown in italics in SEQ ID NO: 63, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyM30104 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Two pilin motifs, discussed above, containing conserved lysine (K) residues have also been identified in SpyM30104. The pilin motif sequences are underlined in SEQ ID NO: 63, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 156 and 227. The pilin sequences, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyM30104 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 63

MSSSDEETLKQYASKYTSNRRGDTSGNLKKQIAKVLTEGYPTNKSDWLNGLTENEKIEVTQDAIWYFTETTVPAD
RSYTNRNVNSQKMKEVYQKLIDTTDIDKYEDVQFDLFVPQDTNLQAVISVEPVIESLPWTSLKPIAQKDITAKKI
WVDAPKEKPIIYFKLYRQLPGEKEVAVDDAELKQINSEGQQEISVTWTNQLVTDEKGMAYIYSVKEVDKNGELLE
PKDYIKKEDGLTVTNTYVKPTSGHYDIEVTFGNGHIDITEDTTPDIVSGENQMKQIEGEDSKPIDEVTENNLIEF
GKNTMPGEEDGTNSNKYEEVEDSRPVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRDIDGKELAGATMELRDSS
GKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEVATAITFTVNEQGQVTVNGKATKGDAHIVMVDAYKPTKGS
GQVIDIEEKLPDEQGHSGSTTEIEDSKSSDVIIGGQGEVVDTTEDTQSGMTGHSGSTTEIEDSKSSDVIIGGQGE
VVDTTEDTQSGMTGHSGSTTKIEDSKSSDVIVGGQGQIVETTEDTQTGMHGDSGRKTEVEDTKLVQSFHFDNKEP
ESNSEIPKKDKSKSNTSLPATGEKQHNKFFWMVTSCSLISSVFVISLKSKKRLSSC

An E box containing a conserved glutamic residue has been identified in SpyM30104. The E-box motif is underlined in SEQ ID NO: 63, below. The conserved glutamic acid (E), at amino acid residue 402, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of SpyM30104. Preferred fragments of SpyM30104 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 63

MSSSDEETLKQYASKYTSNRRGDTSGNLKKQIAKVLTEGYPTNKSDWLNGLTENEKIEVTQDAIWYFTETTVPAD
RSYTNRNVNSQKMKEVYQKLIDTTDIDKYEDVQFDLFVPQDTNLQAVISVEPVIESLPWTSLKPIAQKDITAKKI
WVDAPKEKPIIYFKLYRQLPGEKEVAVDDAELKQINSEGQQEISVTWTNQLVTDEKGMAYIYSVKEVDKNGELLE
PKDYIKKEDGLTVTNTYVKPTSGHYDIEVTFGNGHIDITEDTTPDIVSGENQMKQIEGEDSKPIDEVTENNLIEF
GKNTMPGEEDGTNSNKYEEVEDSRPVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRDIDGKELAGATMELRDSS
GKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEVATAITFTVNEQGQVTVNGKATKGDAHIVMVDAYKPTKGS
GQVIDIEEKLPDEQGHSGSTTEIEDSKSSDVIIGGQGEVVDTTEDTQSGMTGHSGSTTEIEDSKSSDVIIGGQGE
VVDTTEDTQSGMTGHSGSTTKIEDSKSSDVIVGGQGQIVETTEDTQTGMHGDSGRKTEVEDTKLVQSFHFDNKEP
ESNSEIPKKDKSKSNTSLPATGEKQHNKFFWMVTSCSLISSVFVISLKSKKRLSSC

Examples of GAS AI-3 sequences from M3 strain isolate SSI-1 are set forth below.

Sps0099 is a negative transcriptional regulator (Nra). An example of an amino acid sequence for Sps0099 is set forth in SEQ ID NO: 64.

SEQ ID NO: 64

MPYVKKKKDSFLVETYLEQSIRDKSELVLLLFKSPTIIFSHVAKQTGLTAVQLKYYCKELDDFFGNNLDITIKKG
KIICCFVKPVKEFYLHQLYDTSTILKLLVFFIKNGTSSQPLIKFSKKYFLSSSSAYRLRESLIKLLREFGLRVSK
NTIVGEEYRIRYLIAMLYSKFGIVIYPLDHLDNQIIYRFLSQSATNLRTSPWLEEPFSFYNMLLALSWKRHQFAV
SIPQTRIFRQLKKLFIYDCLTRSSRQVIENAFSLTFSQGDLEYLFLIYITTNNSFASLQWTPQHIETCCHIFEKN
DTFRLLLEPILKRLPQLNHSKQDLIKALMYFSKSFLFNLQHFVIEIPSFSLPTYTGNSNLYKALKNIVNQWLAQL
PGKRHLNEKHLQLFCSHIEQILKNKQPALTVVLISSNFINAKLLTDTIPRYFSDKGIHFYSFYLLRDDIYQIPSL
KPDLVITHSRLIPFVKNDLVKGVTVAEFSFDNPDYSIASIQNLIYQLKDKKYQDFLNEQLQ

Sps0100 is thought to be a collagen binding protein (Cbp). It contains a sortase substrate motif VPXTG shown in italics in SEQ ID NO: 65.

SEQ ID NO: 65

MQKRDKTNYGSANNKRRQTTIGLLKVELTFVALIGIVGFSIRAFGAEEQSVPNKQSSVQDYPWYGYDSYSKGYPD
YSPLKTYHNLKVNLDGSKEYQAYCFNLTKHFPSKSDSVRSQWYKKLEGTNENFIKLADKPRIEDGQLQQNILRIL
YNGYPNDRNGIMKGIDPLNAILVTQNAIWYYTDSSYISDTSKAFQQEETDLKLDSQQLQLMRNALKRLINPKEVE
SLPNQVPANYQLSIFQSSDKTFQNLLSAEYVPDTPPKPGEEPPAKTEKTSVIIRKYAEGDYSKLLEGATLKLAQI
EGSGFQEKIFDSNKSGEKVELPNGTYVLSELKPPQGYGVATPITFKVAAEKVLIKNKEGQFVENQNKEIAEPYSV
TAFNDFEEIGYLSDFNNYGKFYYAKNTNGTNQVVYCFNADLHSPPDSYDHGANIDPDVSESKEIKYTHVSGYDLY
KYAATPRDKDADFFLKHIKKILDKGYKKKGDTYKTLTEAQFRAATQLAIYYYTDSADLTTLKTYNDNKGYHGFDK
LDDATLAVVHELITYAEDVTLPMTQNLDFFVPNSSRYQALIGTQYHPNELIDVISMEDKQAPIIPITHKLTISKT
VTGTIADKKKEFNFEIHLKSSDGQAISGTYPTNSGELTVTDGKATFTLKDGESLIVEGLPSGYSYEITETGASDY
EVSVNGKNAPDGKATKASVKEDETVAFENRKDL*VPPTG*LTTDGAIYLWLLLLVPFGLLVWLFGRKGTKK

Sps0101 is referred to as a LepA protein. An example of an amino acid sequence of Sps0101 is set forth as SEQ ID NO: 66

SEQ ID NO: 66

MTNYLNRLNENPLLKAFIRLVLKISIIGFLGYILFQYVFGVMIVNTNQMSPAVSAGDGVLYYRLTDRYHINDVVV
YEVDDTLKVGRIAAQAGDEVNFTQEGGLLINGHPPEKEVPYLTYPHSSGPNFPYKVPTGTYFILNDYREERLDSR
YYGALPINQIKGKISTLLRVRGI

Sps0102 is thought to be a fimbrial protein. It contains a sortase substrate motif QVXTG shown in italics in SEQ ID NO: 67.

SEQ ID NO: 67

MEREKMKKNKLLLATAILATALGTASLNQNVKAETAGVSENAKLIVKKTFDSYTDNEVLMPKADYTFKVEADSTA
SGKTKDGLEIKPGIVNGLTEQIISYTNTDKPDSKVKSTEFDFSKVVFPGIGVYRYTVSEKQGDVEGITYDTKKWT
VDVYVGNKEGGGFEPKFIVSKEQGTDVKKPVNFNNSFATTSLKVKKNVSGNTGELQKEFDFTLTLNESTNFKKDQ
IVSLQKGNEKFEVKIGTPYKFKLKNGESIQLDKLPVGITYKVNEMEANKDGYKTTASLKEGDGQSKMYQLDMEQK
TDESADEIVVTNKRDT*QVPTG*VVGTLAPFAVLSIVAIGGVIYITKRKKA

Sps0103 is a SrtC2 type sortase. An example of Sps0103 is set forth in SEQ ID NO: 68.

SEQ ID NO: 68

MVMTIVQVINKAIDTLILIFCLVVLFLAGFGLWDSYHLYQQADASNFKKFKTAQQQPKFEDLLALNEDVIGWLNI
PGTHIDYPLVQGKTNLEYINKAVDGSVAMSGSLFLDTRNHNDFTDDYSLIYGHHMAGNAMFGEIPKFLKKDFFSK
HNKAIIETKERKKLTVTIFACLKTDAFNQLVFNPNAITNQDQQRQLVDYISKRSKQFKPVKLKHHTKFVAFSTCE
NFSTDNRVIVVGTIQE

Sps0104 is referred to as a hypothetical protein. It contains a sortase substrate motif LPXAG shown in italics in SEQ ID NO: 69.

SEQ ID NO: 69

MLFSVVMILTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTPFSIALESIDAMKTIEEITIAGSGKASFSPLTF
TTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKPKWLVKPIPPRQPN
IPKTP*LPLAG*EVKSLLGILSIVLLGLLVLLYVKKLKSRL

Sps0105 is referred to as a putative multiple sugar metabolism regulator. An example of Sps0105 is set forth in SEQ ID NO: 70.

SEQ ID NO: 70

MALVPHFPINNVRNLLIAIDAFFDTQFETTCQQTIHQLLQHSKQMTADPDIIHRLKHISKASSQLPPVLEHLNHI
MDLVKLGNPQLLKQEINRIPLSSITSSSISALRAEKNLTVIYLTRLLEFSFVENTDVAKHYSLVKYYMALNEEAS
DLLKVLRIRCAAIIHFSESLTNKSISDKRQMYNSVLHYVDSHLYSKLKVSDIAKRLYVSESHLRSVFKKYSNVSL
QHYILSTKIKEAQLLLKRGIPVGEVAKSLYFYDTTHFHKIFKKYTGISSKDYLAKYRDNI

Sps0106 is thought to be a F2 like fibronectic binding protein. It contains a sortase substrate LPXTG (SEQ ID NO: 122) shown in italics in SEQ ID NO: 71.

SEQ ID NO: 71

MTQKNSYKLSFLLSLTGETLGLLLVTEGLSGVSVGHAETRNGANKQGAFEIKKNKSQEEYNYEVYDNRNILQDGE
HKLEIKRVDGTGKTYQGFCFQLTKNFPTAQGVSKKLYKKLSSSDEETLKQYASKYTSNRRGDTSGNLKKQIAKVL
TEGYPTNKSDWLNGLTENEKIEVTQDAIWYFTETTVPADRSYTNRNVNSQKMKEVYQKLIDTTDIDKYEDVQFDL
FVPQDTNLQAVISVEPVIESLPWTSLKPIAQKDITAKKIWVDAPKEKPIIYFKLYRQLPGEKEVAVDDAELKQIN
SEGQQEISVTWTNQLVTDEKGMAYIYSVKEVDKNGELLEPKDYIKKEDGLTVTNTYVKPTSGHYDIEVTFGNGHI
DITEDTTPDIVSGENQMKQIEGEDSKPIDEVTENNLIEFGKNTMPGEEDGTNSNKYEEVEDSRPVDTLSGLSSEQ
GQSGDMTIEEDSATHIKFSKRDIDGKELAGATMELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEV
ATAITFTVNEQGQVTVNGKATKGDAHIVMVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKSSDVIIGGQ
GEVVDTTEDTQSGMTGHSGSTTKIEDSKSSDVIVGGQGQIVETTEDTQTGMHGDSGRKTEVEDTKLVQSFHFDNK
EPESNSEIPKKDKSKSNTS*LPATG*EKQHNKFFWMVTSCSLISSVFVISLKSKKRLSSC

Examples of GAS AI-3 sequences from M5 isolate Manfredo are set forth below.

Orf 77 encodes a negative transcription regulator (Nra). An example of the nucleotide sequence encoding Nra (SEQ ID NO: 88) and an Nra amino acid sequence (SEQ ID NO: 89) are set forth below.

SEQ ID NO: 88

ATGCCTTATGTCAAAAAGAAAAAGGATAGTTTCTTAGTAGAAACATATCTTGAACAGTCTATTAGAGATAAAAGT
GAATTAGTCTTACTGTTATTTAAATCGCCTACTATCATTTTTTCTCATGTTGCTAAACAAACTGGTCTGACGGCT
GTACAATTAAAATATTACTGTAAAGAACTTGATGACTTTTTTGGAAATAATTTAGACATTACCATTAAAAAGGGC
AAAATAATATGTTGTTTTGTCAAACCTGTTAAGGAATTCTACCTTCATCAACTCTATGACACATCAACAATATTA
AAATTATTAGTTTTCTTTATTAAAAATGGAACGTCATCACAACCTCTGATTAAATTTTCAAAAAAGTATTTTCTA
TCAAGCTCCTCAGCTTATCGACTACGGGAATCGCTGATCAAATTACTACGGGAATTTGGCTTGAGAGTCTCAAAA
AATACAATTGTCGGAGAGGAATATCGTATTCGCTATCTTATTGCCATGCTATATAGTAAATTTGGCATTGTCATC
TATCCGTTAGATCATCTAGACAATCAAATTATTTATCGCTTCTTATCACAAAGTGCAACCAATTTAAGAACATCG
CCCTGGCTAGAGGAACCTTTTTCTTTTTATAATATGTTACTTGCCTTGTCATGGAAACGTCACCAATTTGCAGTT
AGCATTCCTCAAACACGTATTTTTCGACAATTAAAAAAGCTTTTTATCTATGATTGTTTAACTCGAAGCAGTCGA
CAAGTAATCGAAAATGCTTTTTCGTTAATGTTCTCACAAGGAGATCTCGATTATCTTTTTTTAATTTATATTACC
ACCAATAATTCCTTTGCCAGCCTACAATGGACTCCACAGCATATTGAAACTTGCTGCCATATTTTTGAAAAAAAT
GACACATTTCGGTTATTGTTAGAGCCCATTCTTAAACGTTTACCGCAATTAAACCATTCTAAACAAGACCTTATT
AAAGCCCTTATGTATTTTTCAAAATCTTTTCTATTTAACCTCCAACATTTCGTCATCGAGATTCCTTCTTTTTCC
TTGCCGACCTATACAGGCAACTCTAATCTTTACAAAGCTTTAAAAAATATTGTAAATCAGTGGCTTGCTCAATTA
CCCGGAAAGCGTCATCTTAACGAAAAGCATCTCCAACTTTTTTGCTCTCATATTGAACAAATCTTAAAAAATAAA
CAACCTGCTTTAACTGTCGTTTTAATATCTAGTAACTTTATAAATGCTAAACTCCTTACAGATACTATCCCACGA
TATTTTTCTGATAAAGGAATTCATTTTTATTCTTTTTACTTATTAAGAGATGATATCTATCAAATTCCAAGCTTA
AAACCAGATTTAGTTATCACTCATAGCCGATTAATTCCTTTTGTTAAGAATGATCTGGTCAAAGGTGTTACTGTT
GCTGAATTTTCTTTTGATAACCCTGACTACTCTATTGCTTCAATTCAAAACTTGATATATCAGCTCAAAGATAAA
AAATATCAAGATTTTCTAAACGAGCAATTACAA

SEQ ID NO: 89

MPYVKKKKDSFLVETYLEQSIRDKSELVLLLFKSPTIIFSHVAKQTGLTAVQLKYYCKELDDFFGNNLDITIKKG
KIICCFVKPVKEFYLHQLYDTSTILKLLVFFIKNGTSSQPLIKFSKKYFLSSSSAYRLRESLIKLLREFGLRVSK
NTIVGEEYRIRYLIAMLYSKFGIVIYPLDHLDNQIIYRFLSQSATNLRTSPWLEEPFSFYNMLLALSWKRHQFAV
SIPQTRIFRQLKKLFIYDCLTRSSRQVIENAFSLMFSQGDLDYLFLIYITTNNSFASLQWTPQHIETCCHIFEKN
DTFRLLLEPILKRLPQLNHSKQDLIKALMYFSKSFLFNLQHFVIEIPSFSLPTYTGNSNLYKALKNIVNQWLAQL
PGKRHLNEKHLQLFCSHIEQILKNKQPALTVVLISSNFINAKLLTDTIPRYFSDKGIHFYSFYLLRDDIYQIPSL
KPDLVITHSRLIPFVKNDLVKGVTVAEFSFDNPDYSIASIQNLIYQLKDKKYQDFLNEQLQ

Orf 78 is thought to be a collagen binding protein (Cbp). An example of the nucleotide sequence encoding Cbp (SEQ ID NO: 90) and a Cbp amino acid sequence (SEQ ID NO: 91) are set forth below.

SEQ ID NO: 90

TTGCAAAAGAGGGATAAAACCAATTATGGAAGCGCTAACAACAAACGACGACAAACGACGATCGGATTACTGAAA
GTATTTTTGACGTTTGTAGCTCTGATAGGAATAGTAGGGTTTTCTATCAGAGCGTTCGGAGCTGAAGAAAAATCT
ACTGAAACTAAAAAAACGTCAGTCATTATTAGAAAATATGCTGAAGGTGACTACTCTAAACTTCTAGAGGGAGCA
ACTTTGCGTTTAACAGGGGAAGATATCCCAGATTTTCAAGAAAAAGTCTTCCAAAGTAATGGAACAGGAGAAAAG
ATTGAATTATCAAATGGGACTTATACCTTAACAGAAACATCATCTCCAGATGGATATAAAATTACGGAGCCGATT

AAGTTTAGAGTAGTGAATAAAAAAGTATTTATCGTCCAAAAAGATGGTTCTCAAGTGGAAAACCCAAACAAAGAA
CTAGGTTCTCCATATACTATAGAGGCATACAATGATTTTGATGAATTTGGCTTACTGTCAACACAAAATTATGCG
AAATTTTATTATGGAAAAAACTATGATGGCAGTTCACAAATTGTTTATTGCTTCAATGCCAACTTGAAATCTCCA
CCTGACTCGGAAGATCATGGTGCTACAATAAATCCTGACTTTACGACTGGTGATATTAGGTACAGTCATATTGCT
GGTTCAGATTTGATAAAATACGCTAATACAGCTAGGGATGAAGATCCTCAATTATTTTTAAAACACGTAAAAAAA
GTAATTGAAAATGGGTATCATAAAAAAGGTCAAGCTATTCCATATAACGGTCTGACTGAGGCACAGTTTCGTGCG
GCTACTCAACTGGCAATTTATTATTTTACAGATAGTGTTGACTTAACTAAGGATAGATTGAAAGACTTCCATGGA
TTTGGAGATATGAATGATCAAACTTTGGGTGTAGCTAAAAAAATTGTAGAATACGCTTTGAGTGATGAAGATTCA
AAACTAACAAATCTTGATTTCTTCGTACCTAATAATAGCAAATACCAATCTCTTATTGGGACAGAATACCATCCA
GATGATTTGGTTGACGTGATTCGTATGGAAGATAAAAAGCAAGAAGTTATTCCAGTAACTCATAGTTTGACGGTG
CAAAAAACAGTAGTCGGTGAGTTGGGAGATAAGACTAAAGGCTTTCAATTTGAACTTGAGTTGAAAGATAAAACT
GGACAGCCTATTGTTAACACTCTAAAAACTAATAATCAAGATTTAGTAGCTAAAGATGGGAAATATTCATTTAAT
CTAAAGCATGGTGACACCATAAGAATAGAAGGATTACCGACGGGATATTCTTATACCCTGAAAGAGACTGAAGCT
AAGGATTATATAGTAACTGTTGATAACAAAGTTAGTCAAGAAGCTCAATCAGCAAGTGAGAATGTCACAGCAGAC
AAAGAAGTCACTTTTGAAAACCGAAAAGATCTTGTCCCACCAACTGGTTTGACAACAGATGGGGCTATCTATCTT
TGGTTATTACTACTTGTTCCATTTGGGTTATTGGTTTGGCTATTTGGTCGTAAAGGGTTAAAAAATGAC

SEQ ID NO: 91

MQKRDKTNYGSANNKRRQTTIGLLKVFLTFVALIGIVGFSIRAFGAEEKSTETKKTSVIIRKYAEGDYSKLLEGA
TLRLTGEDIPDFQEKVFQSNGTGEKIELSNGTYTLTETSSPDGYKITEPIKFRVVNKKVFIVQKDGSQVENPNKE
LGSPYTIEAYNDFDEFGLLSTQNYAKFYYGKNYDGSSQIVYCFNANLKSPPDSEDHGATINPDFTTGDIRYSHIA
GSDLIKYANTARDEDPQLFLKHVKKVIENGYHKKGQAIPYNGLTEAQFRAATQLAIYYFTDSVDLTKDRLKDFHG
FGDMNDQTLGVAKKIVEYALSDEDSKLTNLDFFVPNNSKYQSLIGTEYHPDDLVDVIRMEDKKQEVIPVTHSLTV
QKTVVGELGDKTKGFQFELELKDKTGQPIVNTLKTNNQDLVAKDGKYSFNLKHGDTIRIEGLPTGYSYTLKETEA
KDYIVTVDNKVSQEAQSASENVTADKEVTFENRKDL*VPPTG*LTTDGAIYLWLLLLVPFGLLVWLFGRKGLKND

Orf 78 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 184** VPPTG (shown in italics in SEQ ID NO: 91, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant Orf 78 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Three E boxes containing conserved glutamic residues have been identified in Orf 78. The E-box motifs are underlined in SEQ ID NO: 91, below. The conserved glutamic acid (E) residues, at amino acid residues 112, 395, and 447, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of Orf 78. Preferred fragments of Orf 78 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 91

MQKRDKTNYGSANNKRRQTTIGLLKVFLTFVALIGIVGFSIRAFGAEEKSTETKKTSVIIRKYAEGDYSKLLEGA
TLRLTGEDIPDFQEKVFQSNGTGEKIELSNGTYTLT**E**TSSPDGYKITEPIKFRVVNKKVFIVQKDGSQVENPNKE
LGSPYTIEAYNDFDEFGLLSTQNYAKFYYGKNYDGSSQIVYCFNANLKSPPDSEDHGATINPDFTTGDIRYSHIA
GSDLIKYANTARDEDPQLFLKHVKKVIENGYHKKGQAIPYNGLTEAQFRAATQLAIYYFTDSVDLTKDRLKDFHG
FGDMNDQTLGVAKKIVEYALSDEDSKLTNLDFFVPNNSKYQSLIGTEYHPDDLVDVIRMEDKKQEVIPVTHSLTV
QKTVVGELGDKTKGFQFEL**E**LKDKTGQPIVNTLKTNNQDLVAKDGKYSFNLKHGDTIRIEGLPTGYSYTLK**E**TEA
KDYIVTVDNKVSQEAQSASENVTADKEVTFENRKDLVPPTGLTTDGAIYLWLLLLVPFGLLVWLFGRKGLKND

Orf 79 is thought to be a LepA signal peptidase I. An example of the nucleotide sequence encoding a LepA signal peptidase I (SEQ ID NO: 92) and a LepA signal peptidase I amino acid sequence (SEQ ID NO: 93) are set forth below.

SEQ ID NO: 92

ATGACTAATTACCTAAATCGTTTAAATGAGAATTCACTATTTAAAGCTTTCATACGGTTAGTACTTAAGATTTCT
ATTATTGGGTTTCTAGGTTACATTCTATTTCAGTATGTTTTTGGTGTTATGATTATTAACACTAATGATATGAGT
CCTGCTTTAAGTGCAGGTGACGGTGTTTTATATTATCGTTTGACTGATCGCTATCATATTAATGATGTGGTGGTC
TATGAGGTTGATAACACTTTGAAAGTTGGTCGAATTGTCGCTCAAGCTGGCGATGAGGTTAGTTTTACGCAAGAA
GGAGGACTGTTGATTAATGGGCATCCACCAGAAAAAGAGGTCCCTTACCTGACGTATCCTCACTCAAGTGGCCCA
AACTTTCCCTATAAAGTTCCTACGGGTAAGTATTTCATATTGAATGATTATCGTGAAGAACGTTTGGACAGTCGT
TATTATGGGGCGTTACCCGTCAATCAAATAAAAGGGAAAATCTCAACTCTATTAAGAGTGAGAGGAATT

SEQ ID NO: 93

MTNYLNRLNENSLFKAFIRLVLKISIIGFLGYILFQYVFGVMIINTNDMSPALSAGDGVLYYRLTDRYHINDVVV
YEVDNTLKVGRIVAQAGDEVSFTQEGGLLINGHPPEKEVPYLTYPHSSGPNFPYKVPTGKYFILNDYREERLDSR
YYGALPVNQIKGKISTLLRVRGI

Orf 80 is thought to to be a fimbrial protein. An example of the nucleotide sequence encoding the fimbrial protein (SEQ ID NO: 94) and a fimbrial protein amino acid sequence (SEQ ID NO: 95) are set forth below.

SEQ ID NO: 94

TTGGAGAGAGAAAAAATGAAAAAAAACAAATTATTACTTGCTACTGCAATCTTAGCAACTGCTTTAGGAACAGCT
TCTTTAAATCAAAACGTAAAAGCTGAGACGGCAGGGGTTGTAACAGGAAAATCACTACAAGTTACAAAGACAATG
ACTTATGATGATGAAGAGGTGTTAATGCCCGAAACCGCCTTTACTTTTACTATAGAGCCTGATATGACTGCAAGT
GGAAAAGAAGGCAGCCTAGATATTAAAAATGGAATTGTAGAAGGCTTAGACAAACAAGTAACAGTAAAATATAAG
AATACAGATAAAACATCTCAAAAAACTAAAATAGCACAATTTGATTTTTCTAAGGTTAAATTTCCAGCTATAGGT
GTTTACCGCTATATGGTTTCAGAGAAAAACGATAAAAAAGACGGAATTACGTACGATGATAAAAAGTGGACTGTA
GATGTTTATGTTGGGAATAAGGCCAATAACGAAGAAGGTTTCGAAGTTCTATATATTGTATCAAAAGAAGGTACT
TCTAGTACTAAAAAACCAATTGAATTTACAAACTCTATTAAAACTACTTCCTTAAAAATTGAAAAACAAATAACT
GGCAATGCAGGAGATCGTAAAAAATCATTCAACTTCACATTAACATTACAACCAAGTGAATATTATAAAACTGGA
TCAGTTGTGAAAATCGAACAGGATGGAAGTAAAAAAGATGTGACGATAGGAACGCCTTACAAATTTACTTTGGGA
CACGGTAAGAGTGTCATGTTATCGAAATTACCAATTGGTATCAATTACTATCTTAGTGAAGACGAAGCGAATAAA
GACGGCTACACTACAACGGCAACATTAAAAGAACAAGGCAAAGAAAAGAGTTCCGATTTCACTTTGAGTACTCAA
AACCAGAAAACAGACGAATCTGCTGACGAAATCGTTGTCACAAATAAGCGTGACACTCAAGTTCCAACTGGTGTT
GTAGGGACCCTTGCTCCATTTGCAGTTCTTAGCATTGTGGCTATTGGTGGAGTTATCTATATTACAAAACGTAAA
AAAGCT

SEQ ID NO: 95

MEREKMKKNKLLLATAILATALGTASLNQNVKAETAGVVTGKSLQVTKTMTYDDEEVLMPETAFTFTIEPDMTAS
GKEGSLDIKNGIVEGLDKQVTVKYKNTDKTSQKTKIAQFDFSKVKFPAIGVYRYMVSEKNDKKDGITYDDKKWTV
DVYVGNKANNEEGFEVLYIVSKEGTSSTKKPIEFTNSIKTTSLKIEKQITGNAGDRKKSFNFTLTLQPSEYYKTG
SVVKIEQDGSKKDVTIGTPYKFTLGHGKSVMLSKLPIGINYYLSEDEANKDGYTTTATLKEQGKEKSSDFTLSTQ
NQKTDESADEIVVTNKRDT*QVPTG*VVGTLAPFAVLSIVAIGGVIYITKRKKA

Orf 82 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 140** QVPTG (shown in italics in SEQ ID NO: 95, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant Orf 82 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

An E box containing a conserved glutamic residue has been identified in Orf 80. The E-box motif is underlined in SEQ ID NO: 95, below. The conserved glutamic acid (E), at amino acid residue 270, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is

thought to be important for the formation of oligomeric pilus-like structures of Orf 80. Preferred fragments of Orf 80 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 95

MEREKMKKNKLLLATAILATALGTASLNQNVKAETAGVVTGKSLQVTKTMTYDDEEVLMPETAFTFTIEPDMTAS
GKEGSLDIKNGIVEGLDKQVTVKYKNTDKTSQKTKIAQFDFSKVKFPAIGVYRYMVSEKNDKKDGITYDDKKWTV
DVYVGNKANNEEGFEVLYIVSKEGTSSTKKPIEFTNSIKTTSLKIEKQITGNAGDRKKSFNFTLTLQPSEYYKTG
SVVKIEQDGSKKDVTIGTPYKFTLGHGKSVMLSKLPIGINYYLSEDEANKDGYTTTATLKEQGKEKSSDFTLSTQ
NQKTDESADEIVVTNKRDTQVPTGVVGTLAPFAVLSIVAIGGVIYITKRKKA

Orf 81 is thought to to be a SrtC2 type sortase. An example of the nucleotide sequence encoding the SrtC2 sortase (SEQ ID NO: 96) and a SrtC2 sortase amino acid sequence (SEQ ID NO: 97) are set forth below.

SEQ ID NO: 96

GTGATTAGTCAAAGAATGATGATGACAATTGTACAGGTTATCAATAAAGCCATTGATACTCTCATTCTTATCTTT
TGTTTAGTCGTACTATTTTTAGCTGGTTTTGGTTTGTGGGATTCTTATCATCTCTATCAACAAGCAGACGCTTCT
AATTTCAAAAAATTTAAAACAGCTCAACAACAGCCTAAATTTGAAGACTTGTTAGCTTTGAATGAGGATGTCATT
GGTTGGTTAAATATCCCAGGGACTCATATTGATTATCCTCTAGTTCAGGGAAAAACGAATTTAGAGTATATTAAT
AAAGCAGTTGATGGCAGTGTTGCCATGTCTGGTAGTTTATTTTTAGATACACGGAATCATAATGATTTTACGGAC
GATTACTCTCTGATTTATGGCCATCATATGGCAGGTAATGCCATGTTTGGCGAAATTCCAAAATTTTTAAAAAAG
GATTTTTTCAACAAACATAATAAAGCTATCATTGAAACAAAAGAGAGAAAAAAACTAACCGTCACTATTTTTGCT
TGTCTCAAGACAGATGCCTTTGACCAGTTAGTTTTTAATCCTAATGCTATTACCAATCAAGACCAACAAAAGCAG
CTCGTTGATTATATCAGTAAAAGATCAAAACAATTTAAACCTGTTAAATTGAAGCATCATACAAAGTTCGTTGCT
TTTTCAACGTGTGAAAATTTTTCTACTGACAATCGTGTTATCGTTGTCGGTACTATTCAAGAA

SEQ ID NO: 97

MISQRMMMTIVQVINKAIDTLILIFCLVVLFLAGFGLWDSYHLYQQADASNFKKFKTAQQQPKFEDLLALNEDVI
GWLNIPGTHIDYPLVQGKTNLEYINKAVDGSVAMSGSLFLDTRNHNDFTDDYSLIYGHHMAGNAMFGEIPKFLKK
DFFNKHNKAIIETKERKKLTVTIFACLKTDAFDQLVFNPNAITNQDQQKQLVDYISKRSKQFKPVKLKHHTKFVA
FSTCENFSTDNRVIVVGTIQE

Orf 82 is referred to as a hypothetical protein. It contains a sortase substrate motif LPXAG shown in italics in SEQ ID NO: 99. An example of the nucleotide sequence encoding the hypothetical protein (SEQ ID NO: 98) and a hypothetical protein amino acid sequence (SEQ ID NO: 99) are set forth below.

SEQ ID NO: 98

TTGCTTTTTCAACGTGTGAAAATTTTTCTACTGACAATCGTGTTATCGTTGTCGGTACTATTCAAGAATAACGAA
AGGAGGAGACTTTTGAGAAAATATTGGAAAATGTTATTTTCTGTCGTAATGATATTAACCATGCTGGCCTTTAAT
CAGACTGTTTTAGCAAAAGACAGCACTGTTCAAACTAGCATTAGTGTCGAAAATGTCTTAGAGAGAGCAGGCGAT
AGTACCCCATTTTCGGTTGCATTAGAATCAATTGATGCGATGAAAACAATAGACGAAATAACAATTGCTGGTTCT
GGAAAAGCAAGCTTTTCCCCTCTGACCTTCACAACAGTTGGGCAATATACTTATCGTGTTTATCAGAAGCCTTCA
CAAAATAAAGATTATCAAGCAGATACTACTGTATTTGACGTTCTTGTCTATGTGACCTATGATGAAGATGGGACT
CTAGTCGCAAAAGTTATTTCTCGAAGGGCTGGAGACGAAGAAAAATCAGCGATTACTTTTAAGCCCAAACGGTTA
GTAAAACCAATACCGCCTAGACAACCTAACATCCCTAAAACCCCATTACCATTAGCTGGTGAAGTAAAAAGTTTA
TTGGGTATCTTAAGTATCGTATTACTGGGGTTACTAGTTCTTCTTTATGTTAAAAAACTGAAGAGTAGGCTA

SEQ ID NO: 99

MLFQRVKIFLLTIVLSLSVLFKNNERRRLLRKYWKMLFSVVMILTMLAFNQTVLAKDSTVQTSISVENVLERAGD
STPFSVALESIDAMKTIDEITIAGSGKASFSPLTFTTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGT
LVAKVISRRAGDEEKSAITFKPKRLVKPIPPRQPNIPKTP*LPLAG*EVKSLLGILSIVLLGLLVLLYVKKLKSRL

PC Orf 82 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 185** LPLAG (shown in italics in SEQ ID NO: 99, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant Orf 82 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in Orf 82. The pilin motif sequence is underlined in SEQ ID NO: 99, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 173 and 188. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of Orf 82 include at least one conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 99

MLFQRVKIFLLTIVLSLSVLFKNNERRRLLRKYWKMLFSVVMILTMLAFNQTVLAKDSTVQTSISVENVLERAGD
STPFSVALESIDAMKTIDEITIAGSGKASFSPLTFTTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGT
LVAKVISRRAGDEEKSAITFKPKRLVKPIPPRQPNIPKTPLPLAGEVKSLLGILSIVLLGLLVLLYVKKLKSRL

An E box containing a conserved glutamic residue has been identified in Orf 82. The E-box motif is underlined in SEQ ID NO: 99, below. The conserved glutamic acid (E), at amino acid residue 163, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of Orf 82. Preferred fragments of Orf 82 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 99

MLFQRVKIFLLTIVLSLSVLFKNNERRRLLRKYWKMLFSVVMILTMLAFNQTVLAKDSTVQTSISVENVLERAGD
STPFSVALESIDAMKTIDEITIAGSGKASFSPLTFTTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGT
LVAKVISRRAGDEEKSAITFKPKRLVKPIPPRQPNIPKTPLPLAGEVKSLLGILSIVLLGLLVLLYVKKLKSRL

Orf 83 is thought to to be a multiple sugar metabolism regulator protein. An example of a nucleotide sequence encoding the sugar metabolism regulator protein (SEQ ID NO: 100) and a sugar metabolism regulator protein amino acid sequence (SEQ ID NO: 101) are set forth below.

SEQ ID NO: 100

ATGATACAACTAAGGATGGGGGCAATCTATCAAATGGTTATATTCGATTTAAAACATGTGCAAACATTACACAGC
TTGTCTCAATTACCTATTTCAGTGATGTCACAAGATAAGGCACTTATTCAAGTATATGGTAATGACGACTATTTA
TTATGTTACTATCAATTTTTAAAGCATCTAGCTATTCCTCAAGCTGCACAAGATGTTATTTTTTATGAGGGTTTA
TTTGAAGAGTCCTTTATGATTTTTCCTCTTTGTCACTACATTATTGCCATTGGACCTTTCTATCCTTATTCACTT
AATAAAGACTATCAGGAACAATTAGCTAATAATTTTTTAAAACATTCTTCTCATCGTAGCAAAGAAGAGCTCTTG
TCCTATATGGCACTTGTCCCACATTTTCCAATTAATAATGTGCGGAACCTTTTGATAGCTATTGACGCTTTTTTT
GACACACAATTTGAGACGACTTGCCAACAAACGATTCATCAATTGTTGCAGCATTCAAAACAGATGACTGCTGAT
CCTGATATCATTCATCGCCTTAAGCATATTAGCAAAGCATCTAGCCAATTACCGCCTGTTTTAGAGCACCTAAAT
CATATTATGGATCTGGTAAAGCTAGGCAATCCACAATTGCTCAAGCAAGAAATCAATCGCATCCCCTTATCAAGT
ATCACCTCATCTTCTATTTCTGCTCTAAGGGCGGAAAAGAACCTCACTGTTATCTATTTAACTAGGTTACTGGAA
TTCAGTTTTGTAGAAAATACTGACGTAGCAAAGCATTATAGCCTTGTCAAATACTACATGGCCTTAAATGAAGAA
GCGAGTGACTTGCTCAAAGTTTTGAGAATTCGCTGTGCAGCTATCATCCATTTTTCCGAATCATTAACCAATAAA
AGTATTTCTGATAAACGTCAAATGTACAATAGTGTGCTTCATTATGTCGATAGTCACCTGTATTCCAAATTAAAG

GTATCTGATATCGCTAAGCGCCTATATGTTTCCGAATCTCACTTACGTTCAGTCTTTAAAAAATACTCAAATGTT
TCCTTACAACATTATATTCTAAGTACAAAAATCAAAGAAGCTCAACTACTCTTAAAACGAGGAATTCCTGTTGGA
GAAGTGGCTAAAAGCTTATATTTTTATGACACTACCCATTTTCATAAAATCTTTAAAAAATACACGGGTATTTCT
TCAAAAGACTATCTTGCTAAATACCGAGATAATATT

SEQ ID NO: 101

MIQLRMGAIYQMVIFDLKHVQTLHSLSQLPISVMSQDKALIQVYGNDDYLLCYYQFLKHLAIPQAAQDVIFYEGL
FEESFMIFPLCHYIIAIGPFYPYSLNKDYQEQLANNFLKHSSHRSKEELLSYMALVPHFFPINNVRNLLIAIDAFF
DTQFETTCQQTIHQLLQHSKQMTADPDIIHRLKHISKASSQLPPVLEHLNHIMDLVKLGNPQLLKQEINRIPLSS
ITSSSISALRAEKNLTVIYLTRLLEFSFVENTDVAKHYSLVKYYMALNEEASDLLKVLRIRCAAIIHFSESLTNK
SISDKRQMYNSVLHYVDSHLYSKLKVSDIAKRLYVSESHLRSVFKKYSNVSLQHYILSTKIKEAQLLLKRGIPVG
EVAKSLYFYDTTHFHKIFKKYTGISSKDYLAKYRDNI

Orf 84 is thought to to be a F2-like fibronectin-binding protein. An example of a nucleotide sequence encoding the F2-like fibronectin-binding protein (SEQ ID NO: 102) and a F2-like fibronectin-binding protein amino acid sequence (SEQ ID NO: 103) are set forth below.

SEQ ID NO: 102

ATGACACAAAAAAATAGCTATAAGTTAAGCTTCCTGTTATCCCTAACAGGATTTATTTTAGGTTTATTATTGGTT
TTTATAGGATTGTCCGGAGTATCAGTAGGACATGCGGAAACAAGAAATGGAGCAAACAAACAAGGAGCTTTTGAA
ATCAAGAAAAATAAAAGTCAAGAAGAATATAATTATGAAGTTTATGATAACAGAAACATACTTCAGGATGGGGAA
CATAAACTTGAAATAAAAAGAGTTGATGGGACAGGTAAAACTTATCAAGGTTTTTGCTTTCAGTTAACGAAAAAT
TTTCCCACTGCTCAAGGTGTAAGTAAAAAGCTGTATAAAAAATTGAGTAGTAGTGATGAAGAAACACTAAAGCAA
TATGCCTCTAAGTATACAAGTAATAGGAGAGGAGATACTAGTGGTAATCTTAAAAAGCAAATTGCTAAGGTTCTG
ACAGAAGGTTACCCAACTAACAAAAGTGATTGGTTAAATGGATTGACTGAAAACGAAAAAATAGAAGTAACCCAG
GATGCAATTTGGTATTTTACAGAAACGACAGTTCCGGCTGATAGAAGTTATACGAATCGCAACGTAAATAGTCAA
AAAATGAAAGAAGTGTATCAAAAGCTAATTGATACAACAGATATAGATAAATATGAAGATGTACAATTTGATTTA
TTTGTGCCACAAGATACAAACTTACAGGCAGTAATTAGTGTAGAGCCTGTTATCGAAAGCCTTCCTTGGACATCG
TTGAAGCCAATAGCCCAGAAGGATATCACTGCCAAAAAAATCTGGGTAGATGCACCTAAAGAAAAACCAATTATT
TATTTTAAGCTATATAGACAGCTGCCTGGAGAAAAGGAAGTAGCAGTGGATGACGCTGAGCTAAAACAGATAAAT
AGTGAAGGTCAACAAGAAATATCAGTAACTTGGACAAATCAACTTGTTACAGATGAAAAAGGAATGGCTTACATT
TATTCTGTAAAAGAAGTAGATAAAAATGGCGAGTTACTTGAGCCAAAAGATTATATCAAGAAGGAAGATGGACTT
ACAGTTACTAATACTTATGTAAAGCCAACTAGTGGGCACTATGATATAGAAGTGACATTTGGAAATGGACATATT
GATATTACAGAAGATACTACACCAGATATTGTTTCAGGTGAAAACCAAATGAAGCAAATAGAGGGAGAAGATAGT
AAGCCTATTGATGAAGTAACGGAAAATAATTTAATTGAATTTGGTAAAAACACGATGCCAGGTGAAGAAGATGGC
ACAAATTCTAATAAGTATGAAGAAGTCGAAGACTCACGCCCAGTTGATACCTTGTCAGGTTTATCAAGTGAGCAA
GGTCAGTCCGGTGATATGACAATTGAAGAAGATAGTGCTACCCATATTAAATTCTCAAAACGTGATATTGACGGC
AAAGAGTTAGCTGGTGCAACTATGGAGTTGCGTGATTCATCTGGTAAAACTATTAGTACATGGATTTCAGATGGA
CAAGTGAAAGATTTCTACCTGATGCCAGGAAAATATACATTTGTCGAAACCGCAGCACCAGACGGTTATGAGATA
GCAACTGCTATTACCTTTACAGTTAATGAGCAAGGTCAGGTTACTGTAAATGGCAAAGCAACTAAAGGTGACGCT
CATATTGTCATGGTTGATGCTTACAAGCCAACTAAGGGTTCAGGTCAGGTTATTGATATTGAAGAAAAGCTTCCA
GACGAGCAGGGCCATTCTGGCTCAACTACTGAAATAGAAGATAGCAAGTCTTCAGACGTTATCATTGGTGGTCAG
GGGCAGATTGTCGAGACAACAGAGGATACCCAAACTGGCATGCACGGGGATTCTGGTTGTAAAACGGAAGTCGAA
GATACTAAACTAGTACAATCCTTCCACTTTGATAACAAGGAATCAGAAAGTAACTCTGAGATTCCTAAAAAAGAT
AAGCCAAAGAGTAATACTAGTTTACCAGCAACTGGTGAGAAGCAACATAATATGTTCTTTTGGATGGTTACTTCT
TGCTCACTTATTAGTAGTGTTTTTGTAATATCACTAAAAACTAAAAAACGCCTATCATCATGT

SEQ ID NO: 103

MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGAFEIKKNKSQEEYNYEVYDNRNILQDGE
HKLEIKRVDGTGKTYQGFCFQLTKNFPTAQGVSKKLYKKLSSSDEETLKQYASKYTSNRRGDTSGNLKKQIAKVL
TEGYPTNKSDWLNGLTENEKIEVTQDAIWYFTETTVPADRSYTNRNVNSQKMKEVYQKLIDTTDIDKYEDVQFDL
FVPQDTNLQAVISVEPVIESLPWTSLKPIAQKDITAKKIWVDAPKEKPIIYFKLYRQLPGEKEVAVDDAELKQIN
SEGQQEISVTWTNQLVTDEKGMAYIYSVKEVDKNGELLEPKDYIKKEDGLTVTNTYVKPTSGHYDIEVTFGNGHI
DITEDTTPDIVSGENQMKQIEGEDSKPIDEVTENNLIEFGKNTMPGEEDGTNSNKYEEVEDSRPVDTLSGLSSEQ
GQSGDMTIEEDSATHIKFSKRDIDGKELAGATMELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEI
ATAITFTVNEQGQVTVNGKATKGDAHIVMVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKSSDVIIGGQ
GQIVETTEDTQTGMHGDSGCKTEVEDTKLVQSFHFDNKESESNSEIPKKDKPKSNTS*LPATG*EKQHNMFFWMVTS
CSLISSVFVISLKTKKRLSSC

PCT/US2005/027239

Orf 84 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 181** LPATG (shown in italics in SEQ ID NO: 103, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant Orf 84 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in Orf 84. The pilin motif sequence is underlined in SEQ ID NO: 103, below. A conserved lysine (K) residue is also marked in bold, at amino acid residue 270. The pilin sequence, in particular the conserved lysine residue, is thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of Orf 84 include the conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 103

MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGAFEIKKNKSQEEYNYEVYDNRNILQDGE
HKLEIKRVDGTGKTYQGFCFQLTKNFPTAQGVSKKLYKKLSSSDEETLKQYASKYTSNRRGDTSGNLKKQIAKVL
TEGYPTNKSDWLNGLTENEKIEVTQDAIWYFTETTVPADRSYTNRNVNSQKMKEVYQKLIDTTDIDKYEDVQFDL
FVPQDTNLQAVISVEPVIESLPWTSLKPIAQKDITAKKIWVDAPKEKPIIYFKLYRQLPGEKEVAVDDAELKQIN
SEGQQEISVTWTNQLVTDEKGMAYIYSVKEVDKNGELLEPKDYIKKEDGLTVTNTYVKPTSGHYDIEVTFGNGHI
DITEDTTPDIVSGENQMKQIEGEDSKPIDEVTENNLIEFGKNTMPGEEDGTNSNKYEEVEDSRPVDTLSGLSSEQ
GQSGDMTIEEDSATHIKFSKRDIDGKELAGATMELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEI
ATAITFTVNEQGQVTVNGKATKGDAHIVMVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKSSDVIIGGQ
GQIVETTEDTQTGMHGDSGCKTEVEDTKLVQSFHFDNKESESNSEIPKKDKPKSNTSLPATGEKQHNMFFWMVTS
CSLISSVFVISLKTKKRLSSC

An E box containing a conserved glutamic residue has been identified in Orf 84. The E-box motif is underlined in SEQ ID NO: 103, below. The conserved glutamic acid (E), at amino acid residue 516, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of Orf 84. Preferred fragments of Orf 84 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 103

MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGAFEIKKNKSQEEYNYEVYDNRNILQDGE
HKLEIKRVDGTGKTYQGFCFQLTKNFPTAQGVSKKLYKKLSSSDEETLKQYASKYTSNRRGDTSGNLKKQIAKVL
TEGYPTNKSDWLNGLTENEKIEVTQDAIWYFTETTVPADRSYTNRNVNSQKMKEVYQKLIDTTDIDKYEDVQFDL
FVPQDTNLQAVISVEPVIESLPWTSLKPIAQKDITAKKIWVDAPKEKPIIYFKLYRQLPGEKEVAVDDAELKQIN
SEGQQEISVTWTNQLVTDEKGMAYIYSVKEVDKNGELLEPKDYIKKEDGLTVTNTYVKPTSGHYDIEVTFGNGHI
DITEDTTPDIVSGENQMKQIEGEDSKPIDEVTENNLIEFGKNTMPGEEDGTNSNKYEEVEDSRPVDTLSGLSSEQ
GQSGDMTIEEDSATHIKFSKRDIDGKELAGATMELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEI
ATAITFTVNEQGQVTVNGKATKGDAHIVMVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKSSDVIIGGQ
GQIVETTEDTQTGMHGDSGCKTEVEDTKLVQSFHFDNKESESNSEIPKKDKPKSNTSLPATGEKQHNMFFWMVTS
CSLISSVFVISLKTKKRLSSC

Examples of GAS AI-3 sequences from M18 strain isolate MGAS8232 are set forth below.

SpyM18_0125 is a negative transcriptional regulator (Nra). An example of SpyM18_0125 is set forth in SEQ ID NO: 72.

SEQ ID NO: 72

DEMANDE / PATENT

MPYVKKKKDSFLVETYLEQSIRDKSELVLLFKSPTIIFSHVAKQTGLTAVQLKYYCKELDDFFGNNLDITIKKG
KIICCFVKPVKEFYLHQLYDTSTILKLLVFFIKNGTTSQPLIKFSKKYFLSSSSAYRLRESLIKLLREFGLRVSK
NTIVGEEYRIRYLIAMLYSKFGIVIYPLDHLDNQIIYRFLSQSATNLRTSPWLEEPFSFYNMLLALS

SpyM18_0126 is thought to be a collagen binding protein (CBP). An example of SpyM18_0126 is set forth in SEQ ID NO: 73.

SEQ ID NO: 73

MQKRDKTNYGSANNKRRQTTIGLLKVFLTFVALIGIVGFSIRAFGAEEQSTETKKTSVIIRKYAEGDYSKLLEGA
TLKLAQIEGSGFQEQSFESSTSGQKLQLSDGTYILTETKSPQGYEIAEPITFKVTAGKVFIKGKDGQFVENQNKE
VAEPYSVTAYNDFDDSGFINPKTFTPYGKFYYAKNANGTSQVVYCFNVDLHSPPDSLDKGETIDPDFNEGKEIKY
THILGADLFSYANNPRASTNDELLSQVKKVLEKGYRDDSTTYANLTSVEFRAATQLAIYYFTDSVDLDNLADYHG
FGALTTEALNATKEIVAYAEDRANLPNISNLDFYVPNSNKYQSLIGTQYHPESLVDIIRMEDKQAPIIPITHKLT
ISKTVTGTIADKKKEFNFEIHLKSSDGQAISGTYPTNSGELTVTDGKATFTLKDGESLIVEGLPSGYSYEITETG
ASDYEVSVNGKNAPDGKATKASVKEDETITFENRKDL*VPPTG*LTTDGAIYLWLLLLVLLGLWVWLIGRKGLKND

SpyM18_0126 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 184** VPPTG (shown in italics in SEQ ID NO: 73, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyM18_0126 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in SpyM18_0126. The pilin motif sequence is underlined in SEQ ID NO: 73, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 172 and 179. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyM18_0126 include at least one conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 73

MQKRDKTNYGSANNKRRQTTIGLLKVFLTFVALIGIVGFSIRAFGAEEQSTETKKTSVIIRKYAEGDYSKLLEGA
TLKLAQIEGSGFQEQSFESSTSGQKLQLSDGTYILTETKSPQGYEIAEPITFKVTAGKVFIKGKDGQFVENQNKE
VAEPYSVTAYNDFDDSGFINP**K**TFTPYG**K**FYYAKNANGTSQVVYCFNVDLHSPPDSLDKGETIDPDFNEGKEIKY
THILGADLFSYANNPRASTNDELLSQVKKVLEKGYRDDSTTYANLTSVEFRAATQLAIYYFTDSVDLDNLADYHG
FGALTTEALNATKEIVAYAEDRANLPNISNLDFYVPNSNKYQSLIGTQYHPESLVDIIRMEDKQAPIIPITHKLT
ISKTVTGTIADKKKEFNFEIHLKSSDGQAISGTYPTNSGELTVTDGKATFTLKDGESLIVEGLPSGYSYEITETG
ASDYEVSVNGKNAPDGKATKASVKEDETITFENRKDLVPPTGLTTDGAIYLWLLLLVLLGLWVWLIGRKGLKND

Three E boxes containing conserved glutamic residues have been identified in SpyM18_0126. The E-box motifs are underlined in SEQ ID NO: 73, below. The conserved glutamic acid (E) residues, at amino acid residues 112, 257, and 415, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of SpyM18_0126. Preferred fragments of SpyM18_0126 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 73

MQKRDKTNYGSANNKRRQTTIGLLKVFLTFVALIGIVGFSIRAFGAEEQSTETKKTSVIIRKYAEGDYSKLLEGA
TLKLAQIEGSGFQEQSFESSTSGQKLQLSDGTYILT**E**TKSPQGY**E**IAEPITFKVTAGKVFIKGKDGQFVENQNKE

VAEPYSVTAYNDFDDSGETNPKTETPYGKFYTAKNANGTSQVVYCFNVDLHSPPDSLDKGETIDPDFNEGKEIKY
THILGADLFSYANNPRASTNDELLSQVKKVLEKGYRDDSTTYANLTSVEFRAATQLAIYYFTDSVDLDNLADYHG
FGALTTEALNATKEIVAYAEDRANLPNISNLDFYVPNSNKYQSLIGTQYHPESLVDIIRMEDKQAPIIPITHKLT
ISKTVTGTIADKKKEFNFEIHLKSSDGQAISGTYPTNSGELTVTDGKATFTLKDGESLIVEGLPSGYSYEITETG
ASDYEVSVNGKNAPDGKATKASVKEDETITFENRKDLVPPTGLTTDGAIYLWLLLLVLLGLWVWLIGRKGLKND

SpyM18_0127 is a LepA protein. An example of SpyM18_0127 is shown in SEQ ID NO: 74.

SEQ ID NO: 74

MTNYLNRLNENPLFKAFIRLVLKISIIGFLGYILFQYIFGVMIINTNVMSPALSAGDGILYYRLTDRYHINDVVV
YEVDNTLKVGRIVAQAGDEVSFTQEGGLLINGHPPEKEVPYLTYPHSSGPNFPYKVPTGTYFILNDYREERLDSR
YYGALPINQIKGKISTLLRVRGI

SpyM18_0128 is thought to be a fimbrial protein. An example of SypM18_0128 is shown in SEQ ID NO: 75.

SEQ ID NO: 75

MKKNKLLLATAILATALGTASLNQNVKAETAGVIDGSTLVVKKTFPSYTDDKVLMPKADYTFKVEADDNAKGKTK
DGLDIKPGVIDGLENTKTIHYGNSDKTTAKEKSVNFDFANVKFPGVGVYRYTVSEVNGNKAGIAYDSQQWTVDVY
VVNREDGGFEAKYIVSTEGGQSDKKPVLFKNFFDTTSLKVTKKVTGNTGEHQRSFSFTLLLTPNECFEKGQVVNI
LQGGETKKVVIGEEYSFTLKDKESVTLSQLPVGIEYKVTEEDVTKDGYKTSATLKDGDVTDGYNLGDSKTTDKST
DEIVVTNKRDT*QVPTG*VVGTLAPFAVLSIVAIGGVIYITKRKKA

SpyM18_0128 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 140** QVPTG (shown in italics in SEQ ID NO: 75, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyM18_0128 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in SpyM18_0128. The pilin motif sequence is underlined in SEQ ID NO: 75, below. A conserved lysine (K) residue is also marked in bold, at amino acid residue 57. The pilin sequence, in particular the conserved lysine residue, is thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyM18_0128 include the conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 75

MKKNKLLLATAILATALGTASLNQNVKAETAGVIDGSTLVVKKTFPSYTDDKVLMPKADYTFKVEADDNAKGKTK
DGLDIKPGVIDGLENTKTIHYGNSDKTTAKEKSVNFDFANVKFPGVGVYRYTVSEVNGNKAGIAYDSQQWTVDVY
VVNREDGGFEAKYIVSTEGGQSDKKPVLFKNFFDTTSLKVTKKVTGNTGEHQRSFSFTLLLTPNECFEKGQVVNI
LQGGETKKVVIGEEYSFTLKDKESVTLSQLPVGIEYKVTEEDVTKDGYKTSATLKDGDVTDGYNLGDSKTTDKST
DEIVVTNKRDTQVPTGVVGTLAPFAVLSIVAIGGVIYITKRKKA

An E box containing a conserved glutamic residue has been identified in SpyM18_0128. The E-box motif is underlined in SEQ ID NO: 75, below. The conserved glutamic acid (E), at amino acid residue 266, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of SpyM18_0128.

Preferred fragments of SpyM18_0128 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 75

MKKNKLLLATAILATALGTASLNQNVKAETAGVIDGSTLVVKKTFPSYTDDKVLMPKADYTFKVEADDNAKGKTK
DGLDIKPGVIDGLENTKTIHYGNSDKTTAKEKSVNFDFANVKFPGVGVYRYTVSEVNGNKAGIAYDSQQWTVDVY
VVNREDGGFEAKYIVSTEGGQSDKKPVLFKNFFDTTSLKVTKKVTGNTGEHQRSFSFTLLLTPNECFEKGQVVNI
LQGGETKKVVIGEEYSFTLKDKESVTLSQLPVGIEYKVTEEDVTKDGYKTSATLKDGDVTDGYNLGDSKTTDKST
DEIVVTNKRDTQVPTGVVGTLAPFAVLSIVAIGGVIYITKRKKA

SpyM18_0129 is a SrtC2 type sortase. An example of SpyM18_0129 is shown in SEQ ID NO: 76

SEQ ID NO: 76

MISQRMMMTIVQVINKAIDTLILIFCLVVLFLAGFGLWDSYHLYQQADASNFKKFKTAQQQPKFEDLLALNEDVI
GWLNIPGTHMDYPLVQGKTNLEYINKAVDGSVAMSGSLFLDTRNHNDFTDDYSLIYGHHMAGNAMFGEIPKFLKK
DFFNKHNKAIIETKERKKLTVTIFACLKTDAFDQLVFNPNAITNQDQQRQLVDYISKRSKQFKPVKLKHHTKFVA
FSTCENFSTDNRVIVVGTIQE

SpyM18_0130 is referred to as a hypothetical protein. An example of SpyM18_0130 is shown in SEQ ID NO: 77.

SEQ ID NO: 77

MRKYWKMLFSVVMILTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTSFSVALESIDAMKTIDEITIAGSGKAS
FSPLTFTTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKPKRLVKPI
PPRQPDIPKTP*LPLAG*EVKSLLGILSIVLLGLLVLLYVKKLKSRL

SpyM18_0130 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 185** LPLAG (shown in italics in SEQ ID NO: 77, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyM18_0130 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in SpyM18_0130. The pilin motif sequence is underlined in SEQ ID NO: 77, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 144, 159, and 169. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyM18_0130 include at least one conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 77

MRKYWKMLFSVVMILTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTSFSVALESIDAMKTIDEITIAGSGKAS
FSPLTFTTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKPKRLVKPI
PPRQPDIPKTPLPLAGEVKSLLGILSIVLLGLLVLLYVKKLKSRL

An E box containing a conserved glutamic residue has been identified in SpyM18_0130. The E-box motif is underlined in SEQ ID NO: 77, below. The conserved glutamic acid (E), at amino acid residue 134, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is

thought to be important for the formation of oligomeric pilus-like structures of SpyM18_0130. Preferred fragments of SpyM18_0130 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 77

MRKYWKMLFSVVMILTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTSFSVALESIDAMKTIDEITIAGSGKAS
FSPLTFTTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKPKRLVKPI
PPRQPDIPKTPLPLAGEVKSLLGILSIVLLGLLVLLYVKKLKSRL

SpyM18_0131 is referred to as a putative multiple sugar metabolism regulator. An example of SpyM18_0131 is set forth in SEQ ID NO: 78.

SEQ ID NO: 78

MAIFDLKHVQTLHSLSQLPISVMSQDKALIQVYGNDDYLLCYYQFLKHLAIPQAAQDVIFYEGLFEESFMIFPLC
HYIIAIGPFYPYSLNKDYQEQLANNCLKHSSHRSKEELLSYMALVPHFPINNVRNLLIAIDAFFDTQFETTCQQT
IHQLLQHSKQMTADPDIIHRLKHISKASSQLPPVLEHLNHIMDLVKLGNPQLLKQEINRIPLSSITSSSISALRA
EKNLTVIYLTRLLEFSFVENTDVAKHYSLVKYYMALNEEASDLLKVLRIRCAAIIHFSESLTNKSISDKRQMYNS
VLHYVDSHLYSKLKVSDIAKRLYVSESHLRSVFKKYSNVSLQHYILSTKIKEAQLLLKRGIPVGEVAKSLYFYDT
THFHKIFKKYTGISSKDYLAKYRDNI

SpyM18_0132 is a F2 like fibronectic-binding protein. An example of SpyM18_0132 is set forth in SEQ ID NO: 79.

SEQ ID NO: 79

MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGAFEIKKNKSQEEYNYEVYDNRNILQDGE
HKLEIKRVDGTGKTYQGFCFQLTKNFPTAQGVSKKLYKKLSSSDEETLKQYASKYTSNRRGDTSGNLKKQIAKVL
TEGYPTNKSDWLNGLTENEKIEVTQDAIWYFTETTVPADRSYTNRNVNSQKMKEVYQKLIDTTDIDKYEDVQFDL
FVPQDTNLQAVISVEPVIESLPWTSLKPIAQKDITAKKIWVDAPKEKPIIYFKLYRQLPGEKEVAVDDAELKQIN
SEGQQEISVTWTNQLVTDEKGMAYIYSVKEVDKNGELLEPKDYIKKEDGLTVTNTYVKPTSGHYDIEVTFGNGHI
DITEDTTPDIVSGENQMKQIEGEDSKPIDEVTENNLIEFGKNTMPGEEDGTNSNKYEEVEDSRPVDTLSGLSSEQ
GQSGDMTIEEDSATHIKFSKRDIDGKELAGATMELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEI
ATAITFTVNEQGQVTVNGKATKGDAHIVMVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKSSDVIIGGQ
GQIVETTEDTQTGMHGDSGCKTEVEDTKLVQSFHFDNKESESNSEIPKKDKPKSNTS*LPATG*EKQHNMFFWMVTS
CSLISSVFVISLKTKKRLSSC

SpyM18_0132 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 180** LPATG (shown in italics in SEQ ID NO: 79, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyM18_0132 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in SpyM18_0132. The pilin motif sequence is underlined in SEQ ID NO: 79, below. A conserved lysine (K) residue is also marked in bold, at amino acid residue 270. The pilin sequence, in particular the conserved lysine residue, is thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyM18_0132 include the conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 79

MTQKNSYKESFLLSLTGFILGLLLVFLGLSGVSVGHAETRNGANKQGAFEIKKNKSQEEYNYEVYDNRNILQDGE
HKLEIKRVDGTGKTYQGFCFQLTKNFPTAQGVSKKLYKKLSSSDEETLKQYASKYTSNRRGDTSGNLKKQIAKVL
TEGYPTNKSDWLNGLTENEKIEVTQDAIWYFTETTVPADRSYTNRNVNSQKMKEVYQKLIDTTDIDKYEDVQFDL
FVPQDTNLQAVISVEPVIESLPWTSLKPIAQKDITAKKIWVDAPKEKPIIYFKLYRQLPGEKEVAVDDAELKQIN
SEGQQEISVTWTNQLVTDEKGMAYIYSVKEVDKNGELLEPKDYIKKEDGLTVTNTYVKPTSGHYDIEVTFGNGHI
DITEDTTPDIVSGENQMKQIEGEDSKPIDEVTENNLIEFGKNTMPGEEDGTNSNKYEEVEDSRPVDTLSGLSSEQ
GQSGDMTIEEDSATHIKFSKRDIDGKELAGATMELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEI
ATAITFTVNEQGQVTVNGKATKGDAHIVMVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKSSDVIIGGQ
GQIVETTEDTQTGMHGDSGCKTEVEDTKLVQSFHFDNKESESNSEIPKKDKPKSNTSLPATGEKQHNMFFWMVTS
CSLISSVFVISLKTKKRLSSC

An E box containing a conserved glutamic residue has been identified in SpyM18_0132. The E-box motif is underlined in SEQ ID NO: 79, below. The conserved glutamic acid (E), at amino acid residue 516, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of SpyM18_0132. Preferred fragments of SpyM18_0132 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 79

MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGAFEIKKNKSQEEYNYEVYDNRNILQDGE
HKLEIKRVDGTGKTYQGFCFQLTKNFPTAQGVSKKLYKKLSSSDEETLKQYASKYTSNRRGDTSGNLKKQIAKVL
TEGYPTNKSDWLNGLTENEKIEVTQDAIWYFTETTVPADRSYTNRNVNSQKMKEVYQKLIDTTDIDKYEDVQFDL
FVPQDTNLQAVISVEPVIESLPWTSLKPIAQKDITAKKIWVDAPKEKPIIYFKLYRQLPGEKEVAVDDAELKQIN
SEGQQEISVTWTNQLVTDEKGMAYIYSVKEVDKNGELLEPKDYIKKEDGLTVTNTYVKPTSGHYDIEVTFGNGHI
DITEDTTPDIVSGENQMKQIEGEDSKPIDEVTENNLIEFGKNTMPGEEDGTNSNKYEEVEDSRPVDTLSGLSSEQ
GQSGDMTIEEDSATHIKFSKRDIDGKELAGATMELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEI
ATAITFTVNEQGQVTVNGKATKGDAHIVMVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKSSDVIIGGQ
GQIVETTEDTQTGMHGDSGCKTEVEDTKLVQSFHFDNKESESNSEIPKKDKPKSNTSLPATGEKQHNMFFWMVTS
CSLISSVFVISLKTKKRLSSC

Examples of GAS AI-3 sequences from M49 strain isolate 591 are set forth below.

SpyoM01000156 is a negative transcriptional regulator (Nra). An example of SpyoM01000156 is set forth in SEQ ID NO: 243.

SEQ ID NO: 243

MPYVKKKKDSFLVETYLEQSIRDKSELVLLLFKSPTIIFSHVAKQTGLTAVQLKYYCKELDDFFGNNLDI
TIKKGKIICCFVKPVKEFYLHQLYDTSTILKLLVFFIKNGTSSQPLIKFSKKYFLSSSSAYRLRESLIKL
LREFGLRVSKNTIVGEEYRIRYLIAMLYSKFGIVIYPLDHLDNQIIYRFLSQSATNLRTSPWLEEPFSFY
NMLLALSWKRHQFAVSIPQTRIFRQLKKLFIYDCLTRSSRQVIENAFSLTFSQGDLDYLFLIYITTNNSF
ASLQWTPQHIETCCHIFEKNDTFRLLLEPILKRLPQLNHSKQDLIKALMYFSKSFLFNLQHFVIEIPSFS
LPTYTGNSNLYKALKNIVNQWLAQLPGKRHLNEKHLQLFCSHIEQILKNKQPALTVVLISSNFINAKLLT
DTIPRYFSDKGIHFYSFYLLRDDIYQIPSLKPDLVITHSRLIPFVKNDLVKGVTVAEFSFDNPDYSIASI
QNLIYQLKDKKYQDFLNEQLQ

SpyoM01000155 is thought to be a collagen binding protein (CPA). An example of SpyoM01000155 is set forth in SEQ ID NO: 244.

SEQ ID NO: 244

MQKRDKTNYGSANNKRRQTTIGLLKVFLTFVALIGIVGFSIRAFGAEEQSVPNRQSSIQDYPWYGYDSYP
KGYPDYSPLKTYHNLKVNLEGSKDYQAYCFNLTKHFPSKSDSVRSQWYKKLEGTNENFIKLADKPRIEDG
QLQQNILRILYNGYPNNRNGIMKGIDPLNAILVTQNAIWYYTDSAQINPDESFKTEARSNGINDQQLGLM
RKALKELIDPNLGSKYSNKTPSGYRLNVFESHDKTFQNLLSAEYVPDTPPKPGEEPPAKTEKTSVIIRKY
AEGDYSKLLEGATLKLSQIEGSGFQEKDFQSNSLGETVELPNGTYTLTETSSPDGYKIAEPIKFRVENKK
VFIVQKDGSQVENPNKEVAEPYSVEAYNDFMDEEVLSGFTPYGKFYYAKNKDKSSQVVYCFNADLHSPPD
SYDSGETINPDTSTMKEVKYTHTAGSDLFKYALRPRDTNPEDFLKHIKKVIEKGYKKKGDSYNGLTETQF
RAATQLAIYYFTDSADLKTLKTYNNGKGYHGFESMDEKTLAVTKELITYAQNGSAPQLTNLDFFVPNNSK
YQSLIGTEYHPDDLVDVIRMEDKKQEVIPVTHSLTVKKTVVGELGDKTKGFQFELELKDKTGQPIVNTLK

TNNQDLVAKDGKYSFNLKHGDTIRIEGLPTGYSYTLKETEAKDYIVTVDNKVSQEAQSVGKDITEDKKVT
FENRKDL*VPPTG*LTTDGAIYLWLLLLVPLGLLVWLFGRKGLKND

SpyoM01000155 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 184** VPPTG (shown in italics in SEQ ID NO: 244, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyoM01000155 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Two pilin motifs, discussed above, containing conserved lysine (K) residues have also been identified in SpyoM01000155. The pilin motif sequence is underlined in SEQ ID NO: 244, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 71 and 261. The pilin sequences, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyoM01000155 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 244
MQKRDKTNYGSANNKRRQTTIGLLKVFLTFVALIGIVGFSIRAFGAEEQSVPNRQSSIQDYPWYGYDSYP
KGYPDYSPLKTYHNLKVNLEGSKDYQAYCFNLTKHFPSKSDSVRSQWYKKLEGTNENFIKLADKPRIEDG
QLQQNILRILYNGYPNNRNGIMKGIDPLNAILVTQNAIWYYTDSAQINPDESFKTEARSNGINDQQLGLM
RKALKELIDPNLGSKYSNKTPSGYRLNVFESHDKTFQNLLSAEYVPDTPP**K**PGEEPPAKTEKTSVIIRKY
AEGDYSKLLEGATLKLSQIEGSGFQEKDFQSNSLGETVELPNGTYTLTETSSPDGYKIAEPIKFRVENKK
VFIVQKDGSQVENPNKEVAEPYSVEAYNDFMDEEVLSGFTPYGKFYYAKNKDKSSQVVYCFNADLHSPPD
SYDSGETINPDTSTMKEVKYTHTAGSDLFKYALRPRDTNPEDFLKHIKKVIEKGYKKKGDSYNGLTETQF
RAATQLAIYYFTDSADLKTLKTYNNGKGYHGFESMDEKTLAVTKELITYAQNGSAPQLTNLDFFVPNNSK
YQSLIGTEYHPDDLVDVIRMEDKKQEVIPVTHSLTVKKTVVGELGDKTKGFQFELELKDKTGQPIVNTLK
TNNQDLVAKDGKYSFNLKHGDTIRIEGLPTGYSYTLKETEAKDYIVTVDNKVSQEAQSVGKDITEDKKVT
FENRKDLVPPTGLTTDGAIYLWLLLLVPLGLLVWLFGRKGLKND

Two E boxes containing conserved glutamic residues have been identified in SpyoM01000155. The E-box motifs are underlined in SEQ ID NO: 244, below. The conserved glutamic acid (E) residues, at amino acid residues 329 and 668, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of SpyoM01000155. Preferred fragments of SpyoM01000155 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 244
MQKRDKTNYGSANNKRRQTTIGLLKVFLTFVALIGIVGFSIRAFGAEEQSVPNRQSSIQDYPWYGYDSYP
KGYPDYSPLKTYHNLKVNLEGSKDYQAYCFNLTKHFPSKSDSVRSQWYKKLEGTNENFIKLADKPRIEDG
QLQQNILRILYNGYPNNRNGIMKGIDPLNAILVTQNAIWYYTDSAQINPDESFKTEARSNGINDQQLGLM
RKALKELIDPNLGSKYSNKTPSGYRLNVFESHDKTFQNLLSAEYVPDTPPKPGEEPPAKTEKTSVIIRKY
AEGDYSKLLEGATLKLSQIEGSGFQEKDFQSNSLGETVELPNGTYTLT**E**TSSPDGYKIAEPIKFRVENKK
VFIVQKDGSQVENPNKEVAEPYSVEAYNDFMDEEVLSGFTPYGKFYYAKNKDKSSQVVYCFNADLHSPPD
SYDSGETINPDTSTMKEVKYTHTAGSDLFKYALRPRDTNPEDFLKHIKKVIEKGYKKKGDSYNGLTETQF

RAATQLAIYYFTDSADLKTLKTINNGKGYHGFESMDEKTLAVTKELITYAQNGSAPQLTNLDFFVPNNSK
YQSLIGTEYHPDDLVDVIRMEDKKQEVIPVTHSLTVKKTVVGELGDKTKGFQFELELKDKTGQPIVNTLK
TNNQDLVAKDGKYSFNLKHGDTIRIEGLPTGYSYTLKETEAKDYIVTVDNKVSQEAQSVGKDITEDKKVT
FENRKDLVPPTGLTTDGAIYLWLLLLVPLGLLVWLFGRKGLKND

SpyoM01000154 is a LepA protein. An example of SpyoM01000154 is shown in SEQ ID NO: 245.

SEQ ID NO: 245

MTNYLNRLNENSLFKAFIRLVLKISIIGFLGYILFQYVFGVMIINTNDMSPALSAGDGVLYYRLADRSHI
NDVVVYEVDNTLKVGRIAAQAGDEVNFTQEGGLLINGHPPEKEVPYLTYPHSSGPNFPYKVPTGTYFILN
DYREERLDSRYYGALPINQIKGKISTLLRVRGI

SpyoM01000153 is thought to be a fimbrial protein. An example of SpyoM01000153 is shown in SEQ ID NO: 246.

SEQ ID NO: 246

MKKNKLLLATAILATALGMASMSQNIKAETAGVIDGSTLVVKKTFPSYTDDNVLMPKADYSFKVEADDNA
KGKTKDGLDIKPGVIDGLENTKTIRYSNSDKITAKEKSVNFEFANVKFPGVGVYRYTVAEVNGNKAGITY
DSQQWTVDVYVVNKEGGGFEVKYIVSTEVGQSEKKPVLFKNSFDTTSLKIEKQVTGNTGEHQRLFSFTLL
LTPNECFEKGQVVNILQGGETKKVVIGEEYSFTLKDKESVTLSQLPVGIEYKLTEEDVTKDGYKTSATLK
DGEQSSTYELGKDHKTDKSADEIVVTNKRDT*QVPTG*VVGTLAPFAVLSIVAIGGVIYITKRKKA

SpyoM01000153 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 140** QVPTG (shown in italics in SEQ ID NO: 246, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyoM01000153 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in SpyoM01000153. The pilin motif sequence is underlined in SEQ ID NO: 246, below. A conserved lysine (K) residue is also marked in bold, at amino acid residue 57. The pilin sequence, in particular the conserved lysine residue, is thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyoM01000153 include the conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 246

MKKNKLLLATAILATALGMASMSQNIKAETAGVIDGSTLVVKKTFPSYTDDNVLMPKADYSFKVEADDNA
KGKTKDGLDIKPGVIDGLENTKTIRYSNSDKITAKEKSVNFEFANVKFPGVGVYRYTVAEVNGNKAGITY
DSQQWTVDVYVVNKEGGGFEVKYIVSTEVGQSEKKPVLFKNSFDTTSLKIEKQVTGNTGEHQRLFSFTLL
LTPNECFEKGQVVNILQGGETKKVVIGEEYSFTLKDKESVTLSQLPVGIEYKLTEEDVTKDGYKTSATLK
DGEQSSTYELGKDHKTDKSADEIVVTNKRDTQVPTGVVGTLAPFAVLSIVAIGGVIYITKRKKA

An E box containing a conserved glutamic residue has been identified in SpyoM01000153. The E-box motif is underlined in SEQ ID NO: 246, below. The conserved glutamic acid (E), at amino acid residue 265, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of SpyoM01000153. Preferred fragments of SpyoM01000153 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

PCT/US05/27239

SEQ ID NO: 246

MKKNKLLLATAILATALGMASMSQNIKAETAGVIDGSTLVVKKTFPSYTDDNVLMPKADYSFKVEADDNA
KGKTKDGLDIKPGVIDGLENTKTIRYSNSDKITAKEKSVNFEFANVKFPGVGVYRYTVAEVNGNKAGITY
DSQQWTVDVYVVNKEGGGFEVKYIVSTEVGQSEKKPVLFKNSFDTTSLKIEKQVTGNTGEHQRLFSFTLL
LTPNECFEKGQVVNILQGGETKKVVIGEEYSFTLKDKESVTLSQLPVGIEYKLTEEDVTKDGYKTSATLK
DGEQSSTYELGKDHKTDKSADEIVVTNKRDTQVPTGVVGTLAPFAVLSIVAIGGVIYITKRKKA

SpyoM01000152 is a SrtC2 type sortase. An example of SpyoM01000152 is shown in SEQ ID NO: 247

SEQ ID NO: 247

MMMTIVQVINKAIDTLILIFCLVVLFLAGFGLWDSYHLYQQADASNFKKFKTAQQQPKFEDLLALNEDVI
GWLNIPGTHIDYPLVQGKTNLEYINKAVDGSVAMSGSLFLDTRNHNDFTDDYSLIYGHHMAGNAMFGEIP
KFLKKNFFNKHNKAIIETKERKKLTVTIFACLKTDAFDQLVFNPNAITNQDQQRQLVDYISKRSKQFKPV
KLKHHTKFVAFSTCENFSTDNRVIVVGTIQE

SpyoM01000151 is referred to as a hypothetical protein. An example of SpyoM01000151 is shown in SEQ ID NO: 248.

SEQ ID NO: 248

MLFSVVMMLTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTPFSIALESIDAMKTIEEITIAGSGKASF
SPLTFTTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKPKRL
VKPIPPRQPDIPKTP*LPLAG*EVKSLLGILSIVLLGLLVLLYVKKLKSRL

SpyoM01000151 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 185** LPLAG (shown in italics in SEQ ID NO: 248, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyoM01000151 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in SpyoM01000151. The pilin motif sequence is underlined in SEQ ID NO: 248, below. Conserved lysine (K) residues are also marked in bold, at amino acid residue 138. The pilin sequence, in particular the conserved lysine residue, is thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyoM01000151 include the conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 248

MLFSVVMMLTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTPFSIALESIDAMKTIEEITIAGSGKASF
SPLTFTTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKPKRL
VKPIPPRQPDIPKTPLPLAGEVKSLLGILSIVLLGLLVLLYVKKLKSRL

Two E boxes containing conserved glutamic residues have been identified in SpyoM01000151. The E-box motifs are underlined in SEQ ID NO: 248, below. The conserved glutamic acid (E) residues, at amino acid residues 58 and 128, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of

PCT/US05/27239

oligomeric pilus-like structures of SpyoM01000151. Preferred fragments of SpyoM01000151 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 248

MLFSVVMMLTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTPFSIALESIDAMKTIEEITIAGSGKASF
SPLTFTTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKPKRL
VKPIPPRQPDIPKTPLPLAGEVKSLLGILSIVLLGLLVLLYVKKLKSRL

SpyoM01000150 is referred to as a putative MsmRL. An example of SpyoM01000150 is set forth in SEQ ID NO: 249.

SEQ ID NO: 249

MVIFDLKHVQTLHSLSQLPISVMSQDKALIQVYGNDDYLLCYYQFLKHLAIPQAAQDVIFYEGLFEESFM
IFPLCHYIIAIGPFYPYSLNKDYQEQLANNFLKHSSHRSKEELLSYMALVPHFPINNVRNLLIAIDAFFD
TQFETTCQQTIHQLLQHSKQMTADPDIIHRLKHISKASSQLPPVLEHLNHIMDLVKLGNPQLLKQEINRI
PLSSITSSSISALRAEKNLTVIYLTRLLEFSFVENTDVAKHYSLVKYYMALNEEASDLLKVLRIRCAAII
HFSESLTNKSISDKRQMYNSVLHYVDSHLYSKLKVSDIAKRLYVSESHLRSVFKKYSNVSLQHYILSTKI
KEAQLLLKRGIPVGEVAKSLYFYDTTHFHKIFKKYTGISSKDYLAKYRDNI

SpyoM01000149 is a F2 like fibronectin-binding protein. An example of SpyoM01000149 is set forth in SEQ ID NO: 250.

SEQ ID NO: 250

MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGYFEIKKVDQNNKPLSGATFSLTP
KDGKGKPVQTFTSSEEGIIDAQNLQPGTYTLKEETAPDGYDKTSRTWTVTVYENGYTKLVENPYNGEIIS
KAGSKDVSSSLQLENPKMSVVSKYGEQEKTSNSADFYRNHAAYFKMSFELKQKDKSETINPGDTFVLQLD
RRLNPKGISQDIPKIIYDSENSPLAIGKYDAKTHQLTYTFTNYIAGLDKVQLSAELSLFLENKEVLENTN
ISDFKSTIGGQEITYKGTVNVLYGNESTKESNYITNGLSNVGGSIESYNTETGEFVWYVYVNPNRTNIPY
AVLNLWGFAKRTAQGENDNSSVSSAQLTGYDIYEVPHNYRLPTSYGVDISRLNLRKDLEAKLPQGSTQGA
NKRLRIDFGENLQGKAFVVKVTGKADQSGKELIVQSHLSSFNNWGSYKTLRPNSHVSFTNEIALSPSKGS
GSGTSEFTKPAITVANLKRVAQLRFKKVSTDNVPLPEAAFELRSSNGNSQKLEASSNTQGEIHFKDLTSG
TYDLYETKAPKGYQQVTEKLATVTVDTTKPAEQMVKWEKPHSFVKVEANKEVTIVNHKETLTFSGKKIWE
NDRPDQRPAKIQVQLLQNGQKMPNQIQEVTKDNDWSYHFKDLPKYDAKNQEYKYSVEEVKVPDGYKVSYL
GNDIFNTRETEFVFEQNNFNLEFGNAEIKGQSGSKIIDEDTLTSFKGKKIWKNDTAENRPQAIQVQLYAD
GVAVEGQTKFISGSGNEWSFEFKNLKKYNGTGNDIIYSVKEVTVPTGYDVTYSANDIINTKREVITQQGP
NLEIEETLPLESGASGGTTTVEDSRSVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRDIDGKELAGATM
ELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEIATAITFTVNEQGQVTVNGKATKGDAHIV
MVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKPSDVIIGGQGEVVDTTEDTQSGMTGHSGSTTE
IEDSKSSDVIIGGQGQVVETTEDTQTGMHGDSGCKTEVEDTKLVQFFHFDNKEPESNSEIPKKDKPKSNT
*SLPATG*EKQHNKFFWMVTSCSLISSVFVISLKSKKRLLSC

SpyoM01000149 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 180** LPATG (shown in italics in SEQ ID NO: 250, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant SpyoM01000149 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Two pilin motifs, discussed above, containing conserved lysine (K) residues have also been identified in SpyoM01000149. The pilin motif sequences are underlined in SEQ ID NO: 250, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 157 and 163, and 216 and 224. The pilin sequences, in particular the conserved lysine residues, are thought to be important

for the formation of oligomeric, pilus-like structures. Preferred fragments of SpyoM01000149 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 250

MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGYFEIKKVDQNNKPLSGATFSLTP
KDGKGKPVQTFTSSEEGIIDAQNLQPGTYTLKEETAPDGYDKTSRTWTVTVYENGYTKLVENPYNGEIIS
KAGSKDVSSSLQLENPKMSVVSKYGEQEKTSNSADFYRNHAAYFKMSFELKQKDKSETINPGDTFVLQLD
RRLNPKGISQDIPKIIYDSENSPLAIGKYDAKTHQLTYTFTNYIAGLDKVQLSAELSLFLENKEVLENTN
ISDFKSTIGGQEITYKGTVNVLYGNESTKESNYITNGLSNVGGSIESYNTETGEFVWYVYVNPNRTNIPY
AVLNLWGFAKRTAQGENDNSSVSSAQLTGYDIYEVPHNYRLPTSYGVDISRLNLRKDLEAKLPQGSTQGA
NKRLRIDFGENLQGKAFVVKVTGKADQSGKELIVQSHLSSFNNWGSYKTLRPNSHVSFTNEIALSPSKGS
GSGTSEFTKPAITVANLKRVAQLRFKKVSTDNVPLPEAAFELRSSNGNSQKLEASSNTQGEIHFKDLTSG
TYDLYETKAPKGYQQVTEKLATVTVDTTKPAEQMVKWEKPHSFVKVEANKEVTIVNHKETLTFSGKKIWE
NDRPDQRPAKIQVQLLQNGQKMPNQIQEVTKDNDWSYHFKDLPKYDAKNQEYKYSVEEVKVPDGYKVSYL
GNDIFNTRETEFVFEQNNFNLEFGNAEIKGQSGSKIIDEDTLTSFKGKKIWKNDTAENRPQAIQVQLYAD
GVAVEGQTKFISGSGNEWSFEFKNLKKYNGTGNDIIYSVKEVTVPTGYDVTYSANDIINTKREVITQQGP
NLEIEETLPLESGASGGTTTVEDSRSVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRDIDGKELAGATM
ELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEIATAITFTVNEQGQVTVNGKATKGDAHIV
MVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKPSDVIIGGQGEVVDTTEDTQSGMTGHSGSTTE
IEDSKSSDVIIGGQGQVVETTEDTQTGMHGDSGCKTEVEDTKLVQFFHFDNKEPESNSEIPKKDKPKSNT
SLPATGEKQHNKFFWMVTSCSLISSVFVISLKSKKRLLSC

Two E boxes containing conserved glutamic residues have been identified in SpyoM01000149. The E-box motifs are underlined in SEQ ID NO: 250, below. The conserved glutamic acid (E) residues, at amino acid residues 329 and 668, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of SpyoM01000149. Preferred fragments of SpyoM01000149 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 250

MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGYFEIKKVDQNNKPLSGATFSLTP
KDGKGKPVQTFTSSEEGIIDAQNLQPGTYTLKEETAPDGYDKTSRTWTVTVYENGYTKLVENPYNGEIIS
KAGSKDVSSSLQLENPKMSVVSKYGEQEKTSNSADFYRNHAAYFKMSFELKQKDKSETINPGDTFVLQLD
RRLNPKGISQDIPKIIYDSENSPLAIGKYDAKTHQLTYTFTNYIAGLDKVQLSAELSLFLENKEVLENTN
ISDFKSTIGGQEITYKGTVNVLYGNESTKESNYITNGLSNVGGSIESYNTETGEFVWYVYVNPNRTNIPY
AVLNLWGFAKRTAQGENDNSSVSSAQLTGYDIYEVPHNYRLPTSYGVDISRLNLRKDLEAKLPQGSTQGA
NKRLRIDFGENLQGKAFVVKVTGKADQSGKELIVQSHLSSFNNWGSYKTLRPNSHVSFTNEIALSPSKGS
GSGTSEFTKPAITVANLKRVAQLRFKKVSTDNVPLPEAAFELRSSNGNSQKLEASSNTQGEIHFKDLTSG
TYDLY**E**TKAPKGYQQVTEKLATVTVDTTKPAEQMVKWEKPHSFVKVEANKEVTIVNHKETLTFSGKKIWE
NDRPDQRPAKIQVQLLQNGQKMPNQIQEVTKDNDWSYHFKDLPKYDAKNQEYKYSVEEVKVPDGYKVSYL
GNDIFNTRETEFVFEQNNFNLEFGNAEIKGQSGSKIIDEDTLTSFKGKKIWKNDTAENRPQAIQVQLYAD
GVAVEGQTKFISGSGNEWSFEFKNLKKYNGTGNDIIYSVKEVTVPTGYDVTYSANDIINTKREVITQQGP
NLEIEETLPLESGASGGTTTVEDSRSVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRDIDGKELAGATM
ELRDSSGKTISTWISDGQVKDFYLMPGKYTFV**E**TAAPDGYEIATAITFTVNEQGQVTVNGKATKGDAHIV
MVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKPSDVIIGGQGEVVDTTEDTQSGMTGHSGSTTE
IEDSKSSDVIIGGQGQVVETTEDTQTGMHGDSGCKTEVEDTKLVQFFHFDNKEPESNSEIPKKDKPKSNT
SLPATGEKQHNKFFWMVTSCSLISSVFVISLKSKKRLLSC

As discussed above, applicants have also determined the nucleotide and encoded amino acid sequence of fimbrial structural subunits in several other GAS AI-3 strains of bacteria. Examples of sequences of these fimbrial structural subunits are set forth below.

M3 strain isolate ISS 3040 is a GAS AI-3 strain of bacteria. ISS3040_fimbrial is thought to be a fimbrial structural subunit of M3 strain isolate ISS 3040. An example of a nucleotide sequence

encoding the ISS3040_fimbrial protein (SEQ ID NO: 263) and an ISS3040_fimbrial protein amino acid sequence (SEQ ID NO: 264) are set forth below.

SEQ ID NO: 263

```
gagacggcaggagtgtccgaaaatgcaaaattaatagtaaaaaagacatttgactcttat
acagacaatgaagttttaatgccaaaagctgattatacttttaaagtagaggcagatagt
acagctagtggcaaaacgaaagacggtttagagattaagccaggtattgttaatggttta
acagaacagattatcagctatactaatactgataaaccagatagtaaagttaaaagtaca
gagtttgatttttcaaaagtagtattccctggtattggtgtttaccgctatactgtttca
gaaaaacaaggtgatgttgaaggaattacctacgatactaagaagtggacagtagatgtt
tatgttggaaacaaagaaggtggtggttttgaacctaagtttattgtatctaaggaacaa
ggaacagacgtcaaaaaaccagttaattttaacaactcgtttgcaactacttcgttaaaa
gttaagaagaatgtatcggggaatactggagaattgcaaaaagaatttgactttacattg
acgcttaatgaaagcacgaatttaaaaaagatcaaattgtttctttacaaaaaggaaac
gagaaatttgaagttaagattggtactccctacaagtttaaactcaaaaatggggaatct
attcaactagacaagttaccagttggtattacttataaagtcaatgaaatggaagctaat
aaagatgggtataaaacaacagcatccttgaaagagggagatggtcaatctaaaatgtat
caattggatatggaacaaaaaacagacgaatctgctgacgaaatcgttgtcacaaataag
cgtgacactcaagttccaactggtgttgtaggcacccttgctccatttgcagttcttagc
```

SEQ ID NO: 264

```
ETAGVSENAKLIVKKTFDSYTDNEVLMPKADYTFKVEADSTASG
KTKDGLEIKPGIVNGLTEQIISYTNTDKPDSKVKSTEFDFSKVVFPGIGVYRYTVSEK
QGDVEGITYDTKKWTVDVYVGNKEGGGFEPKFIVSKEQGTDVKKPVNFNNSFATTSLK
VKKNVSGNTGELQKEFDFTLTLNESTNFKKDQIVSLQKGNEKFEVKIGTPYKFKLKNG
ESIQLDKLPVGITYKVNEMEANKDGYKTTASLKEGDGQSKMYQLDMEQKTDESADEIV
VTNKRDTQVPTGVVGTLAPFAVLS
```

M44 strain isolate ISS 3776 is a GAS AI-3 strain of bacteria. ISS3776_fimbrial is thought to be a fimbrial structural subunit of M44 isolate ISS 3776. An example of a nucleotide sequence encoding the ISS3776_fimbrial protein (SEQ ID NO: 253) and an ISS3776_fimbrial protein amino acid sequence (SEQ ID NO: 254) are set forth below.

SEQ ID NO: 253

```
ttggagagagaaaaaatgaaaaaaaacaaattattacttgctactgcaatcttagcaact
gctttaggaacagcttctttaaatcaaaacgtaaaagctgagacggcaggggttgtaaca
ggaaaatcactacaagttacaaagacaatgacttatgatgatgaagaggtgttaatgccc
gaaaccgcctttacttttactatagagcctgatatgactgcaagtggaaaagaaggcagc
ctagatattaaaaatggaattgtagaaggcttagacaaacaagtaacagtaaaatataag
aatacagataaaacatctcaaaaaactaaaatagcacaatttgatttttctaaggttaaa
tttccagctataggtgtttaccgctatatggtttcagagaaaaacgataaaaaagacgga
attacgtacgatgataaaaagtggactgtagatgtttatgttgggaataaggccaataac
gaagaaggtttcgaagttctatatattgtatcaaaagaaggtacttctagtactaaaaaa
ccaattgaatttacaaactctattaaaactacttccttaaaaattgaaaaacaaataact
ggcaatgcaggagatcgtaaaaaatcattcaacttcacattaacattacaaccaagtgaa
tattataaaactggatcagttgtgaaaatcgaacaggatggaagtaaaaaagatgtgacg
ataggaacgccttacaaatttactttgggacacggtaagagtgtcatgttatcgaaatta
ccaattggtatcaattactatcttagtgaagacgaagcgaataaagacggctacactaca
acggcaacattaaaagaacaaggcaaagaaaagagttccgatttcactttgagtactcaa
aaccagaaaacagacgaatctgctgacgaaatcgttgtcacaaataagcgtgacactcaa
gttccaactggtgttgtagggacccttgctccatttgcagttcttagcattgtggctatt
ggtggagttatctatattacaaaacgtaaaaaagcttaa
```

SEQ ID NO: 254

```
MEREKMKKNKLLLATAILATALGTASLNQNVKAETAGVVTGKSL
QVTKTMTYDDEEVLMPETAFTFTIEPDMTASGKEGSLDIKNGIVEGLDKQVTVKYKNT
DKTSQKTKIAQFDFSKVKFPAIGVYRYMVSEKNDKKDGITYDDKKWTVDVYVGNKANN
EEGFEVLYIVSKEGTSSTKKPIEFTNSIKTTSLKIEKQITGNAGDRKKSFNFTLTLQP
SEYYKTGSVVKIEQDGSKKDVTIGTPYKFTLGHGKSVMLSKLPIGINYYLSEDEANKD
```

GYTITATLKEQGKEKSSDFTLSTQNQKTDESADEIVVTNKRDTQVPTGVVGTLAPFAV
LSIVAIGGVIYITKRKKA

M77 strain isolate ISS4959 is a GAS AI-3 strain of bacteria. ISS4959_fimbrial is thought to be a fimbrial structural subunit of M77 strain ISS 4959. An example of a nucleotide sequence encoding the ISS4959_fimbrial protein (SEQ ID NO: 271) and an ISS4959_fimbrial protein amino acid sequence (SEQ ID NO: 272) are set forth below.

SEQ ID NO: 271

gtaacagtaaaatataagaatacagataaaacatctcaaaaaactaaaatagcacaattt
gatttttctaaggttaaatttccagctataggtgtttaccgctatatggtttcagagaaa
aacgataaaaaagacggaattacgtacgatgataaaaagtggacngtagatgtttatgtt
gggaataaggccaataacgaagaaggtttcgaagttctatatattgtatcaaaagaaggt
acttctagtnctaaaaaaccaattgaatttacaaactctattaaaactacttccttaaaa
attgaaaaacaaataactggcaatgcaggagatcgtaaaaaatcattcaacttcacattn
acattacanccaagtgaatattataaaactggatcagttgtgaaaatcgaacaggatgga
agtaaaaaagatgtgacgataggaacgccttacaaatttactttgggacacggtaagagt
gtcatgttatcgaaattnccaattggtatcaattactatcttagtgaagacgaagcgaat
aaagacggntacactacancggcaacattaaaagaacaaggcaaagaaaagagttccgat
ttcactttgagtactcaaaaccagaaaacagacgaatctgctg

SEQ ID NO: 272

VTVKYKNTDKTSQKTKIAQFDFSKVKFPAIGVYRYMVSEKNDKK
DGITYDDKKWTVDVYVGNKANNEEGFEVLYIVSKEGTSSXKKPIEFTNSIKTTSLKIE
KQITGNAGDRKKSFNFTXTLXPSEYYKTGSVVKIEQDGSKKDVTIGTPYKFTLGHGKS
VMLSKXPIGINYYLSEDEANKDGYTTXATLKEQGKEKSSDFTLSTQNQKTDESA

Examples of GAS AI-4 sequences from M12 strain isolate A735 are set forth below.

19224133 is thought to be a RofA regulatory protein. An example of a nucleotide sequence encoding the RofA regulatory protein (SEQ ID NO: 104) and a RofA regulatory protein amino acid sequence (SEQ ID NO: 105) are set forth below.

SEQ ID NO: 104

ATGACCATCCAAAAAAGGATGATATCTTGCCAATTTACACATCCTTCTAAAGAAACTTATCTTTACCAACTCTAT
GCATCATCTAATGTCTTACAATTACTAGCGTTTTTAATAAAAAATGGTTCCCACTCTCGTCCCCTTACGGATTTT
GCAAGAAGTCATTTTTTATCAAACTCCTCAGCTTATCGGATGCGCGAAGCATTGATTCCTTTATTAAGAAACTTT
GAATTAAAACTCTCTAAGAACAAGATTGTCGGTGAGGAATATCGTATCCGTTACCTCATCGCTCTGCTATATAGT
AAGTTTGGCATTAAAGTTTATGACTTGACGCAGCAAGACAAAAACATTATTCATAGCTTTTTATCCCATAGTTCC
ACCCACCTTAAAACTTCTCCTTGGTTATCGGAATCGTTTTCTTTCTATGACATTTTATTAGCTTTATCGTGGAAG
CGGCATCAATTTTCGGTAACTATTCCCCAAACCAGAATTTTTCAACAATTAAAAAAACTTTTTGTCTACGATTCT
TTGAAAAAAAGTAGCCGTGATATTATCGAAACTTACTGCCAACTAAACTTTTCAGCAGGAGATTTGGACTACCTC
TATTTAATTTATATCACCGCTAATAATTCTTTTGCGAGCTTACAATGGACACCTGAGCATATCAGACAATGTTGT
CAACTTTTTGAAGAAAATGATACTTTTCGCCTGCTTTTAAATCCTATCATCACTCTTTTACCTAACCTAAAAGAG
CAAAAGGCTAGTTTAGTAAAAGCTCTTATGTTTTTTTTCAAAATCATTCTTGTTTAATCTGCAACATTTTATTCCT
GAGACCAACTTATTCGTTTCTCCGTACTATAAAGGAAACCAAAAACTCTATACGTCCTTAAAGTTAATTGTCGAA
GAGTGGATGGCCAAACTTCCTGGTAAGCGTTACTTGAACCATAAGCATTTTCATCTTTTTTGCCACTATGTCGAG
CAAATTCTAAGAAATATCCAACCTCCTTTAGTTGTTGTTTTCGTAGCCAGTAATTTTATCAATGCTCATCTCCTA
ACAGATTCTTTCCCAAGGTATTTCTCGGATAAAAGCATTGATTTTCATTCCTATTATCTATTGCAAGATAATGTT
TATCAAATTCCTGATTTAAAGCCAGATTTGGTCATCACTCACAGTCAACTGATTCCTTTTGTTCACCATGAACTT
ACAAAAGGAATTGCTGTTGCTGAAATATCTTTTGATGAATCGATTCTGTCTATCCAAGAATTGATGTATCAAGTT
AAAGAGGAAAAATTCCAAGCTGATTTAACCAAACAATTAACATAA

SEQ ID NO: 105

MTIQKRMISCQFTHPSKETYLYQLYASSNVLQLLAFLIKNGSHSRPLTDFARSHFLSNSSAYRMREALIPLLRNF
ELKLSKNKIVGEEYRIRYLIALLYSKFGIKVYDLTQQDKNIIHSFLSHSSTHLKTSPWLSESFSFYDILLALSWK
RHQFSVTIPQTRIFQQLKKLFVYDSLKKSSRDIIETYCQLNFSAGDLDYLYLIYITANNSFASLQWTPEHIRQCC
QLFEENDTFRLLLNPIITLLPNLKEQKASLVKALMFFSKSFLFNLQHFIPETNLFVSPYYKGNQKLYTSLKLIVE

EWMAKLPGKRYLNHKHFHLFCHYVEQILRNIQPPLVVVFVASNFINAHLLTDSFPRYFSDKSIDFHSYYLLQDNV
YQIPDLKPDLVITHSQLIPFVHHELTKGIAVAEISFDESILSIQELMYQVKEEKFQADLTKQLT

19224134 is thought to be a protein F fibronectin binding protein. An example of a nucleotide sequence encoding the protein F fibronectin binding protein (SEQ ID NO: 106) and a protein F fibronectin binding protein amino acid sequence (SEQ ID NO: 107) are set forth below.

SEQ ID NO: 106

ATGGTAAGCTCATATATGTTTGCGAGAGGAGAGAAAATGAATAACAAAATGTTTTTGAACAAAGAAGCCGGTTTT
TTGGTACACACAAAAAGAAAAAGGCGATTTGCTGTCACTTTAGTGGGAGTCTTTTTTCTGCTTTTGGCATGTGCG
GGTGCTATCGGTTTTGGTCAAGTAGCCTATGCTGCGGATGAGAAGACTGTGCCGAATTTTAAAAGCCCAGATCCA
GATTATCCCTGGTATGGTTATGATTCGTATAGAGGAATATTTGCAAGATATCACAATTTAAAAGTAAATCTAAAA
GGAAGTAAGGAGTATCAAGCGTATTGTTTTAACCTAACAAAATACTTTCCTCGCCCCACTTATAGTACTACAAAT
AATTTTTACAAGAAAATTGATGGGAGTGGATCAGCGTTCAAATCTTATGCAGCGAATCCTAGGGTTTTAGATGAG
AATTTAGATAAATTAGAAAAAAATATACTGAATGTAATTTATAATGGATATAAAAGTAATGCAAATGGTTTTATG
AATGGTATAGAAGATCTTAATGCTATACTAGTAACTCAAAACGCTATTTGGTACTATTCAGATAGTGCTCCATTA
AATGATGTTAATAAAATGTGGGAAAGAGAGGTTCGGAATGGGGAGATTAGTGAGTCACAAGTTACTTTAATGCGT
GAGGCATTGAAAAAACTAATTGATCCCAATTTAGAAGCTACTGCAGCTAATAAAATCCCATCAGGATATCGTTTA
AATATCTTTAAGTCTGAAAATGAAGATTACCAAAATCTTTTAAGTGCTGAATATGTACCTGATGATCCCCCTAAA
CCTGGTGATACGTCAGAACATAATCCTAAAACTCCCGAGTTGGATGGCACTCCAATTCCCGAGGACCCAAAACGT
CCAGATGAGAGTTCAGAACCTGCGCTTCCCCCATTAATGCCAGAGCTAGATGGTGAAGAAGTCCCAGAAGTTCCA
AGCGAGAGCTTAGAACCTGCGCTTCCCCCATTGATGCCAGAGCTAGATGGTGAAGAAGTCCCAGAAGTTCCAAGC
GAGAGCTTAGAACCTGCGCTTCCCCCATTGATGCCAGAGCTAGATGGTGAAGAAGTCCCAGAAGTTCCAAGCGAG
AGCTTAGAACCTGCGCTTCCCCCATTAATGCCAGAGCTAGATGGTGAAGAAGTCCCAGAAGTTCCAAGCGAGAGC
TTAGAACCTGCGCTTCCCCCATTGATGCCAGAGTTAGATGGTGAAGAAGTCCCTGAAAAACCTAGTGTTGACTTA
CCTATTGAAGTTCCTCGTTATGAGTTTAACAATAAAGACCAGTCACCTCTAGCGGGTGAGTCTGGTGAGACGGAG
TATATTACCGAAGTCTATGGAAATCAACAGAACCCTGTTGATATTGATAAAAAACTTCCGAATGAAACAGGTTTT
TCAGGAAATATGGTTGAGACAGAAGATACGAAAGAGCCAGAAGTGTTGATGGGAGGTCAAAGTGAGTCTGTTGAA
TTTACTAAAGACACTCAAACAGGCATGAGTGGTCAAACAACTCCTCAGGTTGAGACAGAAGATACGAAAGAGCCA
GAAGTGTTGATGGGAGGTCAAAGTGAGTCTGTTGAATTTACTAAAGACACTCAAACAGGCATGAGTGGTCAAACA
ACTCCTCAGGTTGAGACAGAAGATACGAAAGAGCCAGGAGTGTTGATGGGAGGCCAAAGTGAGTCTGTTGAATTT
ACTAAAGACACTCAAACAGGCATGAGTGGTCAAACAACTCCTCAGGTTGAGACAGAAGACACGAAAGAGCCAGGA
GTGTTGATGGGAGGTCAAAGTGAGTCTGTTGAATTTACTAAAGACACTCAAACAGGCATGAGCGGTTTCAGTGAA
ACAGTGACCATTGTTGAAGATACGCGTCCGAAGTTAGTGTTCCATTTTGACAATAATGAGCCCAAAGTGGAAGAG
AATCGGGAAAAGCCTACAAAAAATATAACACCTATCCTTCCTGCAACAGGAGATATTGAGAATGTTTTGGCCTTT
CTTGGAATCCTTATTTTGTCAGTACTTTCTATTTTTAGCCTTTTAAAAAACAAACAAAACAATAAAGTCTGA

SEQ ID NO: 107

MVSSYMFARGEKMNNKMFLNKEAGFLVHTKRKRRFAVTLVGVFFLLLACAGAIGFGQVAYAADEKTVPNFKSPDP
DYPWYGYDSYRGIFARYHNLKVNLKGSKEYQAYCFNLTKYFPRPTYSTTNNFYKKIDGSGSAFKSYAANPRVLDE
NLDKLEKNILNVIYNGYKSNANGFMNGIEDLNAILVTQNAIWYYSDSAPLNDVNKMWEREVRNGEISESQVTLMR
EALKKLIDPNLEATAANKIPSGYRLNIFKSENEDYQNLLSAEYVPDDPPKPGDTSEHNPKTPELDGTPIPEDPKR
PDESSEPALPPLMPELDGEEVPEVPSESLEPALPPLMPELDGEEVPEVPSESLEPALPPLMPELDGEEVPEVPSE
SLEPALPPLMPELDGEEVPEVPSESLEPALPPLMPELDGEEVPEKPSVDLPIEVPRYEFNNKDQSPLAGESGETE
YITEVYGNQQNPVDIDKKLPNETGFSGNMVETEDTKEPEVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDTKEP
EVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDTKEPGVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDTKEPG
VLMGGQSESVEFTKDTQTGMSGFSETVTIVEDTRPKLVFHFDNNEPKVEENREKPTKNITPI*LPATG*DIENVLAF
LGILILSVLSIFSLLKNKQNNKV

19224134 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 181** LPATG (shown in italics in SEQ ID NO: 107, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant 19224134 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular

domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in 19224134. The pilin motif sequence is underlined in SEQ ID NO: 107, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 275, 285, and 299. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of 19224134 include at least one conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 107

MVSSYMFARGEKMNNKMFLNKEAGFLVHTKRKRRFAVTLVGVFFLLLACAGAIGFGQVAYAADEKTVPNFKSPDP
DYPWYGYDSYRGIFARYHNLKVNLKGSKEYQAYCFNLTKYFPRPTYSTTNNFYKKIDGSGSAFKSYAANPRVLDE
NLDKLEKNILNVIYNGYKSNANGFMNGIEDLNAILVTQNAIWYYSDSAPLNDVNKMWEREVRNGEISESQVTLMR
EALKKLIDPNLEATAANKIPSGYRLNIFKSENEDYQNLLSAEYVPDDPP**K**PGDTSEHNP**K**TPELDGTPIPEDP**K**R
PDESSEPALPPLMPELDGEEVPEVPSESLEPALPPLMPELDGEEVPEVPSESLEPALPPLMPELDGEEVPEVPSE
SLEPALPPLMPELDGEEVPEVPSESLEPALPPLMPELDGEEVPEKPSVDLPIEVPRYEFNNKDQSPLAGESGETE
YITEVYGNQQNPVDIDKKLPNETGFSGNMVETEDTKEPEVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDTKEP
EVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDTKEPGVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDTKEPG
VLMGGQSESVEFTKDTQTGMSGFSETVTIVEDTRPKLVFHFDNNEPKVEENREKPTKNITPILPATGDIENVLAF
LGILILSVLSIFSLLKNKQNNKV

Two E boxes containing conserved glutamic residues have been identified in 19224134. The E-box motifs are underlined in SEQ ID NO: 107, below. The conserved glutamic acid (E) residues, at amino acid residues 487 and 524, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of 19224134. Preferred fragments of 19224134 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 107

MVSSYMFARGEKMNNKMFLNKEAGFLVHTKRKRRFAVTLVGVFFLLLACAGAIGFGQVAYAADEKTVPNFKSPDP
DYPWYGYDSYRGIFARYHNLKVNLKGSKEYQAYCFNLTKYFPRPTYSTTNNFYKKIDGSGSAFKSYAANPRVLDE
NLDKLEKNILNVIYNGYKSNANGFMNGIEDLNAILVTQNAIWYYSDSAPLNDVNKMWEREVRNGEISESQVTLMR
EALKKLIDPNLEATAANKIPSGYRLNIFKSENEDYQNLLSAEYVPDDPPKPGDTSEHNPKTPELDGTPIPEDPKR
PDESSEPALPPLMPELDGEEVPEVPSESLEPALPPLMPELDGEEVPEVPSESLEPALPPLMPELDGEEVPEVPSE
SLEPALPPLMPELDGEEVPEVPSESLEPALPPLMPELDGEEVPEKPSVDLPIEVPRYEFNNKDQSPLAGESGETE
YITEVYGNQQNPVDIDKKLPNETGFSGNMVETEDTKEPEVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDTKEP
EVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDTKEPGVLMGGQSESVEFTKDTQTGMSGQTTPQVETEDTKEPG
VLMGGQSESVEFTKDTQTGMSGFSETVTIVEDTRPKLVFHFDNNEPKVEENREKPTKNITPILPATGDIENVLAF
LGILILSVLSIFSLLKNKQNNKV

19224135 is thought to be a capsular polysaccharide adhesin (Cpa) protein. An example of a nucleotide sequence encoding the Cpa protein (SEQ ID NO: 108) and a Cpa protein amino acid sequence (SEQ ID NO: 109) are set forth below.

SEQ ID NO: 108

ATGAATAACAAAAAATTGCAAAAGAAGCAAGATGCTCCTCGGGTATCAAACAGAAAGCCAAAACAATTAACTGTC
ACTTTAGTGGGAGTATTTTTAATGTTTTTGACCTTGGTAAGTTCCATGAGAGGTGCTCAAAGCATATTTGGAGAG
GAAAAGAGAATTGAAGAAGTCAGTGTTCCTAAAATAAAAAGTCCAGATGATGCCTACCCTTGGTATGGCTATGAT
TCATATGACTCTAGTCATCCTTACTATGAACGTTTTAAAGTAGCACATGATTTAAGGGTTAATTTAAATGGAAGT
AAGAGCTACCAAGTATATTGCTTTAATATCAATTCTCATTATCCGAATAGAAAAAATGCTTTTTCTAAACAATGG
TTTAAGAGAGTTGATGGGACAGGTGATGTGTTCACAAATTATGCTCAGACACCTAAGATTCGTGGAGAATCATTG
AATAATAAACTTTTAAGTATTATGTACAACGCTTATCCTAAAAATGCTAATGGCTATATGGATAAGATAGAACCA

TTAAATGCTATTTAGTAACTCAACAAGCTGTTTGGTACTATTCTGACAGTTCTTATGGTAATATAAAAACGTTA
TGGGCATCTGAGCTTAAAGACGGAAAAATAGATTTTGAACAAGTAAAATTAATGCGTGAAGCTTACTCAAAACTA
ATTAGTGATGATTTAGAAGAAACATCTAAAAATAAGCTACCTCAAGGATCTAAACTGAATATTTTTGTTCCGCAA
GATAAATCTGTTCAAAATTTATTAAGTGCAGAGTACGTGCCTGAATCCCCTCCGGCACCAGGTCAGTCTCCAGAA
CCGCCAGTGCAAACAAAAAAAACATCAGTCATTATCAGAAAATATGCGGAAGGTGACTACTCTAAACTTCTAGAG
GGAGCAACTTTGCGTTTAACAGGGGAAGATATCCTAGATTTTCAAGAAAAAGTCTTCCAAAGTAATGGAACAGGA
GAAAAGATTGAATTATCAAATGGGACTTATACCTTAACAGAAACATCATCTCCAGATGGATATAAAATTGCGGAG
CCGATTAAGTTTAGAGTAGTGAATAAAAAAGTATTTATCGTCCAAAAAGATGGTTCTCAAGTGGAAAATCCAAAC
AAAGAAGTAGCAGAGCCATACTCAGTGGAAGCGTACAGCGATATGCAAGATAGTAACTATATTAATCCAGAAACG
TTCACTCCTTATGGGAAATTTTATTACGCTAAAAATAAGGATAAAAGTTCACAAGTTGTCTACTGTTTTAATGCT
GATTTACACTCTCCACCTGAATCAGAGGATGGGGGAGGAACTATAGATCCTGATATTAGTACGATGAAAGAAGTC
AAGTACACACATACGGCAGGTAGTGATTTGTTTAAATACGCGCTAAGACCGAGAGATACAAATCCAGAAGACTTC
TTAAAGCACATTAAAAAAGTAATTGAAAAAGGCTACAATAAAAAAGGTGATAGCTATAATGGATTAACAGAAACA
CAGTTTCGCGCGGCTACTCAGCTTGCTATCTATTACTTTACAGACAGCACTGACTTAAAAACCTTAAAAACTTAT
AACAATGGGAAAGGTTACCATGGATTTGAATCTATGGATGAAAAAACCCTAGCTGTAACAAAAGAATTAATTAAT
TACGCTCAAGATAATAGTGCCCCTCAACTAACAAATCTTGATTTCTTCGTACCTAATAATAGCAAATACCAATCT
CTTATTGGGACAGAATACCATCCAGATGATTTGGTTGACGTGATTCGTATGGAAGATAAAAAGCAAGAAGTTATT
CCAGTAACTCACAGTTTGACAGTGAAAAAAACAGTAGTCGGTGAGTTGGGAGATAAAACTAAAGGCTTCCAATTT
GAACTTGAGTTGAAAGATAAAACTGGACAGCCTATTGTTAACACTCTAAAAACTAATAATCAAGATTTAGTAGCT
AAAGATGGGAAATATTCATTTAATCTAAAGCATGGTGACACCATAAGAATAGAAGGATTACCGACGGGATATTCT
TATACTCTGAAAGAGACTGAAGCTAAGGATTATATAGTAACCGTTGATAACAAAGTTAGTCAAGAAGCTCAATCA
GCAAGTGAGAATGTCACAGCAGACAAAGAAGTCACTTTTGAAAACCGTAAAGATCTTGTCCCACCAACTGGTTTT
ATTACTGATGGTGGAACCTATCTGTGGTTATTATTGCTTGTCCCATTTGGTTTGTTAGTGTGGTTCTTTGGTCGT
AAAGGACTAAAAAATGACTAA

SEQ ID NO: 109

MNNKKLQKKQDAPRVSNRKPKQLTVTLVGVFLMFLTLVSSMRGAQSIFGEEKRIEEVSVPKIKSPDDAYPWYGYD
SYDSSHPYYERFKVAHDLRVNLNGSKSYQVYCFNINSHYPNRKNAFSKQWFKRVDGTGDVFTNYAQTPKIRGESL
NNKLLSIMYNAYPKNANGYMDKIEPLNAILVTQQAVWYYSDSSYGNIKTLWASELKDGKIDFEQVKLMREAYSKL
ISDDLEETSKNKLPQGSKLNIFVPQDKSVQNLLSAEYVPESPPAPGQSPEPPVQTKKTSVIIRKYAEGDYSKLLE
GATLRLTGEDILDFQEKVFQSNGTGEKIELSNGTYTLTETSSPDGYKIAEPIKFRVVNKKVFIVQKDGSQVENPN
KEVAEPYSVEAYSDMQDSNYINPETFTPYGKFYYAKNKDKSSQVVYCFNADLHSPPESEDGGGTIDPDISTMKEV
KYTHTAGSDLFKYALRPRDTNPEDFLKHIKKVIEKGYNKKGDSYNGLTETQFRAATQLAIYYFTDSTDLKTLKTY
NNGKGYHGFESMDEKTLAVTKELINYAQDNSAPQLTNLDFFVPNNSKYQSLIGTEYHPDDLVDVIRMEDKKQEVI
PVTHSLTVKKTVVGELGDKTKGFQFELELKDKTGQPIVNTLKTNNQDLVAKDGKYSFNLKHGDTIRIEGLPTGYS
YTLKETEAKDYIVTVDNKVSQEAQSASENVTADKEVTFENRKDL*VPPTG*FITDGGTYLWLLLLVPFGLLVWFFGR
KGLKND

19224135 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 184** VPPTG (shown in italics in SEQ ID NO: 109, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant 19224135 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in 19224135. The pilin motif sequence is underlined in SEQ ID NO: 109, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 164 and 172. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of 19224135 include at least one conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 109

MNNKKLQKKQDAPRVSNRKPKQLTVTLVGVFLMFLTLVSSMRGAQSIFGEEKRIEEVSVPKIKSPDDAYPWYGYD
SYDSSHPYYERFKVAHDLRVNLNGSKSYQVYCFNINSHYPNRKNAFSKQWFKRVDGTGDVFTNYAQTPKIRGESL
NNKLLSIMYNAYPKNANGYMDKIEPLNAILVTQQAVWYYSDSSYGNIKTLWASELKDGKIDFEQVKLMREAYSKL
ISDDLEETSKNKLPQGSKLNIFVPQDKSVQNLLSAEYVPESPPAPGQSPEPPVQTKKTSVIIRKYAEGDYSKLLE
GATLRLTGEDILDFQEKVFQSNGTGEKIELSNGTYTLTETSSPDGYKIAEPIKFRVVNKKVFIVQKDGSQVENPN
KEVAEPYSVEAYSDMQDSNYINPETFTPYGKFYYAKNKDKSSQVVYCFNADLHSPPESEDGGGTIDPDISTMKEV
KYTHTAGSDLFKYALRPRDTNPEDFLKHIKKVIEKGYNKKGDSYNGLTETQFRAATQLAIYYFTDSTDLKTLKTY
NNGKGYHGFESMDEKTLAVTKELINYAQDNSAPQLTNLDFFVPNNSKYQSLIGTEYHPDDLVDVIRMEDKKQEVI
PVTHSLTVKKTVVGELGDKTKGFQFELELKDKTGQPIVNTLKTNNQDLVAKDGKYSFNLKHGDTIRIEGLPTGYS
YTLKETEAKDYIVTVDNKVSQEAQSASENVTADKEVTFENRKDLVPPTGFITDGGTYLWLLLLVPFGLLVWFFGR
KGLKND

An E box containing a conserved glutamic residue has been identified in 19224135. The E-box motif is underlined in SEQ ID NO: 109, below. The conserved glutamic acid (E), at amino acid residue 339, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of 19224135. Preferred fragments of 19224135 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 109

MNNKKLQKKQDAPRVSNRKPKQLTVTLVGVFLMFLTLVSSMRGAQSIFGEEKRIEEVSVPKIKSPDDAYPWYGYD
SYDSSHPYYERFKVAHDLRVNLNGSKSYQVYCFNINSHYPNRKNAFSKQWFKRVDGTGDVFTNYAQTPKIRGESL
NNKLLSIMYNAYPKNANGYMDKIEPLNAILVTQQAVWYYSDSSYGNIKTLWASELKDGKIDFEQVKLMREAYSKL
ISDDLEETSKNKLPQGSKLNIFVPQDKSVQNLLSAEYVPESPPAPGQSPEPPVQTKKTSVIIRKYAEGDYSKLLE
GATLRLTGEDILDFQEKVFQSNGTGEKIELSNGTYTLT**E**TSSPDGYKIAEPIKFRVVNKKVFIVQKDGSQVENPN
KEVAEPYSVEAYSDMQDSNYINPETFTPYGKFYYAKNKDKSSQVVYCFNADLHSPPESEDGGGTIDPDISTMKEV
KYTHTAGSDLFKYALRPRDTNPEDFLKHIKKVIEKGYNKKGDSYNGLTETQFRAATQLAIYYFTDSTDLKTLKTY
NNGKGYHGFESMDEKTLAVTKELINYAQDNSAPQLTNLDFFVPNNSKYQSLIGTEYHPDDLVDVIRMEDKKQEVI
PVTHSLTVKKTVVGELGDKTKGFQFELELKDKTGQPIVNTLKTNNQDLVAKDGKYSFNLKHGDTIRIEGLPTGYS
YTLKETEAKDYIVTVDNKVSQEAQSASENVTADKEVTFENRKDLVPPTGFITDGGTYLWLLLLVPFGLLVWFFGR
KGLKND

19224136 is thought to be a LepA protein. An example of a nucleotide sequence encoding the LepA protein (SEQ ID NO: 110) and a LepA protein amino acid sequence (SEQ ID NO: 111) are set forth below.

SEQ ID NO: 110

ATGACTAATTACCTAAATCGCTTAAATGAGAATCCACTATTTAAAGCTTTCATACGGTTAGTACTTAAGATTTCT
ATTATTGGATTTCTAGGTTACATTCTATTTCAGTATGTTTTTGGCGTCATGATTGTTAACACAAATCAGATGAGT
CCTGCTGTAAGTGCTGGTGATGGAGTCTTATATTATCGTTTGACTGATCGCTATCATATTAATGATGTGGTGGTC
TATGAGGTTGATAACACTTTGAAAGTTGGTCGAATTGCCGCTCAAGCTGGCGATGAGGTTAGTTTTACGCAAGAA
GGAGGACTGTTGATTAATGGGCATCCACCAGAAAAAGAGGTCCCTTACCTGACGTATCCTCACTCAAGTGGTCCA
AACTTTCCCTATAAAGTTCCTACGGGTACGTATTTCATATTGAATGATTATCGTGAAGAACGTTTGGACAGTCGT
TATTATGGGGCGTTACCCATCAATCAAATCAAAGGGAAAATCTCAACTCTATTAAGAGTGAGAGGAATTTAA

SEQ ID NO: 111

MTNYLNRLNENPLFKAFIRLVLKISIIGFLGYILFQYVFGVMIVNTNQMSPAVSAGDGVLYYRLTDRYHINDVVV
YEVDNTLKVGRIAAQAGDEVSFTQEGGLLINGHPPEKEVPYLTYPHSSGPNFPYKVPTGTYFILNDYREERLDSR
YYGALPINQIKGKISTLLRVRGI

19224137 is thought to be a fimbrial protein. An example of a nucleotide sequence encoding the fimbrial protein (SEQ ID NO: 112) and a fimbrial protein amino acid sequence (SEQ ID NO: 113) are set forth below.

SEQ ID NO: 112

ATGAAAAAAATAATTACTTGCTACTGCAATCTTAGCAACTGCTTTAGGAACAGCTTCTTTAAATCAAAAC
GTAAAAGCTGAGACGGCAGGGGTTGTTAGCAGTGGTCAATTAACAATAAAAAAATCAATTACAAATTTTAATGAT
GATACACTTTTGATGCCTAAGACAGACTATACTTTTAGCGTTAATCCGGATAGTGCGGCTACAGGTACTGAAAGT
AATTTACCAATTAAACCAGGTATTGCTGTTAACAATCAAGATATTAAGGTTTCTTATTCTAATACTGATAAGACA
TCAGGTAAAGAAAAACAAGTTGTTGTTGACTTTATGAAAGTTACTTTTCCTAGCGTTGGTATTTACCGTTATGTT
GTTACCGAGAATAAAGGGACAGCAGAAGGAGTTACATATGATGATACAAAATGGTTAGTTGACGTCTATGTTGGT
AATAATGAAAAGGGAGGTCTTGAACCAAAGTATATTGTATCTAAAAAAGGAGATTCTGCTACTAAAGAACCAATC
CAGTTTAATAATTCATTCGAAACAACGTCATTAAAAATTGAAAAGGAAGTTACTGGTAATACAGGAGATCATAAA
AAAGCATTTACCTTTACATTAACATTGCAACCAAATGAATACTATGAGGCAAGTTCGGTTGTGAAAATTGAAGAG
AACGGACAAACGAAAGATGTGAAAATTGGGGAGGCATATAAGTTTACTTTGAACGATAGTCAGAGTGTGATATTG
TCTAAATTACCAGTTGGTATTAATTATAAAGTTGAAGAAGCAGAAGCTAATCAAGGTGGATATACTACAACAGCA
ACTTTAAAAGATGGAGAAAAGTTATCTACTTATAACTTAGGTCAGGAACATAAAACAGACAAGACTGCTGATGAA
ATCGTTGTCACAAATAACCGTGACACTCAAGTTCCAACTGGTGTTGTAGGCACCCTTGCTCCATTTGCAGTTCTT
AGCATTGTGGCTATTGGTGGAGTTATCTATATTACAAAACGTAAAAAAGCTTAA

SEQ ID NO: 113

MKKNKLLLATAILATALGTASLNQNVKAETAGVVSSGQLTIKKSITNFNDDTLLMPKTDYTFSVNPDSAATGTES
NLPIKPGIAVNNQDIKVSYSNTDKTSGKEKQVVVDFMKVTFPSVGIYRYVVTENKGTAEGVTYDDTKWLVDVYVG
NNEKGGLEPKYIVSKKGDSATKEPIQFNNSFETTSLKIEKEVTGNTGDHKKAFTFTLTLQPNEYYEASSVVKIEE
NGQTKDVKIGEAYKFTLNDSQSVILSKLPVGINYKVEEAEANQGGYTTTATLKDGEKLSTYNLGQEHKTDKTADE
IVVTNNRDT*QVPTG*VVGTLAPFAVLSIVAIGGVIYITKRKKA

19224137 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 140** QVPTG (shown in italics in SEQ ID NO: 113, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant 19224137 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in 19224137. The pilin motif sequence is underlined in SEQ ID NO: 113, below. A conserved lysine (K) residue is also marked in bold, at amino acid residue 160. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of 19224137 include the conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 113

MKKNKLLLATAILATALGTASLNQNVKAETAGVVSSGQLTIKKSITNFNDDTLLMPKTDYTFSVNPDSAATGTES
NLPIKPGIAVNNQDIKVSYSNTDKTSGKEKQVVVDFMKVTFPSVGIYRYVVTENKGTAEGVTYDDTKWLVDVYVG
NNEKGGLEP**K**YIVSKKGDSATKEPIQFNNSFETTSLKIEKEVTGNTGDHKKAFTFTLTLQPNEYYEASSVVKIEE
NGQTKDVKIGEAYKFTLNDSQSVILSKLPVGINYKVEEAEANQGGYTTTATLKDGEKLSTYNLGQEHKTDKTADE
IVVTNNRDTQVPTGVVGTLAPFAVLSIVAIGGVIYITKRKKA

An E box containing a conserved glutamic residue has been identified in 19224137. The E-box motif is underlined in SEQ ID NO: 113, below. The conserved glutamic acid (E), at amino acid residue 263, is marked in bold. The E box motif, in particular the conserved glutamic acid residue, is thought to be important for the formation of oligomeric pilus-like structures of 19224137. Preferred fragments of 19224137 include the conserved glutamic acid residue. Preferably, fragments include the E box motif.

SEQ ID NO: 113

MKKNKLLLATALATALGTASLNQNVKAETAGVVSSGQLTIKKSITNFNDDTLLMPKTDYTFSVNPDSAATGTES
NLPIKPGIAVNNQDIKVSYSNTDKTSGKEKQVVVDFMKVTFPSVGIYRYVVTENKGTAEGVTYDDTKWLVDVYVG
NNEKGGLEPKYIVSKKGDSATKEPIQFNNSFETTSLKIEKEVTGNTGDHKKAFTFTLTLQPNEYYEASSVVKIEE
NGQTKDVKIGEAYKFTLNDSQSVILSKLPVGINYKVEEAEANQGGYTTTATLKDGEKLSTYNLGQEHKTDKTADE
IVVTNNRDTQVPTGVVGTLAPFAVLSIVAIGGVIYITKRKKA

19224138 is thought to be a SrtC2-type sortase. An example of a nucleotide sequence encoding the SrtC2 sortase (SEQ ID NO: 114) and a SrtC2 sortase amino acid sequence (SEQ ID NO: 115) are set forth below.

SEQ ID NO: 114

ATGATGATGACAATTGTACAGGTTATCAATAAAGCCATTGATACTCTCATTCTTATCTTTTGTTTAGTCGTACTA
TTTTTAGCTGGTTTTGGTTTGTGGGATTCTTATCATCTCTATCAACAAGCAGACGCTTCTAATTTCAAAAAATTT
AAAACAGCTCAACAACAGCCTAAATTTGAAGACTTGTTAGCTTTGAATGAGGATGTCATTGGTTGGTTAAATATC
CCGGGGACTCATATTGATTATCCTCTAGTTCAGGGAAAAACGAATTTAGAGTATATTAATAAAGCAGTTGATGGC
AGTGTTGCCATGTCTGGTAGTTTATTTTTAGATACACGGAATCATAATGATTTTACGGACGATTACTCTCTGATT
TATGGCCATCATATGGCAGGTAATGCCATGTTTGGCGAAATTCCAAAATTTTTAAAAAAGGATTTTTTCAACAAA
CATAATAAAGCTATCATTGAAACAAAAGAGAGAAAAAAACTAACCGTCACTATTTTTGCTTGTCTCAAGACAGAT
GCCTTTGACCAGTTAGTTTTTAATCCTAATGCTATTACCAATCAAGACCAACAAAGGCAGCTCGTTGATTATATC
AGTAAAAGATCAAAACAATTTAAACCTGTTAAATTGAAGCATCATACAAAGTTCGTTGCTTTTTCAACGTGTGAA
AATTTTTCTACTGACAATCGTGTTATCGTTGTCGGTACTATTCAAGAATAA

SEQ ID NO: 115

MMMTIVQVINKAIDTLILIFCLVVLFLAGFGLWDSYHLYQQADASNFKKFKTAQQQPKFEDLLALNEDVIGWLNI
PGTHIDYPLVQGKTNLEYINKAVDGSVAMSGSLFLDTRNHNDFTDDYSLIYGHHMAGNAMFGEIPKFLKKDFFNK
HNKAIIETKERKKLTVTIFACLKTDAFDQLVFNPNAITNQDQQRQLVDYISKRSKQFKPVKLKHHTKFVAFSTCE
NFSTDNRVIVVGTIQE

19224139 is an open reading frame that encodes a sortase substrate motif LPXAG shown in italics in SEQ ID NO: 117. An example of a nucleotide sequence of the open reading frame (SEQ ID NO: 116) and the amino acid sequence encoded by the open reading frame (SEQ ID NO: 117) are set forth below.

SEQ ID NO: 116

ATGTTATTTTCTGTCGTAATGATATTAACCATGCTGGCCTTTAATCAGACTGTTTTAGCAAAAGACAGCACTGTT
CAAACTAGCATTAGTGTCGAAAATGTCTTAGAGAGAGCAGGCGATAGTACCCCATTTTCGATTGCATTAGAATCA
ATTGATGCGATGAAAACAATAGAAGAAATAACAATTGCTGGTTCTGGAAAAGCAAGCTTTTCCCCTCTGACCTTC
ACAACAGTTGGGCAATATACTTATCGTGTTTATCAGAAGCCTTCACAAAATAAAGATTATCAAGCAGATACTACT
GTATTTGACGTTCTTGTCTATGTGACCTATGATGAAGATGGGACTCTAGTCGCAAAAGTTATTTCTCGAAGGGCT
GGAGACGAAGAAAAATCAGCGATTACTTTTAAGCCCAAACGGTTAGTAAAACCAATACCGCCTAGACAACCTAAC
ATCCCTAAAACCCCATTACCATTAGCTGGTGAAGTAAAAAGTTTATTGGGTATCTTAAGTATCGTATTACTGGGG
TTACTAGTTCTTCTTTATGTTAAAAAACTGAAGAG

SEQ ID NO: 117

MLFSVVMILTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTPFSIALESIDAMKTIEEITIAGSGKASFSPLTF
TTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKPKRLVKPIPPRQPN
IPKTP*LPLAG*EVKSLLGILSIVLLGLLVLLYVKKLKSKL

19224139 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 185** LPLAG (shown in italics in SEQ ID NO: 117, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant 19224139 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular

domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

A pilin motif, discussed above, containing a conserved lysine (K) residue has also been identified in 19224139. The pilin motif sequence is underlined in SEQ ID NO: 117, below. A conserved lysine (K) residue is also marked in bold, at amino acid residue 138. The pilin sequence, in particular the conserved lysine residue, is thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of 19224139 include the conserved lysine residue. Preferably, fragments include the pilin sequence.

SEQ ID NO: 117

MLFSVVMILTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTPFSIALESIDAMKTIEEITIAGSGKASFSPLTF
TTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGDEEKSAITFKP**K**RLVKPIPPRQPN
IPKTPLPLAGEVKSLLGILSIVLLGLLVLLYVKKLKSKL

Two E boxes containing conserved glutamic residues have been identified in 19224139. The E-box motifs are underlined in SEQ ID NO: 117, below. The conserved glutamic acid (E) residues, at amino acid residues 58 and 128, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of 19224139. Preferred fragments of 19224139 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 117

MLFSVVMILTMLAFNQTVLAKDSTVQTSISVENVLERAGDSTPFSIALESIDAMKTI**E**EITIAGSGKASFSPLTF
TTVGQYTYRVYQKPSQNKDYQADTTVFDVLVYVTYDEDGTLVAKVISRRAGD**E**EKSAITFKPKRLVKPIPPRQPN
IPKTPLPLAGEVKSLLGILSIVLLGLLVLLYVKKLKSKL

19224140 is thought to be a MsmRL protein. An example of a nucleotide sequence encoding the MsmRL protein (SEQ ID NO: 118) and a MsmRL protein amino acid sequence (SEQ ID NO: 119) are set forth below.

SEQ ID NO: 118

ATGGTTATATTCGATTTAAAACATGTGCAAACATTACACAGCTTGTCTCAATTACCTATTTCAGTGATGTCACAA
GATAAGGCACTTATTCAAGTATATGGTAATGACGACTATTTATTATGTTACTATCAATTTTTAAAGCATCTAGCT
ATTCCTCAAGCTGCACAAGATGTTATTTTTTATGAGGGTTTATTTGAAGAGTCCTTTATGATTTTTCCTCTTTGT
CACTACATTATTGCCATTGGACCTTTCTACCCTTATTCACTTAATAAAGACTATCAGGAACAATTAGCTAATAAT
TTTTTAAAACATTCTTCTCATCGTAGCAAAGAAGAGCTCTTATCCTATATGGCACTTGTCCCACATTTTCCAATT
AATAATGTGCGGAACCTTTTGATAGCTATTGACGCTTTTTTTTGACACACAATTTGAGACGACTTGCCAACAAACA
ATTCATCAATTGTTGCAGCATTCAAAACAGATGACTGCTGATCCTGATATCATTCATCGCCTTAAGCATATTAGC
AAAGCATCTAGCCAACTACCGCCTGTTTTAGAGCACCTAAATCATATTATGGATCTGGTAAAGCTAGGCAATCCA
CAATTGCTCAAGCAAGAAATCAATCGCATCCCCTTATCAAGTATCACCTCATCTTCTATTTCTGCTCTAAGGGCG
GAAAAGAACCTCACTGTTATCTATTTAACTAGGTTACTGGAATTCAGTTTTGTAGAAAATACTGACGTAGCAAAG
CATTATAGCCTTGTCAAATACTACATGGCCTTAAATGAAGAAGCGAGTGACTTGCTCAAAGTTTTGAGAATTCGC
TGTGCAGCCATCATCCATTTTTCCGAATCATTAACCAATAAAAGTATTTCTGATAAACGTCAAATGTACAATAGT
GTGCTTCATTATGTCGATAGTCACCTGTATTCCAAATTAAAGGTATCTGATATCGCTAAGCGCCTATATGTTTCC
GAATCTCACTTACGTTCAGTCTTTAAAAAATACTCAAATGTTTCCTTACAACATTATATTCTAAGTACAAAAATC
AAAGAAGCTCAACTACTCTTAAAACGAGGAATTCCTGTTGGAGAAGTGGCTAAAAGCTTATATTTTTATGACACT
ACCCATTTTCATAAAATCTTTAAAAAATACACGGGTATTTCTTCAAAAGACTATCTTGCTAAATACCGAGATAAT
ATTTAA

SEQ ID NO: 119

MVIFDLKHVQTLHSLSQLPISVMSQDKALIQVYGNDDYLLCYYQFLKHLAIPQAAQDVIFYEGLFEESFMIFPLC
HYIIAIGPFYPYSLNKDYQEQLANNFLKHSSHRSKEELLSYMALVPHFPINNVRNLLIAIDAFFDTQFETTCQQT

IHQLLQHSKQMTADPDTHRLKHISKASSQLPPVLEHLNHIMDLVKLGNPQLLKQEINRIPLSSITSSSISALRA
EKNLTVIYLTRLLEFSFVENTDVAKHYSLVKYYMALNEEASDLLKVLRIRCAAIIHFSESLTNKSISDKRQMYNS
VLHYVDSHLYSKLKVSDIAKRLYVSESHLRSVFKKYSNVSLQHYILSTKIKEAQLLLKRGIPVGEVAKSLYFYDT
THFHKIFKKYTGISSKDYLAKYRDNI

19224141 is thought to be a protein F2 fibronectin binding protein. An example of a nucleotide sequence encoding the protein F2 fibronectin binding protein (SEQ ID NO: 120) and a protein F2 fibronectin binding protein amino acid sequence (SEQ ID NO: 121) are set forth below.

SEQ ID NO: 120

ATGACACAAAAAAATAGCTATAAGTTAAGCTTCCTGTTATCCCTAACAGGATTTATTTTAGGTTTATTATTGGTT
TTTATAGGATTGTCCGGAGTATCAGTAGGACATGCGGAAACAAGAAATGGAGCAAACAAACAAGGATCTTTTGAA
ATCAAGAAAGTCGACCAAAACAATAAGCCTTTACCGGGAGCAACTTTTTCACTGACATCAAAGGATGGCAAGGGA
ACATCTGTTCAAACGTTCACTTCAAATGATAAAGGTATTGTAGATGCTCAAAATCTCCAACCAGGGACTTATACC
TTAAAAGAAGAAACAGCACCAGATGGTTATGATAAAACCAGCCGGACTTGGACAGTGACTGTTTATGAGAACGGC
TATACCAAGTTGGTTGAAAATCCCTATAATGGGGAAATCATCAGTAAAGCAGGGTCAAAAGATGTTAGTAGTTCT
TTACAGTTGGAAAATCCCAAAATGTCAGTTGTTTCTAAATATGGGAAAACAGAGGTTAGTAGTGGCGCAGCGGAT
TTCTACCGCAACCATGCCGCCTATTTTAAAATGTCTTTTGAGTTGAAACAAAAGGATAAATCTGAAACAATCAAC
CCAGGTGATACCTTTGTGTTACAGCTGGATAGACGTCTCAATCCTAAAGGTATCAGTCAAGATATCCCTAAAATC
ATTTACGACAGTGCAAATAGTCCGCTTGCGATTGGAAAATACCATGCTGAGAACCATCAACTTATCTATACTTTC
ACAGATTATATTGCGGGTTTAGATAAAGTCCAGTTGTCTGCAGAATTGAGCTTATTCCTAGAGAATAAGGAAGTG
TTGGAAAATACTAGTATCTCAAATTTTAAGAGTACCATAGGTGGGCAGGAGATCACCTATAAAGGAACGGTTAAT
GTTCTTTATGGAAATGAGAGCACTAAAGAAAGCAATTATATTACTAATGGATTGAGCAATGTGGGTGGGAGTATT
GAAAGCTACAACACCGAAACGGGAGAATTTGTCTGGTATGTTTATGTCAATCCAAACCGTACCAATATTCCTTAT
GCGACCATGAATTTATGGGGATTTGGAAGGGCTCGTTCAAATACAAGCGACTTAGAAAACGACGCTAATACAAGT
AGTGCTGAGCTTGGAGAGATTCAGGTCTATGAAGTACCTGAAGGAGAAAAATTACCATCAAGTTATGGGGTTGAT
GTTACAAAACTTACTTTAAGAACGGATATCACAGCAGGCCTAGGAAATGGTTTTCAAATGACCAAACGTCAGCGA
ATTGACTTTGGAAATAATATCCAAAATAAAGCATTTATCATCAAAGTAACAGGGAAAACAGACCAATCTGGTAAG
CCATTGGTTGTTCAATCCAATTTGGCAAGTTTTCGTGGTGCTTCTGAATATGCTGCTTTTACTCCAGTTGGAGGA
AATGTCTACTTCCAAAACGAAATTGCCTTGTCTCCTTCTAAGGGTAGTGGTTCTGGGAAAAGTGAATTTACTAAG
CCCTCTATTACAGTAGCAAATCTAAAACGAGTGGCTCAGCTTCGCTTTAAGAAAATGTCAACTGACAATGTGCCA
TTGCCAGAAGCGGCTTTTGAGCTGCGTTCATCAAATGGTAATAGTCAGAAATTAGAAGCCAGTTCAAACACACAA
GGAGAGGTTCACTTTAAGGACCTGACCTCGGGCACATATGACCTGTATGAAACAAAAGCGCCAAAAGGTTATCAG
CAGGTGACAGAGAAATTGGCGACCGTTACTGTTGATACTACCAAACCTGCTGAGGAAATGGTCACTTGGGGAAGC
CCACATTCGTCTGTAAAAGTAGAAGCTAACAAAGAAGTCACGATTGTCAACCATAAAGAAACCCTTACGTTTTCA
GGGAAGAAAATTTGGGAGAATGACAGACCAGATCAACGCCCAGCAAAGATTCAAGTGCAACTGTTGCAAAATGGT
CAAAAGATGCCTAACCAGATTCAAGAAGTAACGAAGGATAACGATTGGTCTTATCACTTCAAAGACTTGCCTAAG
TACGATGCCAAGAATCAGGAGTATAAGTACTCAGTTGAAGAAGTAAATGTTCCAGACGGCTACAAGGTGTCGTAT
TTAGGAAATGATATATTTAACACCAGAGAAACAGAATTTGTGTTTGAACAGAATAACTTTAACCTTGAATTTGGA
AATGCTGAAATAAAAGGTCAATCTGGGTCAAAAATCATTGATGAAGACACGCTAACGTCTTTCAAAGGTAAGAAA
ATTTGGAAAAATGATACGGCAGAAAATCGTCCCCAAGCCATTCAAGTGCAGCTTTATGCTGATGGAGTGGCTGTG
GAAGGTCAAACCAAATTTATTTCTGGCTCAGGTAATGAGTGGTCATTTGAGTTTAAAAACTTGAAGAAGTATAAT
GGAACAGGTAATGACATCATTTACTCAGTTAAAGAAGTAACTGTTCCAACAGGTTATGATGTGACTTACTCAGCT
AATGATATTATTAATACCAAACGTGAGGTTATTACACAACAAGGACCGAAACTAGAGATTGAAGAAACGCTTCCG
CTAGAATCAGGTGCTTCAGGCGGTACCACTACTGTCGAAGACTCACGCCCAGTTGATACCTTATCAGGTTTATCA
AGTGAGCAAGGTCAGTCCGGTGATATGACAATTGAAGAAGATAGTGCTACCCATATTAAATTCTCAAAACGTGAT
ATTGACGGCAAAGAGTTAGCTGGTGCAACTATGGAGTTGCGTGATTCATCTGGTAAAACTATTAGTACATGGATT
TCAGATGGACAAGTGAAAGATTTCTACCTGATGCCAGGAAAATATACATTTGTCGAAACCGCAGCACCAGACGGT
TATGAGATAGCAACTGCTATTACCTTTACAGTTAATGAGCAAGGTCAGGTTACTGTAAATGGCAAAGCAACTAAA
GGTGACACTCATATTGTCATGGTTGATGCTTACAAGCCAACTAAGGGTTCAGGTCAGGTTATTGATATTGAAGAA
AAGCTTCCAGACGAGCAAGGTCATTCTGGTTCAACTACTGAAATAGAAGACAGTAAATCTTCAGACCTTATCATT
GGCGGTCAAGGTGAAGTTGTTGACACAACAGAAGACACACAAAGTGGTATGACGGGCCATTCTGGCTCAACTACT
GAAATAGAAGATAGCAAGTCTTCAGACGTTATCATTGGTGGTCAGGGGCAGGTTGTCGAGACAACAGAGGATACC
CAAACTGGCATGTACGGGGATTCTGGTTGTAAAACGGAAGTCGAAGATACTAAACTAGTACAATCCTTCCACTTT
GATAACAAGGAACCAGAAAGTAACTCTGAGATTCCTAAAAAAGATAAGCCAAAGAGTAATACTAGTTTACCAGCA
ACTGGTGAGAAGCAACATAATATGTTCTTTTGGATGGTTACTTCTTGCTCACTTATTAGTAGTGTTTTTGTAATA
TCACTAAAATCCAAAAAACGCCTATCATCATGTTAA

SEQ ID NO: 121

MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGSFEIKKVDQNNKPLPGATFSLTSKDGKG
TSVQTFTSNDKGIVDAQNLQPGTYTLKEETAPDGYDKTSRTWTVTVYENGYTKLVENPYNGEIISKAGSKDVSSS
LQLENPKMSVVSKYGKTEVSSGAADFYRNHAAYFKMSFELKQKDKSETINPGDTFVLQLDRRLNPKGISQDIPKI
IYDSANSPLAIGKYHAENHQLIYTFTDYIAGLDKVQLSAELSLFLENKEVLENTSISNFKSTIGGQEITYKGTVN
VLYGNESTKESNYITNGLSNVGGSIESYNTETGEFVWYVYVNPNRTNIPYATMNLWGFGRARSNTSDLENDANTS
SAELGEIQVYEVPEGEKLPSSYGVDVTKLTLRTDITAGLGNGFQMTKRQRIDFGNNIQNKAFIIKVTGKTDQSGK
PLVVQSNLASFRGASEYAAFTPVGGNVYFQNEIALSPSKGSGSGKSEFTKPSITVANLKRVAQLRFKKMSTDNVP
LPEAAFELRSSNGNSQKLEASSNTQGEVHFKDLTSGTYDLYETKAPKGYQQVTEKLATVTVDTTKPAEEMVTWGS
PHSSVKVEANKEVTIVNHKETLTFSGKKIWENDRPDQRPAKIQVQLLQNGQKMPNQIQEVTKDNDWSYHFKDLPK
YDAKNQEYKYSVEEVNVPDGYKVSYLGNDIFNTRETEFVFEQNNFNLEFGNAEIKGQSGSKIIDEDTLTSFKGKK
IWKNDTAENRPQAIQVQLYADGVAVEGQTKFISGSGNEWSFEFKNLKKYNGTGNDIIYSVKEVTVPTGYDVTYSA
NDIINTKREVITQQGPKLEIEETLPLESGASGGTTTVEDSRPVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRD
IDGKELAGATMELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEIATAITFTVNEQGQVTVNGKATK
GDTHIVMVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKSSDLIIGGQGEVVDTTEDTQSGMTGHSGSTT
EIEDSKSSDVIIGGQGQVVETTEDTQTGMYGDSGCKTEVEDTKLVQSFHFDNKEPESNSEIPKKDKPKSNTS*LPA*
*T*GEKQHNMFFWMVTSCSLISSVFVISLKSKKRLSSC

19224141 contains an amino acid motif indicative of a cell wall anchor: **SEQ ID NO: 181** LPATG (shown in italics in SEQ ID NO: 121, above). In some recombinant host cell systems, it may be preferable to remove this motif to facilitate secretion of a recombinant 19224141 protein from the host cell. Alternatively, in other recombinant host cell systems, it may be preferable to use the cell wall anchor motif to anchor the recombinantly expressed protein to the cell wall. The extracellular domain of the expressed protein may be cleaved during purification or the recombinant protein may be left attached to either inactivated host cells or cell membranes in the final composition.

Two pilin motifs, discussed above, containing conserved lysine (K) residues have also been identified in 19224141. The pilin motif sequences are underlined in SEQ ID NO: 121, below. Conserved lysine (K) residues are also marked in bold, at amino acid residues 157 and 163 and at amino acid residues 216, 224, and 238. The pilin sequence, in particular the conserved lysine residues, are thought to be important for the formation of oligomeric, pilus-like structures. Preferred fragments of 19224141 include at least one conserved lysine residue. Preferably, fragments include at least one pilin sequence.

SEQ ID NO: 121

MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGSFEIKKVDQNNKPLPGATFSLTSKDGKG
TSVQTFTSNDKGIVDAQNLQPGTYTLKEETAPDGYDKTSRTWTVTVYENGYTKLVENPYNGEIISKAGSKDV<u>SSS</u>
<u>LQLENPKMSVVSK</u>YGKTEVSSGAADFYRNHAAYFKMSFELKQKDKSETINPGDTFV<u>LQLDRRLNPK</u>GISQDIP<u>KI</u>
<u>IYDSANSPLAIGK</u>YHAENHQLIYTFTDYIAGLDKVQLSAELSLFLENKEVLENTSISNFKSTIGGQEITYKGTVN
VLYGNESTKESNYITNGLSNVGGSIESYNTETGEFVWYVYVNPNRTNIPYATMNLWGFGRARSNTSDLENDANTS
SAELGEIQVYEVPEGEKLPSSYGVDVTKLTLRTDITAGLGNGFQMTKRQRIDFGNNIQNKAFIIKVTGKTDQSGK
PLVVQSNLASFRGASEYAAFTPVGGNVYFQNEIALSPSKGSGSGKSEFTKPSITVANLKRVAQLRFKKMSTDNVP
LPEAAFELRSSNGNSQKLEASSNTQGEVHFKDLTSGTYDLYETKAPKGYQQVTEKLATVTVDTTKPAEEMVTWGS
PHSSVKVEANKEVTIVNHKETLTFSGKKIWENDRPDQRPAKIQVQLLQNGQKMPNQIQEVTKDNDWSYHFKDLPK
YDAKNQEYKYSVEEVNVPDGYKVSYLGNDIFNTRETEFVFEQNNFNLEFGNAEIKGQSGSKIIDEDTLTSFKGKK
IWKNDTAENRPQAIQVQLYADGVAVEGQTKFISGSGNEWSFEFKNLKKYNGTGNDIIYSVKEVTVPTGYDVTYSA
NDIINTKREVITQQGPKLEIEETLPLESGASGGTTTVEDSRPVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRD
IDGKELAGATMELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEIATAITFTVNEQGQVTVNGKATK
GDTHIVMVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKSSDLIIGGQGEVVDTTEDTQSGMTGHSGSTT
EIEDSKSSDVIIGGQGQVVETTEDTQTGMYGDSGCKTEVEDTKLVQSFHFDNKEPESNSEIPKKDKPKSNTSLPA
TGEKQHNMFFWMVTSCSLISSVFVISLKSKKRLSSC

Two E boxes containing conserved glutamic residues have been identified in 19224141. The E-box motifs are underlined in SEQ ID NO: 121, below. The conserved glutamic acid (E) residues, at

amino acid residues 567 and 944, are marked in bold. The E box motifs, in particular the conserved glutamic acid residues, are thought to be important for the formation of oligomeric pilus-like structures of 19224141. Preferred fragments of 19224141 include at least one conserved glutamic acid residue. Preferably, fragments include at least one E box motif.

SEQ ID NO: 121

MTQKNSYKLSFLLSLTGFILGLLLVFIGLSGVSVGHAETRNGANKQGSFEIKKVDQNNKPLPGATFSLTSKDGKG
TSVQTFTSNDKGIVDAQNLQPGTYTLKEETAPDGYDKTSRTWTVTVYENGYTKLVENPYNGEIISKAGSKDVSSS
LQLENPKMSVVSKYGKTEVSSGAADFYRNHAAYFKMSFELKQKDKSETINPGDTFVLQLDRRLNPKGISQDIPKI
IYDSANSPLAIGKYHAENHQLIYTFTDYIAGLDKVQLSAELSLFLENKEVLENTSISNFKSTIGGQEITYKGTVN
VLYGNESTKESNYITNGLSNVGGSIESYNTETGEFVWYVYVNPNRTNIPYATMNLWGFGRARSNTSDLENDANTS
SAELGEIQVYEVPEGEKLPSSYGVDVTKLTLRTDITAGLGNGFQMTKRQRIDFGNNIQNKAFIIKVTGKTDQSGK
PLVVQSNLASFRGASEYAAFTPVGGNVYFQNEIALSPSKGSGSGKSEFTKPSITVANLKRVAQLRFKKMSTDNVP
LPEAAFELRSSNGNSQKLEASSNTQGEVHFKDLTSGTYDLYETKAPKGYQQVTEKLATVTVDTTKPAEEMVTWGS
PHSSVKVEANKEVTIVNHKETLTFSGKKIWENDRPDQRPAKIQVQLLQNGQKMPNQIQEVTKDNDWSYHFKDLPK
YDAKNQEYKYSVEEVNVPDGYKVSYLGNDIFNTRETEFVFEQNNFNLEFGNAEIKGQSGSKIIDEDTLTSFKGKK
IWKNDTAENRPQAIQVQLYADGVAVEGQTKFISGSGNEWSFEFKNLKKYNGTGNDIIYSVKEVTVPTGYDVTYSA
NDIINTKREVITQQGPKLEIEETLPLESGASGGTTTVEDSRPVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRD
IDGKELAGATMELRDSSGKTISTWISDGQVKDFYLMPGKYTFVETAAPDGYEIATAITFTVNEQGQVTVNGKATK
GDTHIVMVDAYKPTKGSGQVIDIEEKLPDEQGHSGSTTEIEDSKSSDLIIGGQGEVVDTTEDTQSGMTGHSGSTT
EIEDSKSSDVIIGGQGQVVETTEDTQTGMYGDSGCKTEVEDTKLVQSFHFDNKEPESNSEIPKKDKPKSN
TSLPATGEKQHNMFFWMVTSCSLISSVFVISLKSKKRLSSC

As discussed above, applicants have also determined the nucleotide and encoded amino acid sequence of fimbrial structural subunits in several other GAS AI-4 strains of bacteria. Examples of sequences of these fimbrial structural subunits are set forth below.

M12 strain isolate 20010296 is a GAS AI-4 strain of bacteria. 20010296_fimbrial is thought to be a fimbrial structural subunit of M12 strain isolate 20010296. An example of a nucleotide sequence encoding the 20010296_fimbrial protein (SEQ ID NO: 257) and a 20010296_fimbrial protein amino acid sequence (SEQ ID NO: 258) are set forth below.

SEQ ID NO: 257

agcagtggtcaattaacaataaaaaaatcaattacaaattttaatgatgatacacttttg
atgcctaagacagactatacttttagcgttaatccggatagtgcggctacaggtactgaa
agtaatttaccaattaaaccaggtattgctgttaacaatcaagatattaaggtttcttat
tctaatactgataagacatcaggtaaagaaaaacaagttgttgttgactttatgaaagtt
acttttcctagcgttggtatttaccgttatgttgttaccgagaataaagggacagcagaa
ggagttacatatgatgatacaaaatggttagttgacgtctatgttggtaataatgaaaag
ggaggtcttgaaccaaagtatattgtatctaaaaaaggagattctgctactaaagaacca
atccagtttaataattcattcgaaacaacgtcattaaaaattgaaaaggaagttactggt
aatacaggagatcataaaaaaagcatttaactttacattaacattgcaaccaaatgaatac
tatgaggcaagttcggttgtgaaaattgaagagaacggacaaacgaaagatgtgaaaatt
ggggaggcatataagtttactttgaacgatagtcagagtgtgatattgtctaaattacca
gttggtattaattataaagttgaagaagcagaagctaatcaaggtggatatactacaaca
gcaactttaaaagatggagaaaagttatctacttataacttaggtcaggaacataaaaca
gacaagactgctgatgaaatcgt

SEQ ID NO: 258

SSGQLTIKKSITNFNDDTLLMPKTDYTFSVNPDSAATGTESNLP
IKPGIAVNNQDIKVSYSNTDKTSGKEKQVVVDFMKVTFPSVGIYRYVVTENKGTAEGV
TYDDTKWLVDVYVGNNEKGGLEPKYIVSKKGDSATKEPIQFNNSFETTSLKIEKEVTG
NTGDHKKAFNFTLTLQPNEYYEASSVVKIEENGQTKDVKIGEAYKFTLNDSQSVILSK
LPVGINYKVEEAEANQGGYTTTATLKDGEKLSTYNLGQEHKTDKTADEIV

PC M12 strain isolate 20020069 is a GAS AI-4 strain of bacteria. 20020069_fimbrial is thought to be a fimbrial structural subunit of M12 strain isolate 20020069. An example of a nucleotide sequence encoding the 20020069_fimbrial protein (SEQ ID NO: 259) and a 20020069_fimbrial protein amino acid sequence (SEQ ID NO: 260) are set forth below.

SEQ ID NO: 259

agcagtggtcaattaacaataaaaaaatcaattacaaattttaatgatgatacacttttg
atgcctaagacagactatacttttagcgttaatccggatagtgcggctacaggtactgaa
agtaatttaccaattaaaccaggtattgctgttaacaatcaagatattaaggtttcttat
tctaatactgataagacatcaggtaaagaaaaacaagttgttgctttatgaaagtt
actttcctagcgttggtatttaccgttatgttgttaccgagaataaagggacagcagaa
ggagttacatatgatgatacaaaatggttagttgacgtctatgttggtaataatgaaaag
ggaggtcttgaaccaaagtatattgtatctaaaaaaggagattctgctactaaagaacca
atccagtttaataattcattcgaaacaacgtcattaaaaattgaaaaggaagttactggt
aatacaggagatcataaaaaagcatttaactttacattaacattgcaaccaaatgaatac
tatgaggcaagttcggttgtgaaaattgaagagaacggacaaacgaaagatgtgaaaatt
ggggaggcatataagtttactttgaacgatagtcagagtgtgatattgtctaaattacca
gttggtattaattataaagttgaagaagcagaagctaatcaaggtggatatactacaaca
gcaactttaaaagatggagaaaagttatctacttataacttaggtcaggaacataaaaca
gacaagactgctgatgaaatcgt

SEQ ID NO: 260

SSGQLTIKKSITNFNDDTLLMPKTDYTFSVNPDSAATGTESNLP
IKPGIAVNNQDIKVSYSNTDKTSGKEKQVVVDFMKVTFPSVGIYRYVVTENKGTAEGV
TYDDTKWLVDVYVGNNEKGGLEPKYIVSKKGDSATKEPIQFNNSFETTSLKIEKEVTG
NTGDHKKAFNFTLTLQPNEYYEASSVVKIEENGQTKDVKIGEAYKFTLNDSQSVILSK
LPVGINYKVEEAEANQGGYTTTATLKDGEKLSTYNLGQEHKTDKTADEIV

M12 strain isolate CDC SS 635 is a GAS AI-4 strain of bacteria. CDC SS 635_fimbrial is thought to be a fimbrial structural subunit of M12 strain isolate CDC SS 635. An example of a nucleotide sequence encoding the CDC SS 635_fimbrial protein (SEQ ID NO: 261) and a CDC SS 635_fimbrial protein amino acid sequence (SEQ ID NO: 262) are set forth below.

SEQ ID NO: 261

gagacggcaggggttgttagcagtggtcaattaacaataaaaaaatcaattacaaatttt
aatgatgatacacttttgatgcctaagacagactatacttttagcgttaatccggatagt
gcggctacaggtactgaaagtaatttaccaattaaaccaggtattgctgttaacaatcaa
gatattaaggtttcttattctaatactgataagacatcaggtaaagaaaaacaagttgtt
gttgactttatgaaagttacttttcctagcgttggtatttaccgttatgttgttaccgag
aataaagggacagcagaaggagttacatatgatgatacaaaatggttagttgacgtctat
gttggtaataatgaaaagggaggtcttgaaccaaagtatattgtatctaaaaaaggagat
tctgctactaaagaaccaatccagtttaataattcattcgaaacaacgtcattaaaaatt
gaaaaggaagttactggtaatacaggagatcataaaaaagcatttaactttacattaaca
ttgcaaccaaatgaatactatgaggcaagttcggttgtgaaaattgaagagaacggacaa
acgaaagatgtgaaaattggggaggcatataagtttactttgaacgatagtcagagtgtg
atattgtctaaattaccagttggtattaattataaagttgaagaagcagaagctaatcaa
ggtggatatactacaacagcaactttaaaagatggagaaaagttatctacttataactta
ggtcaggaacataaaacagacaagactgctgatgaaatcgttgtcacaaataaccgtgac
act

SEQ ID NO: 262

ETAGVVSSGQLTIKKSITNFNDDTLLMPKTDYTFSVNPDSAATG
TESNLPIKPGIAVNNQDIKVSYSNTDKTSGKEKQVVVDFMKVTFPSVGIYRYVVTENK
GTAEGVTYDDTKWLVDVYVGNNEKGGLEPKYIVSKKGDSATKEPIQFNNSFETTSLKI
EKEVTGNTGDHKKAFNFTLTLQPNEYYEASSVVKIEENGQTKDVKIGEAYKFTLNDSQ
SVILSKLPVGINYKVEEAEANQGGYTTTATLKDGEKLSTYNLGQEHKTDKTADEIVVT
NNRDT

M5 strain isolate ISS4883 is a GAS AI-4 strain of bacteria. ISS4883_fimbrial is thought to be a fimbrial structural subunit of M5 strain isolate ISS 4883. An example of a nucleotide sequence encoding the ISS4883_fimbrial protein (SEQ ID NO: 265) and an ISS4883_fimbrial protein amino acid sequence (SEQ ID NO: 266) are set forth below.

SEQ ID NO: 265

```
gagacggcaggggttgtaacaggaaaatcactacaagttacaaagacaatgacttatgat
gatgaagaggtgttaatgcccgaaaccgcctttacttttactatagagcctgatatgact
gcaagtggaaaagaaggcgacctagatattaaaaatggaattgtagaaggcttagacaaa
caagtaacagtaaaatataagaatacagataaaacatctcaaaaaactaaaatagcacaa
tttgatttttctaaggttaaatttccagctataggtgtttaccgctatatggtttcagag
aaaaacgataaaaaagacggaattaggtacgatgataaaaagtggactgtagatgtttat
gttgggaataaggccaataacgaagaaggtttcgaagttctatatattgtatcaaaagaa
ggtacttctagtactaaaaaaccaattgaatttacaaactctattaaaactacttcctta
aaaattgaaaaacaaataactggcaatgcaggagatcgtaaaaaatcattcaacttcaca
ttaacattacaaccaagtgaatattataaaaccggatcagttgtgaaaatcgaacaggat
ggaagtaaaaaagatgtgacgataggaacgccttacaaatttactttgggacacggtaag
agtgtcatgttatcgaaattaccaattggtatcaattactatcttagtgaagacgaagcg
aataaagacggttacactacaacggcaacattaaaagaacaaggcaaagaaaagagttcc
gatttcactttgagtactcaaaaccagaaaacagacgaatctgctgacgaaatcgttgtc
acaaataagcgtgacactctcgag
```

SEQ ID NO: 266

```
ETAGVVTGKSLQVTKTMTYDDEEVLMPETAFTFTIEPDMTASGK
EGDLDIKNGIVEGLDKQVTVKYKNTDKTSQKTKIAQFDFSKVKFPAIGVYRYMVSEKN
DKKDGIRYDDKKWTVDVYVGNKANNEEGFEVLYIVSKEGTSSTKKPIEFTNSIKTTSL
KIEKQITGNAGDRKKSFNFTLTLQPSEYYKTGSVVKIEQDGSKKDVTIGTPYKFTLGH
GKSVMLSKLPIGINYYLSEDEANKDGYTTTATLKEQGKEKSSDFTLSTQNQKTDESAD
EIVVTNKRDTLE
```

M50 strain isolate ISS4538 is a GAS AI-4 strain of bacteria. ISS4538_fimbrial is thought to be a fimbrial structural subunit of M50 strain ISS 4538. An example of a nucleotide sequence encoding the ISS4538_fimbrial protein (SEQ ID NO: 255) and an ISS4538_fimbrial protein amino acid sequence (SEQ ID NO: 256) are set forth below.

SEQ ID NO: 255

```
atgaaaaaaaataaattattacttgctactgcaatcttagcaactgctttaggaacagct
tctttaaatcaaaacgtaaaagctgagacggcaggggttgttagcagtggtcaattaaca
ataaaaaaatcaattacaaattttaatgatgatacacttttgatgcctaagacagactat
acttttagcgttaatccggatagtgcggctacaggtactgaaagtaatttaccaattaaa
ccaggtattgctgttaacaatcaagatattaaggtttcttattctaatactgataagaca
tcaggtaaagaaaaacaagttgttgttgactttatgaaagttacttttcctagcgttggt
atttaccgttatgttgttaccgagaataaagggacagcagaaggagttacatatgatgat
acaaaatggttagttgacgtctatgttggtaataatgaaaagggaggtcttgaaccaaag
tatattgtatctaaaaaaggagattctgctactaaagaaccaatccagtttaataattca
ttcgaaacaacgtcattaaaaattgaaaagaaagttactggtaatacaggagatcataaa
aaagcatttaactttacattaacattgcaaccaaatgaatactatgaggcaagttcggtt
gtgaaaattgaagagaacggacaaacgaaagatgtgaaaattggggaggcatataagttt
actttgaacgatagtcagagtgtgatattgtctaaattaccagttggtattaattataaa
gttgaagaagcagaagctaatcaaggtggatatactacaacagcaactttaaaagatgga
gaaaagttatctacttataacttaggtcaggaacataaaacagacaagactgctgatgaa
atcgttgtcacaaataancgngacactcnagttccaacnggtgtngtaggcaccccncct
ccattcncagttcttancattgnggctantggtggngtnatntatnttacaaaacgnaaa
aaagnataa
```

SEQ ID NO: 256

```
MKKNKLLLATAILATALGTASLNQNVKAETAGVVSSGQLTIKKS
ITNFNDDTLLMPKTDYTFSVNPDSAATGTESNLPIKPGIAVNNQDIKVSYSNTDKTSG
```

KEKQVVVDFMKVTPPSVGLYRYVVTENKGTAEGVTYDDTKWLVDVYVGNNEKGGLEPK
YIVSKKGDSATKEPIQFNNSFETTSLKIEKKVTGNTGDHKKAFNFTLTLQPNEYYEAS
SVVKIEENGQTKDVKIGEAYKFTLNDSQSVILSKLPVGINYKVEEAEANQGGYTTTAT
LKDGEKLSTYNLGQEHKTDKTADEIVVTNXRDTXVPTGVVGTPPPFXVLXIXAXGGVX
YXTKRKKX

There may be an upper limit to the number of GAS proteins which will be in the compositions of the invention. Preferably, the number of GAS proteins in a composition of the invention is less than 20, less than 19, less than 18, less than 17, less than 16, less than 15, less than 14, less than 13, less than 12, less than 11, less than 10, less than 9, less than 8, less than 7, less than 6, less than 5, less than 4, or less than 3. Still more preferably, the number of GAS proteins in a composition of the invention is less than 6, less than 5, or less than 4. Still more preferably, the number of GAS proteins in a composition of the invention is 3.

The GAS proteins and polynucleotides used in the invention are preferably isolated, *i.e.*, separate and discrete, from the whole organism with which the molecule is found in nature or, when the polynucleotide or polypeptide is not found in nature, is sufficiently free of other biological macromolecules so that the polynucleotide or polypeptide can be used for its intended purpose.

Examples Other Gram positive bacterial Adhesin Island Sequences

The Gram positive bacteria AI polypeptides of the invention can, of course, be prepared by various means (*e.g.* recombinant expression, purification from a gram positive bacteria, chemical synthesis *etc.*) and in various forms (*e.g.* native, fusions, glycosylated, non-glycosylated *etc.*). They are preferably prepared in substantially pure form (*i.e.* substantially free from other streptococcal or host cell proteins) or substantially isolated form.

The Gram positive bacteria AI proteins of the invention may include polypeptide sequences having sequence identity to the identified Gram positive bacteria proteins. The degree of sequence identity may vary depending on the amino acid sequence (a) in question, but is preferably greater than 50% (*e.g.* 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.5% or more). Polypeptides having sequence identity include homologs, orthologs, allelic variants and mutants of the identified Gram positive bacteria proteins. Typically, 50% identity or more between two proteins is considered to be an indication of functional equivalence. Identity between proteins is preferably determined by the Smith-Waterman homology search algorithm as implemented in the MPSRCH program (Oxford Molecular), using an affinity gap search with parameters *gap open penalty=12* and *gap extension penalty=1*.

The Gram positive bacteria adhesin island polynucleotide sequences may include polynucleotide sequences having sequence identity to the identified Gram positive bacteria adhesin island polynucleotide sequences. The degree of sequence identity may vary depending on the polynucleotide sequence in question, but is preferably greater than 50% (*e.g.* 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.5% or more).

PCT/US2005/027239

The Gram positive bacteria adhesin island polynucleotide sequences of the invention may include polynucleotide fragments of the identified adhesin island sequences. The length of the fragment may vary depending on the polynucleotide sequence of the specific adhesin island sequence, but the fragment is preferably at least 10 consecutive polynucleotides, (*e.g.* at least 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200 or more).

The Gram positive bacteria adhesin island amino acid sequences of the invention may include polypeptide fragments of the identified Gram positive bacteria proteins. The length of the fragment may vary depending on the amino acid sequence of the specific Gram positive bacteria antigen, but the fragment is preferably at least 7 consecutive amino acids, (*e.g.* 8, 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200 or more). Preferably the fragment comprises one or more epitopes from the sequence. The fragment may comprise at least one T-cell or, preferably, a B-cell epitope of the sequence. T- and B-cell epitopes can be identified empirically (e.g., using PEPSCAN [Geysen *et al.* (1984) *PNAS USA* 81:3998-4002; Carter (1994) *Methods Mol. Biol.* 36:207-223, or similar methods], or they can be predicted (*e.g.*, using the Jameson-Wolf antigenic index [Jameson, BA *et al.* 1988, *CABIOS* 4(1):1818-186], matrix-based approaches [Raddrizzani and Hammer (2000) *Brief Bioinform.* 1(2):179-189], TEPITOPE [De Lalla *et al.* (199) *J. Immunol.* 163:1725-1729], neural networks [Brusic *et al.* (1998) *Bioinformatics* 14(2):121-130], OptiMer & EpiMer [Meister *et al.* (1995) *Vaccine* 13(6):581-591; Roberts *et al.* (1996) *AIDS Res. Hum. Retroviruses* 12(7):593-610], ADEPT [Maksyutov & Zagrebelnaya (1993) *Comput. Appl. Biosci.* 9(3):291-297], Tsites [Feller & de la Cruz (1991) *Nature* 349(6311):720-721], hydrophilicity [Hopp (1993) *Peptide Research* 6:183-190], antigenic index [Welling *et al.* (1985)*FEBS Lett.* 188:215-218] or the methods disclosed in Davenport *et al.* (1995) *Immunogenetics* 42:392-297, etc. Other preferred fragments include (1) the N-terminal signal peptides of each identified Gram positive bacteria protein, (2) the identified Gram positive bacteria protein without their N-terminal signal peptides, (3) each identified Gram positive bacteria protein wherein up to 10 amino acid residues (e.g. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25 or more) are deleted from the N-terminus and/or the C-terminus *e.g.* the N-terminal amino acid residue may be deleted. Other fragments omit one or more domains of the protein (*e.g.* omission of a signal peptide, of a cytoplasmic domain, of a transmembrane domain, or of an extracellular domain), and (4) the polypeptides, but without their N-terminal amino acid residue.

As indicated in the above text, nucleic acids and polypeptides of the invention may include sequences that:

(a) are identical (*i.e.*, 100% identical) to the sequences disclosed in the sequence listing;

(b) share sequence identity with the sequences disclosed in the sequence listing;

(c) have 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 single nucleotide or amino acid alterations (deletions, insertions, substitutions), which may be at separate locations or may be contiguous, as compared to the sequences of (a) or (b);

(d) when aligned with a particular sequence from the sequence listing using a pairwise alignment algorithm, a moving window of *x* monomers (amino acids or nucleotides)

PCT/US2005/027239

moving from start (N-terminus or 5′) to end (C-terminus or 3′), such that for an alignment that extends to *p* monomers (where *p*>*x*) there are *p*-*x*+1 such windows, each window has at least *x*·*y* identical aligned monomers, where: *x* is slected from 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 150, 200; *y* is selected from 0.50, 0.60, 0.70, 0.75, 0.80, 0.85, 0.90, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99; and if *x*·*y* is not an integer then it is rounded up to the nearest integer. The preferred pairwise alignment algorithm is the Needleman-Wunsch global alignment algorithm [Needlman &Wunsch (1970) *J. Mol. Biol.* 48, 443-453], using default parameters (*e.g.*, with Gap opening penalty = 10.0, and with Gap extension penalty = 0.5, using the EBLOSUM62 scoring matrix). This algorithm is conveniently implemented in the *needle* tool in the EMBOSS package [Rice *et al.* (2000) *Trends Genet.* 16:276-277].

The nucleic acids and polypeptides of the inention may additionally have further sequences to the N-terminus/5′ and/or C-terminus/3′ of these sequences (a) to (d).

All of the Gram positive bacterial sequences referenced herein are publicly available through PubMed on GenBank.

Streptococcus pneumoniae Adhesin Island Sequences

As discussed above, a *S. pneumoniae* AI sequence is present in the TIGR4 *S. pneumoniae* genome. Examples of *S. pneumoniae* AI sequences are set forth below.

SrtD (Sp0468) is a sortase. An example of an amino acid sequence of SrtD is set forth in SEQ ID NO: 80.

SEQ ID NO: 80
MSRTKLRALLGYLLMLVACLIPIYCFGQMVLQSLGQVKGHATFVKSMTTEMYQEQQNHSLAYNQRLASQNRIVDP
FLAEGYEVNYQVSDDPDAVYGYLSIPSLEIMEPVYLGADYHHLGMGLAHVDGTPLPLDGTGIRSVIAGHRAEPSH
VFFRHLDQLKVGDALYYDNGQEIVEYQMMDTEIILPSEWEKLESVSSKNIMTLITCDPIPTFNKRLLVNFERVAV
YQKSDPQTAAVARVAFTKEGQSVSRVATSQWLYRGLVVLAFLGILFVLWKLARLLRGK

SrtC (Sp0467) is a sortase. An example of an amino acid sequence of SrtC is set forth in SEQ ID NO: 81.

SEQ ID NO: 81
MSRYYYRIESNEVIKEFDETVSQMDKAELEERWRLAQAFNATLKPSEILDPFTEQEKKKGVSEYANMLKVHERIG
YVEIPAIDQEIPMYVGTSEDILQKGAGLLEGASLPVGGENTHTVITAHRGLPTAELFSQLDKMKKGDIFYLHVLD
QVLAYQVDQIVTVEPNDFEPVLIQHGEDYATLLTCTPYMINSHRLLVRGKRIPYTAPIAERNRAVRERGQFWLWL
LLGAMAVILLLLYRVYRNRRIVKGLEKQLEGRHVKD

SrtB (SP0466) is a sortase. An example of an amino acid sequence of SrtB is set forth in SEQ ID NO: 82.

SEQ ID NO: 82
MAVMAYPLVSRLYYRVESNQQIADFDKEKATLDEADIDERMKLAQAFNDSLNNVVSGDPWSEEMKKKGRAEYARM
LEIHERMGHVEIPVIDVDLPVYAGTAEEVLQQGAGHLEGTSLPIGGNSTHAVITAHTGLPTAKMFTDLTKLKVGD
KFYVHNIKEVMAYQVDQVKVIEPTNFDDLLIVPGHDYVTLLTCTPYMINTHRLLVRGHRIPYVAEVEEEFIAANK
LSHLYRYLFYVAVGLIVILLWIIRRLRKKKKQPEKALKALKAARKEVKVEDGQQ

Sp0465 is a hypothetical protein. An example of an amino acid sequence of Sp0465 is set forth in SEQ ID NO: 83.

SEQ ID NO: 83
MFLPFLSASLYLQTHHFIAFPNRQSYLLRETRKSHFFLIHHPF

RrgC (SP0464) is a cell wall surface anchor family protein. RrgC contains a sortase substrate motif VPXTG (SEQ ID NO: 137), shown in italics in SEQ ID NO: 84.

SEQ ID NO: 84
MISRIFFVMALCFSLVWGAHAVQAQEDHTLVLQLENYQEVVSQLPSRDGHRLQVWKLDDSYSYDDRVQIVRDLHS
WDENKLSSFKKTSFEMTFLENQIEVSHIPNGLYYVRSIIQTDAVSYPAEFLFEMTDQTVEPLVIVAKKTDTMTTK
VKLIKVDQDHNRLEGVGFKLVSVARDVSEKEVPLIGEYRYSSSGQVGRTLYTDKNGEIFVTNLPLGNYRFKEVEP
LAGYAVTTLDTDVQLVDHQLVTITVVNQKLPRGNVDFMKVDGRTNTSLQGAMFKVMKEESGHYTPVLQNGKEVVV
TSGKDGRFRVEGLEYGTYYLWELQAPTGYVQLTSPVSFTIGKDTRKELVTVVKNNKRPRID*VPDTG*EETLYILML
VAILLFGSGYYLTKKPNN

RrgB (Sp0463) is a cell wall surface anchor protein. RrgB contains a sortase substrate motif IPXTG (SEQ ID NO: 133), shown in italics in SEQ ID NO: 85.

SEQ ID NO: 85
MKSINKFLTMLAALLLTASSLFSAATVFAAGTTTTSVTVHKLLATDGDMDKIANELETGNYAGNKVGVLPANAKE
IAGVMFVWTNTNNEIIDENGQTLGVNIDPQTFKLSGAMPATAMKKLTEAEGAKFNTANLPAAKYKIYEIHSLSTY
VGEDGATLTGSKAVPIEIELPLNDVVDAHVYPKNTEAKPKIDKDFKGKANPDTPRVDKDTPVNHQVGDVVEYEIV
TKIPALANYATANWSDRMTEGLAFNKGTVKVTVDDVALEAGDYALTEVATGFDLKLTDAGLAKVNDQNAEKTVKI
TYSATLNDKAIVEVPESNDVTFNYGNNPDHGNTPKPNKPNENGDLTLTKTWVDATGAPIPAGAEATFDLVNAQTG
KVVQTVTLTTDKNTVTVNGLDKNTEYKFVERSIKGYSADYQEITTAGEIAVKNWKDENPKPLDPTEPKVVTYGKK
FVKVNDKDNRLAGAEFVIANADNAGQYLARKADKVSQEEKQLVVTTKDALDRAVAAYNALTAQQQTQQEKEKVDK
AQAAYNAAVIAANNAFEWVADKDNENVVKLVSDAQGRFEITGLLAGTYYLEETKQPAGYALLTSRQKFEVTATSY
SATGQGIEYTAGSGKDDATKVVNKKIT*IPQTG*GIGTIIFAVAGAAIMGIAVYAYVKNNKDEDQLA

RrgA (Sp0462) is a cell wall surface anchor protein. RrgA contains a sortase substrate motif YPXTG (SEQ ID NO: 186), indicated in italics in SEQ ID NO: 86.

SEQ ID NO: 86
MLNRETHMKKVRKIFQKAVAGLCCISQLTAFSSIVALAETPETSPAIGKVVIKETGEGGALLGDAVFELKNNTDG
TTVSQRTEAQTGEAIFSNIKPGTYTLTEAQPPVGYKPSTKQWTVEVEKNGRTTVQGEQVENREEALSDQYPQTGT
YPDVQTPYQIIKVDGSEKNGQHKALNPNPYERVIPEGTLSKRIYQVNNLDDNQYGIELTVSGKTVYEQKDKSVPL
DVVILLDNSNSMSNIRNKNARRAERAGEATRSLIDKITSDSENRVALVTYASTIFDGTEFTVEKGVADKNGKRLN
DSLFWNYDQTSFTTNTKDYSYLKLTNDKNDIVELKNKVPTEAEDHDGNRLMYQFGATFTQKALMKADEILTQQAR
QNSQKVIFHITDGVPTMSYPINFNHATFAPSYQNQLNAFFSKSPNKDGILLSDFITQATSGEHTIVRGDGQSYQM
FTDKTVYEKGAPAAFPVKPEKYSEMKAAGYAVIGDPINGGYIWLNWRESILAYPFNSNTAKITNHGDPTRWYYNG
NIAPDGYDVFTVGIGINGDPGTDEATATSFMQSISSKPENYTNVTDTTKILEQLNRYFHTIVTEKKSIENGTITD
PMGELIDLQLGTDGRFDPADYTLTANDGSRLENGQAVGGPQNDGGLLKNAKVLYDTTEKRIRVTGLYLGTDEKVT
LTYNVRLNDEFVSNKFYDTNGRTTLHPKEVEQNTVRDFPIPKIRDVRKYPEITISKEKKLGDIEFIKVNKNDKKP
LRGAVFSLQKQHPDYPDIYGAIDQNGTYQNVRTGEDGKLTFKNLSDGKYRLFENSEPAGYKPVQNKPIVAFQIVN
GEVRDVTSIVPQDIPAGYEFTNDKHYITNEPIPPKRE*YPRTG*GIGMLPFYLIGCMMMGGVLLYTRKHP

RlrA (Sp0461) is a transcriptional regulator. An example of an amino acid sequence for RlrA is set forth in SEQ ID NO: 87.

SEQ ID NO: 87
MLNKYIEKRITDKITILNILLDIRSIELDELSTLTSLQSKSLLSILQELQETFEEELTFNLDTQQVQLIEHHSHQ
TNYYFHQLYNQSTILKILRFFLLQGNQSFNEFTQKEYISIATGYRVRQKCGLLLRSVGLDLVKNQVVGPEYRIRF
LIALLQFHFGIEIYDLNDGSMDWVTHMIVQSNSQLSHELLEITPDEYVHFSILVALTWKRREFPLEFPESKEFEK
LKNLFMYPILMEHCQTYLEPHANMTFTQEELDYIFLVYCSANSSFSKDKWNQEKKTHTIQLILQHTRGKHLLSKF
KNILGNDISNSLSFLTALTFLTRTFLFGLQNLVPYYNYYEHYGIESDKPLYHISKAIVQEWMTEQKIEGVIDQHR
LYLFSLYLTETIFSSLPAIPIFIILNNQADVNLIKSIILRNFTDKVASVTGYNILISPPPSEEHLTEPLIIITTK
EYLPYVKKQYPKGKHHFLTIALDLHVSQQRLIYQTIVDIRKEAFDKRVAMIAKKAHYLL

As discussed above, a *S. pneumoniae* AI sequence is present in the *S. pneumoniae* strain 670 genome. Examples of *S. pneumoniae* AI sequences are set forth below.

PCT/US2005/027239

Orf1_670 is a transposase. An example of an amino acid sequence of orf1_670 is set forth in SEQ ID NO: 171.

SEQ ID NO: 171

MEHINHTTLLIGIKDKNITLNKAIQHDTHIEVFATLDYHPPKCKHCKGKQIKYDFQKPSKIPFIEIGGFPSLIHL
KKRRFQCKSCRKVTVAETTLVQKNCQISEMVRQKIAQLLLNREALTHIASKLAISTSTSTVYRKLKQFHFQEDYT
TLPEILSWDEFSYQKGKLAFIAQDFNTKKIMTILDNRRQTTIRNHFFKYSKEARKKVKVVTVDMSGSYIPLIKKL
FPNAKIVLDRFHIVQHMSRALNQTRINIMKQFDDKSLEYRALKYYWKFILKDSRKLSLKPFYARTFRETLTPREC
LKKIFTLVPELKDYYDLYQLLLFHLQEKNTDQFWGLIQDTLPHLNRTFKTTLSTFICYKNYITNAIELPYSNAKL
EATNKLIKDIKRNAFGFRNFENFKKRIFIALNIKKERTKFVLSRA

Orf2_670 is a transcriptional regulator. An example of an amino acid sequence of Orf2_670 is set forth in SEQ ID NO: 172.

SEQ ID NO: 172

MLNKYIEKRITDKITILNILLDIRSIELDELSTLTSLQSKSLLSILQELQETFEEELTFNLDTQQVQLIEHHSHQ
TNYYFHQLYNQSTILKILRFFLLQGNQSFNEFTQKEYISIATGYRVRQKCGLLLRSVGLDLVKNQVVGPEYRIRF
LIALLQFHFGIEIYDLNDGSMDWVTHMIVQSNSQLSHELLEITPDEYVHFSILVALTWKRREFPLEFPESKEFEK
LKNLFMYPILMEHCQTYLEPHANMTFTQEELDYIFLVYCSANSSFSKDKWNQEKKTHTIQLILQHTRGKHLLSKF
KNILGNDISNSLSFLTALTFLTRTFLFGLQNLVPYYNYYEHYGIESDKPLYHISKAIVQEWMTEQKIEGVIDQHR
LYLFSLYLTETIFSSLPAIPIFIILNNQADVNLIKSIILRNFTDKVASVTGYNILISPPPSEEHLTEPLIIITTK
EYLPYVKKQYPKGKHHFLTIALDLHVSQQRLIYQTIVDIRKEAFDKRVAMIAKKAHYLL

Orf3_670 is a cell wall surface anchor family proten. An example of an amino acid sequence of Orf3_670 is set forth in SEQ ID NO: 173.

SEQ ID NO: 173

MLNRETHMKKVRKIFQKAVAGLCCISQLTAFSSIVALAETPETSPAIGKVVIKETGEGGALLGDAVFELKNNTDG
TTVSQRTEAQTGEAIFSNIKPGTYTLTEAQPPVGYKPSTKQWTVEVEKNGRTTVQGEQVENREEALSDQYPQTGT
YPDVQTPYQIIKVDGSEKNGQHKALNPNPYERVIPEGTLSKRIYQVNNLDDNQYGIELTVSGKTTVETKEASTPL
DVVILLDNSNSMSNIRHNHAHRAEKAGEATRALVDKITSNPDNRVALVTYGSTIFDGSEATVEKGVADANGKILN
DSALWTFDRTTFTAKTYNYSFLNLTSDPTDIQTIKDRIPSDAEELNKDKLMYQFGATFTQKALMTADDILTKQAR
PNSKKVIFHITDGVPTMSYPINFKYTGTTQSYRTQLNNFKAKTPNSSGILLEDFVTWSADGEHKIVRGDGESYQM
FTKKPVTDQYGVHQILSITSMEQRAKLVSAGYRFYGTDLYLYWRDSILAYPFNSSTDWITNHGDPTTWYYNGNMA
QDGYDVFTVGVGVNGDPGTDEATATRFMQSISSSPDNYTNVADPSQILQELNRYFYTIVNEKKSIENGTITDPMG
ELIDFQLGADGRFDPADYTLTANDGSSLVNNVPTGGPQNDGGLLKNAKVFYDTTEKRIRVTGLYLGTGEKVTLTY
NVRLNDQFVSNKFYDTNGRTTLHPKEVEKNTVRDFPIPKIRDVRKYPEITIPKEKKLGEIEFIKINKNDKKPLRD
AVFSLQKQHPDYPDIYGAIDQNGTYQNVRTGEDGKLTFKNLSDGKYRLFENSEPAGYKPVQNKPIVAFQIVNGEV
RDVTSIVPQDIPAGYEFTNDKHYITNEPIPPKREYPRTGGIGMLPFYLIGCMMMGGVLLYTRKHP

Orf4_670 is a cell wall surface anchor family protein. An example of an amino acid sequence of orf4_670 is set forth in SEQ ID NO: 174.

SEQ ID NO: 174

MKSINKFLTMLAALLLTASSLFSAATVFAADNVSTAPDAVTKTLTIHKLLLSEDDLKTWDTNGPKGYDGTQSSLK
DLTGVVAEEIPNVYFELQKYNLTDGKEKENLKDDSKWTTVHGGLTTKDGLKIETSTLKGVYRIREDRTKTTYVGP
NGQVLTGSKAVPALVTLPLVNNNGTVIDAHVFPKNSYNKPVVDKRIADTLNYNDQNGLSIGTKIPYVVNTTIPSN
ATFATSFWSDEMTEGLTYNEDVTITLNNVAMDQADYEVTKGNNGFNLKLTEAGLAKINGKDADQKIQITYSATLN
SLAVADIPESNDITYHYGNHQDHGNTPKPTKPNNGQITVTKTWDSQPAPEGVKATVQLVNAKTGEKVGAPVELSE
NNWTYTWSGLDNSIEYKVEEEYNGYSAEYTVESKGKLGVKNWKDNNPAPINPEEPRVKTYGKKFVKVDQKDTRLE
NAQFVVKKADSNKYIAFKSTAQQAADEKAAATAKQKLDAAVAAYTNAADKQAAQALVDQAQQEYNVAYKEAKFGY
VEVAGKDEAMVLTSNTDGQFQISGLAAGTYKLEEIKAPEGFAKIDDVEFVVGAGSWNQGEFNYLKDVQKNDATKV
VNKKITIPQTGGIGTIIFAVAGAAIMGIAVYAYVKNNKDEDQLA

Orf5_670 is a cell wall surface anchor family protein. An example of an amino acid sequence of orf5_670 is set forth in SEQ ID NO: 175.

SEQ ID NO: 175

MTMQKMQKMISRIFFVMALCFSLVWGAHAVQAQEDHTLVLQLENYQEVVSQLPSRDGHRLQVWKLDDSYSYDDRV
QIVRDLHSWDENKLSSFKKTSFEMTFLENQIEVSHIPNGLYYVRSIIQTDAVSYPAEFLFEMTDQTVEPLVIVAK

DEMANDE OU BREVET VOLUMINEUX

LA PRÉSENTE PARTIE DE CETTE DEMANDE OU CE BREVET COMPREND PLUS D'UN TOME.

CECI EST LE TOME 1 DE 2
CONTENANT LES PAGES 1 À 202

NOTE : Pour les tomes additionels, veuillez contacter le Bureau canadien des brevets

JUMBO APPLICATIONS/PATENTS

THIS SECTION OF THE APPLICATION/PATENT CONTAINS MORE THAN ONE VOLUME

THIS IS VOLUME 1 OF 2
CONTAINING PAGES 1 TO 202

NOTE: For additional volumes, please contact the Canadian Patent Office

NOM DU FICHIER / FILE NAME :

NOTE POUR LE TOME / VOLUME NOTE:

WE CLAIM:

1. An immunogenic composition comprising a purified Group B Streptococcus (GBS) adhesin island (AI) polypeptide in oligomeric form.

2. The immunogenic composition of claim 1 wherein the GBS AI polypeptide is selected from a GBS AI-1.

3. The immunogenic composition of claim 1 wherein the GBS AI polypeptide is selected from a GBS AI-2.

4. The immunogenic composition of claim 2 wherein the GBS AI polypeptide is selected from the group consisting of GBS 80, GBS 104, GBS 52, and fragments thereof.

5. The immunogenic composition of claim 3 wherein the GBS AI polypeptide is selected from the group consisting of GBS 59, GBS 67, GBS 150, 01521, 01523, 01524, and fragments thereof.

6. The immunogenic composition of claim 4 wherein the GBS AI polypeptide is GBS 80.

7. The immunogenic composition of any of claims 1-6 wherein the oligomeric form is a hyperoligomer.

8 (22). An immunogenic composition comprising a purified Gram positive bacteria adhesin island (AI) polypeptide in an oligomeric form.

9 (23). The immunogenic composition of claim 8 wherein the Gram positive bacteria is of a genus selected from the group consisting of *Streptococcus*, *Enterococcus*, *Staphylococcus*, *Clostridium*, *Corynebacterium*, or *Listeria*.

10 (24). The immunogenic composition of claim 9 wherein the Gram positive bacteria is of the genus *Streptococcus*.

11 (35). The immunogenic composition of claim 10 wherein the genus *Streptococcus* bacteria is Group A Streptococcus (GAS) bacteria and the Gram positive bacteria AI polypeptide is a GAS AI polypeptide.

12 (36). The immunogenic composition of claim 11 wherein the GAS AI polypeptide is selected from a GAS AI-1.

13 (37). The immunogenic composition of claim 11 wherein the GAS AI polypeptide is selected from a GAS AI-2.

14 (38). The immunogenic composition of claim 11 wherein the GAS AI polypeptide is selected from a GAS AI-3.

15 (39). The immunogenic composition of claim 11 wherein the GAS AI polypeptide is selected from a GAS AI-4.

16 (66). The immunogenic composition of any one of claims 8-15 wherein the oligomeric form is a hyperoligomer.

17. An immunogenic composition comprising a first and a second Group B Streptococcus (GBS) adhesin island (AI) polypeptide.

18. The immunogenic composition of claim 17 wherein the first GBS AI polypeptide is encoded by a GBS AI-1.

19. The immunogenic composition of claim 18 wherein the second GBS AI polypeptide is encoded by a GBS AI-2.

20. The immunogenic composition of claim 18 wherein the first GBS AI polypeptide is selected from the group consisting of GBS 80, GBS 104, GBS 52, and fragments thereof.

21. The immunogenic composition of claim 19 wherein the second GBS AI polypeptide is selected from the group consisting of GBS 59, GBS 67, GBS 150, 01521, 01523, 01524, and fragments thereof, and wherein the first and the second GBS AI polypeptide are not the same polypeptide.

22. The immunogenic composition of claim 19 wherein the first GBS AI polypeptide is GBS 80 and the second GBS AI polypeptide is GBS 67.

23. An immunogenic composition comprising a first and a second Gram positive bacteria adhesin island (AI) polypeptide.

24. The immunogenic composition of claim 23 wherein the Gram positive bacteria is *Streptococcus*, *Enterococcus*, *Staphylococcus*, *Clostridium*, *Corynebacterium*, or *Listeria*.

25. The immunogenic composition of claim 23 wherein the first Gram positive bacteria AI polypeptide is a first Group A Streptococcus (GAS) AI polypeptide.

26. The immunogenic composition of claim 25 wherein the first GAS AI polypeptide is a first GAS AI-1 polypeptide.

27. The immunogenic composition of claim 25 wherein the first GAS AI polypeptide is a first GAS AI-2 polypeptide.

28. The immunogenic composition of claim 25 wherein the first GAS AI polypeptide is a first GAS AI-3 polypeptide.

29. The immunogenic composition of claim 25 wherein the first GAS AI polypeptide is a first GAS AI-4 polypeptide.

30. The immunogenic composition of any one of claims 25-29 wherein the second Gram positive bacteria AI polypeptide is a second GAS AI polypeptide.

31. The immunogenic composition of claim 30 wherein the second GAS AI polypeptide is a second GAS AI-1 polypeptide.

32. The immunogenic composition of claim 30 wherein the second GAS AI polypeptide is a second GAS AI-2 polypeptide.

33. The immunogenic composition of claim 30 wherein the second GAS AI polypeptide is a second GAS AI-3 polypeptide.

34. The immunogenic composition of claim 30 wherein the second GAS AI polypeptide is a second GAS AI-4 polypeptide.

35. A modified Gram positive bacterium adapted to produce increased levels of AI surface protein.

36. The modified Gram positive bacterium of claim 35 wherein the AI surface protein is in oligomeric form.

37. The modified Gram positive bacterium of claim 36 wherein the oligomeric form is a hyperoligomer.

38. The modified Gram positive bacterium of any one of claims 35-37 which is a non-pathogenic Gram positive bacterium.

39. The modified Gram positive bacterium of claim 38 wherein the non-pathogenic Gram positive bacterium is *Lactococcus lactis*.

40. A method for manufacturing an oligomeric adhesin island (AI) surface antigen comprising:

culturing a Gram positive bacterium that expresses an oligomeric AI surface antigen and isolating the expressed oligomeric AI surface antigen.

DEMANDE OU BREVET VOLUMINEUX

LA PRÉSENTE PARTIE DE CETTE DEMANDE OU CE BREVET COMPREND PLUS D'UN TOME.

CECI EST LE TOME 1 DE 4
CONTENANT LES PAGES 1 À 152

NOTE : Pour les tomes additionels, veuillez contacter le Bureau canadien des brevets

JUMBO APPLICATIONS/PATENTS

THIS SECTION OF THE APPLICATION/PATENT CONTAINS MORE THAN ONE VOLUME

THIS IS VOLUME 1 OF 4
CONTAINING PAGES 1 TO 152

NOTE: For additional volumes, please contact the Canadian Patent Office

NOM DU FICHIER / FILE NAME :

NOTE POUR LE TOME / VOLUME NOTE:

FIGURE 1: Adhesion Island 1

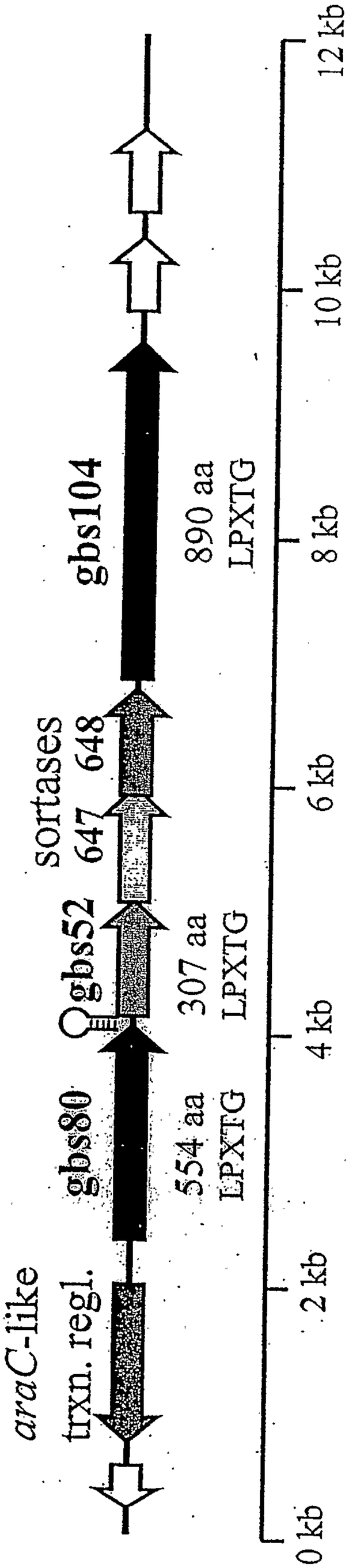


Figure 2: Conservation of AI-1 in GBS serotypes and strain isolates

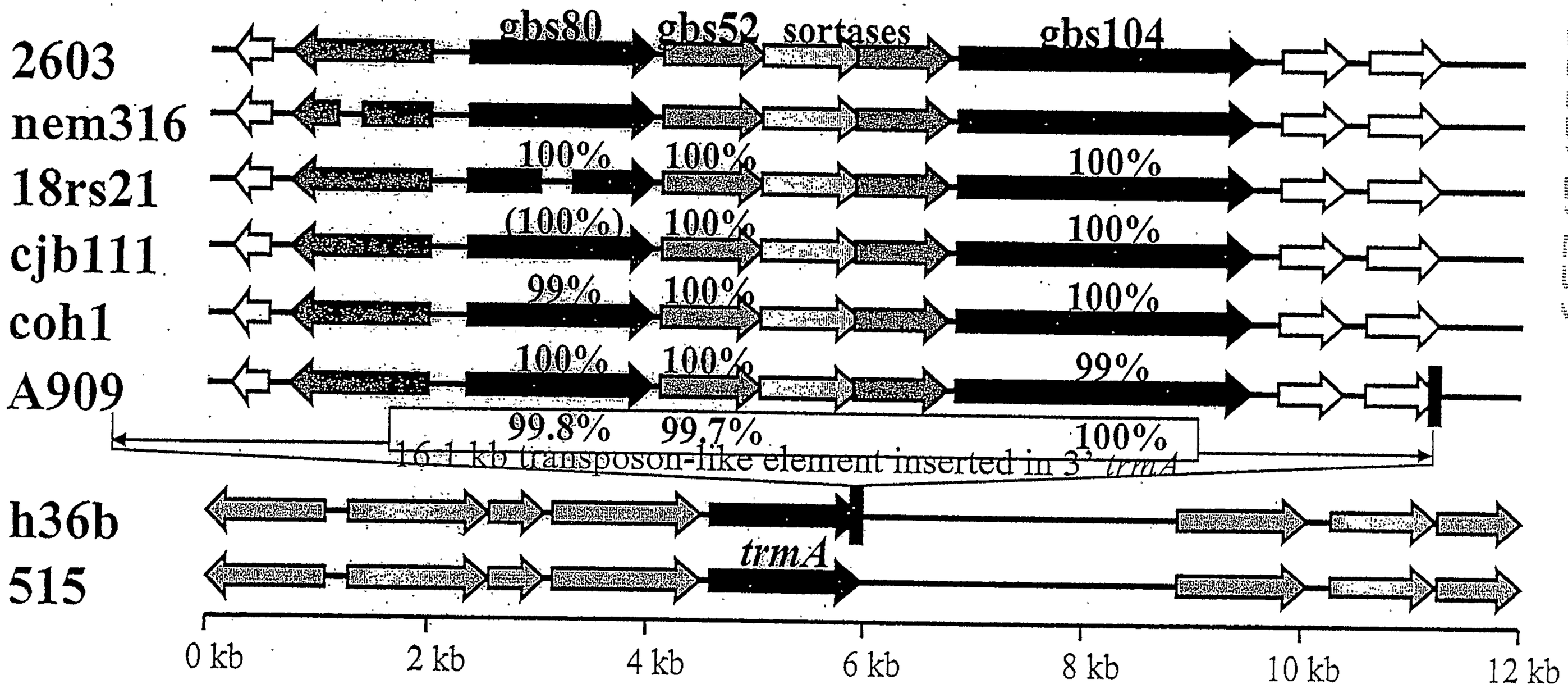
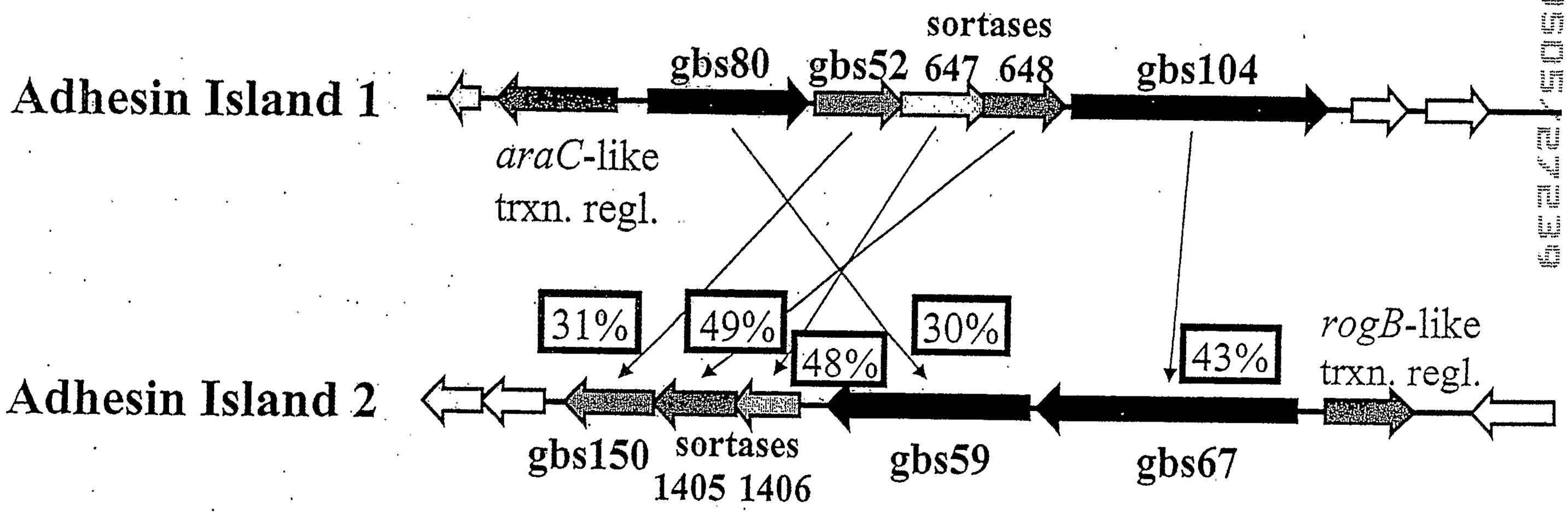
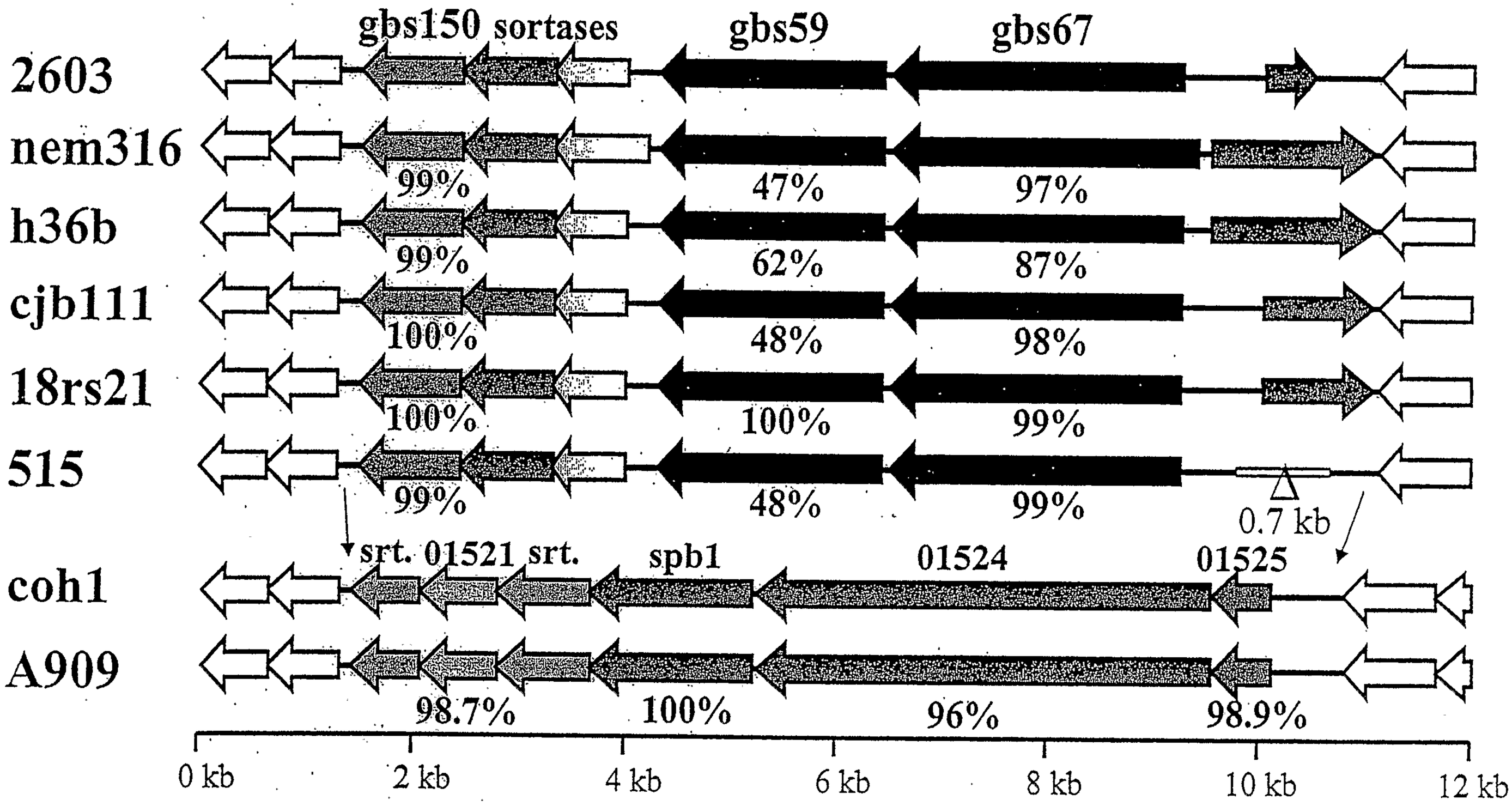


FIGURE 3: Correlation of AI-1 and AI-2 within GBS serotype V, strain isolate 2603 genome



WO 2006/078318 PCT/US2005/027239
3/487
CA 02575548 2007-01-29
PCT/US2005/027239

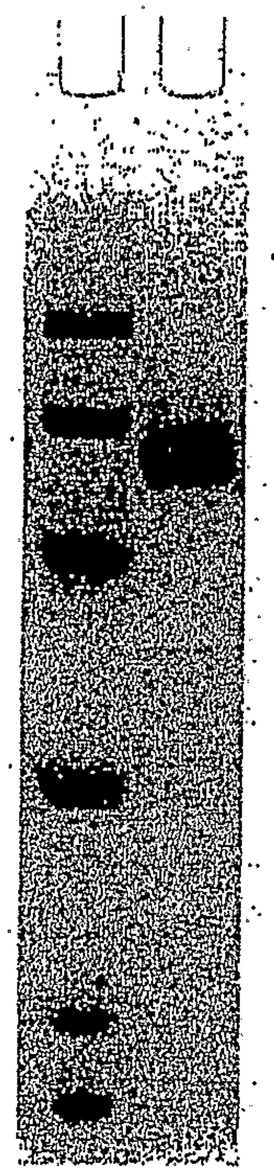
Figure 4: Identification and Variance of AI-2 in Several GBS Serotypes and Strain Isolates



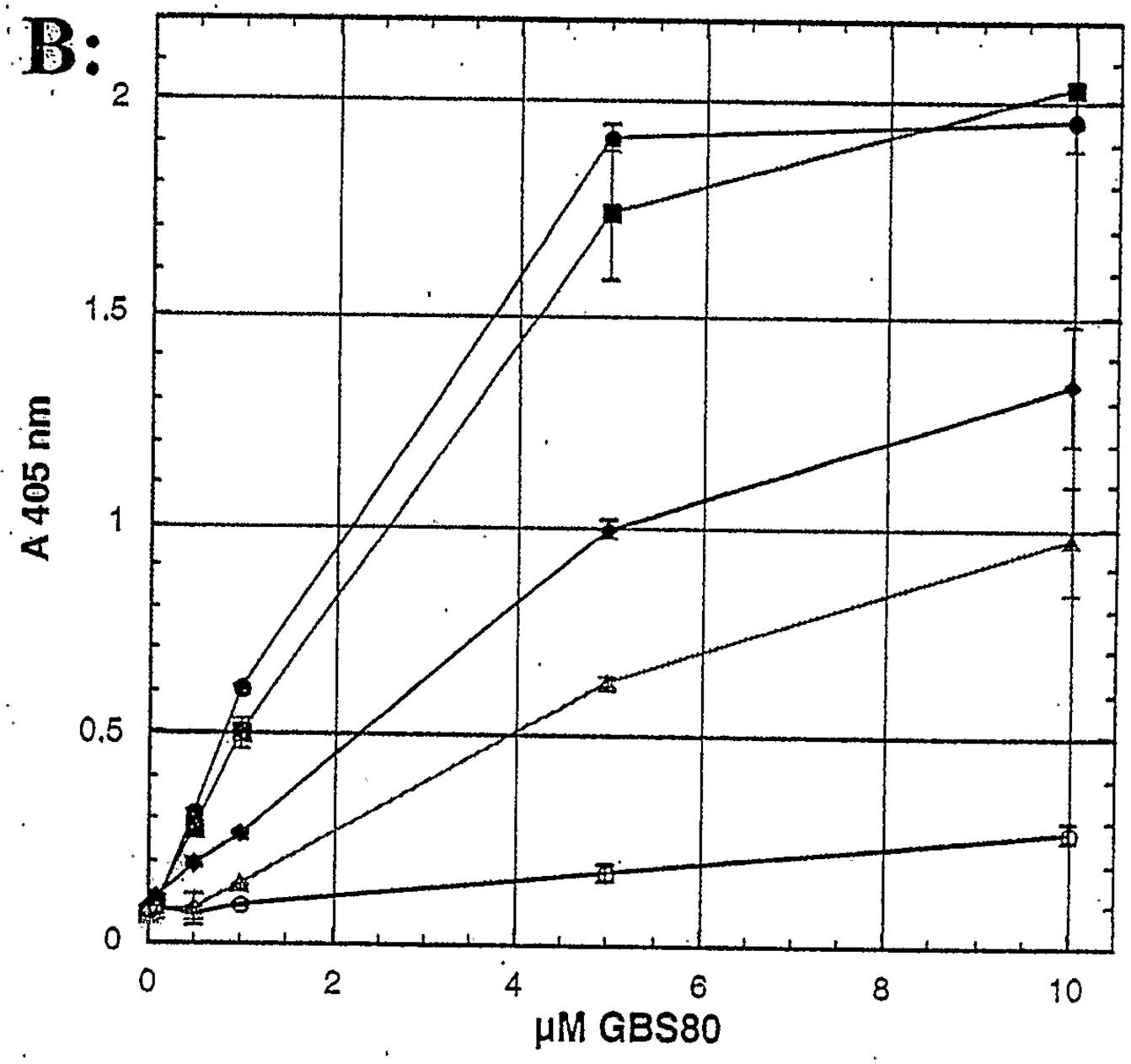
WO 2006/078318 PCT/US2005/027239 4/487
CA 02575548 2007-01-29
PCT/US2005/027239

Figure 5: Purified gbs80 protein binds fibronectin and fibrinogen in an ELISA

A:



B:



WO 2006/078318 PCT/US2005/027239
CA 2575548 2007-01-29
5/487
PCT/US2005/027239

Figure 6: Adhesion Island I is an operon by RT-PCR

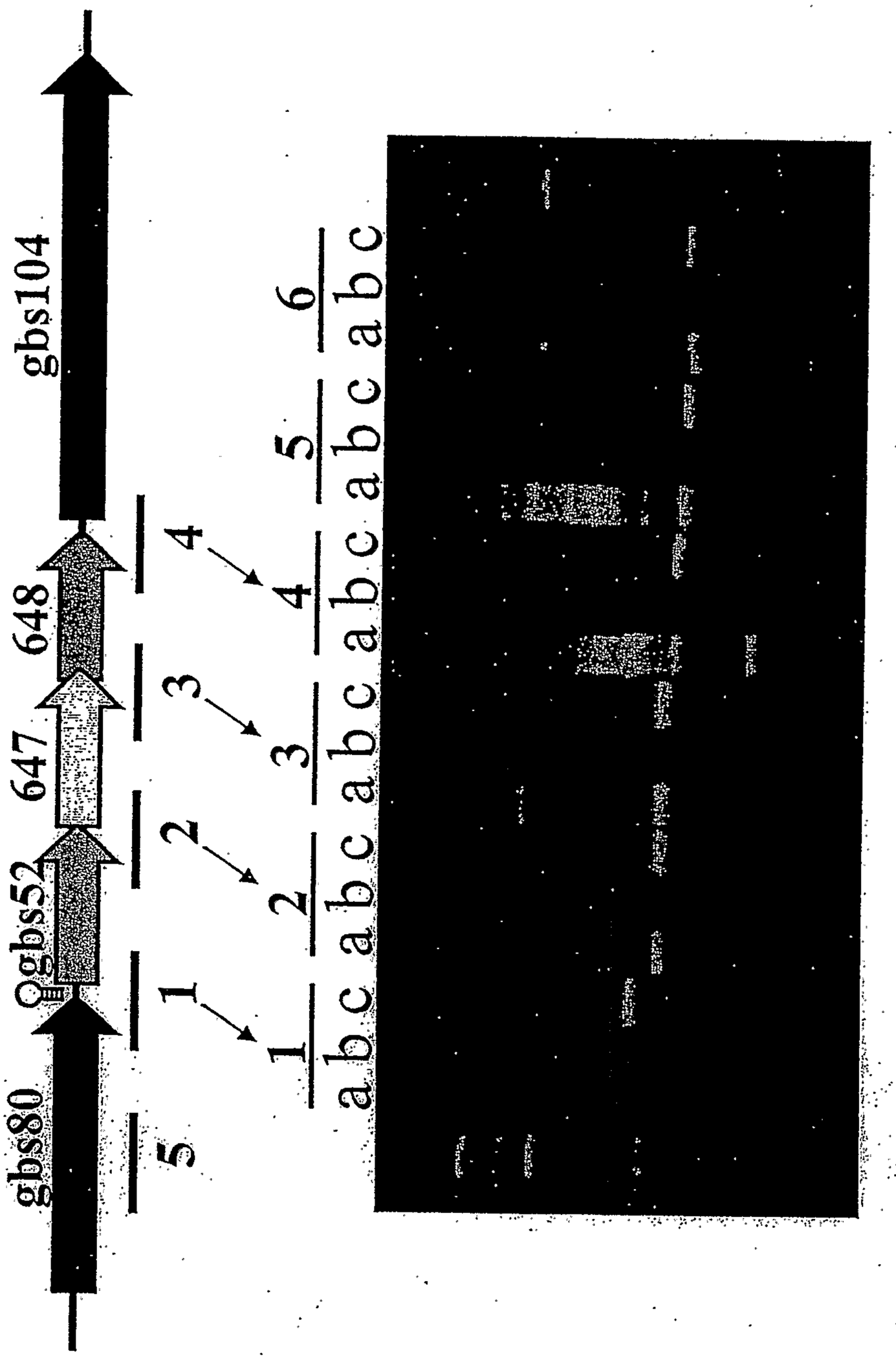
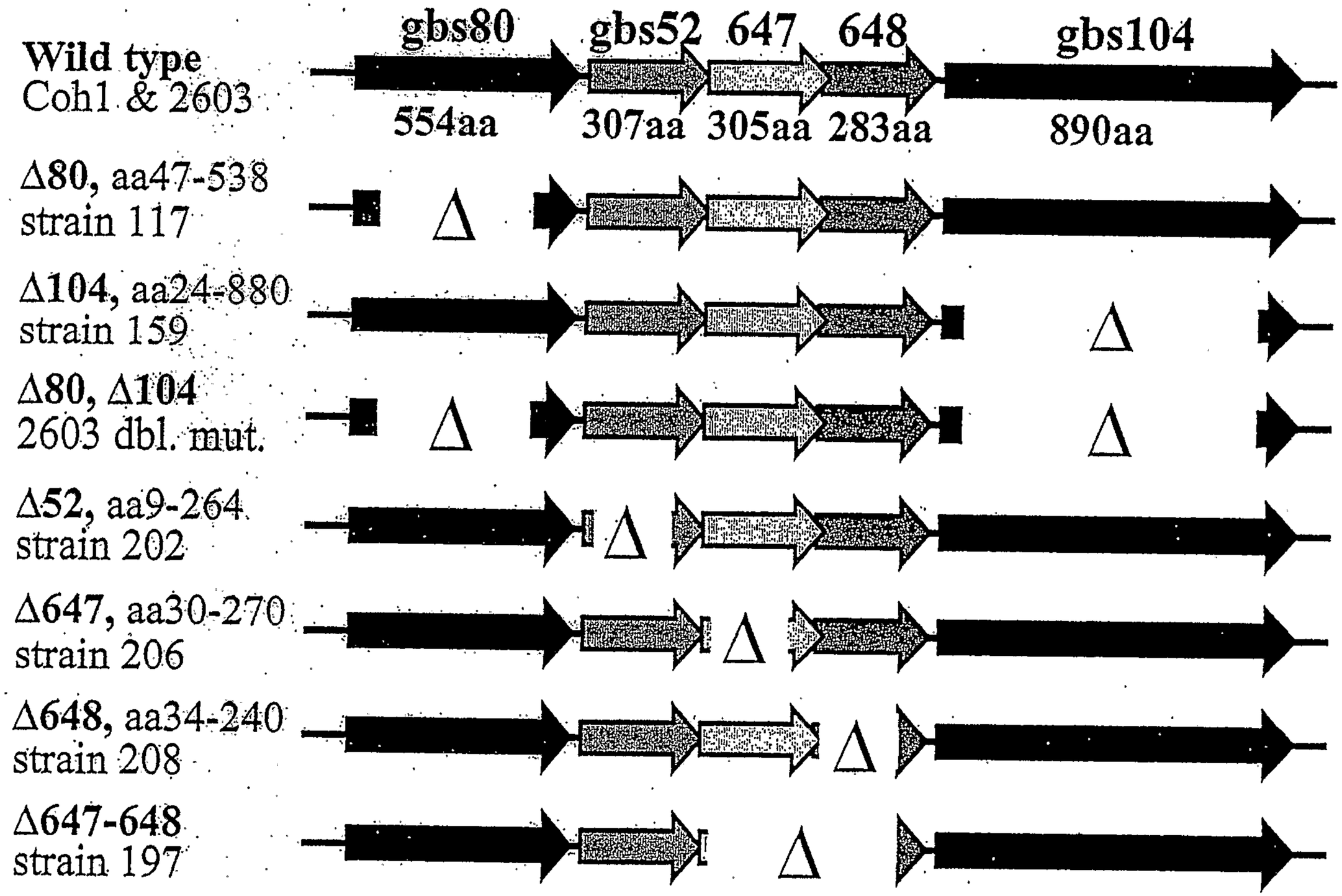


Figure 7: In frame deletions of AI-1 genes



WO 2006/078318 PCT/US2005/027239 7/487 CA 02575548 2007-01-29

Figure 8: gbs80 is required for surface localization of gbs104

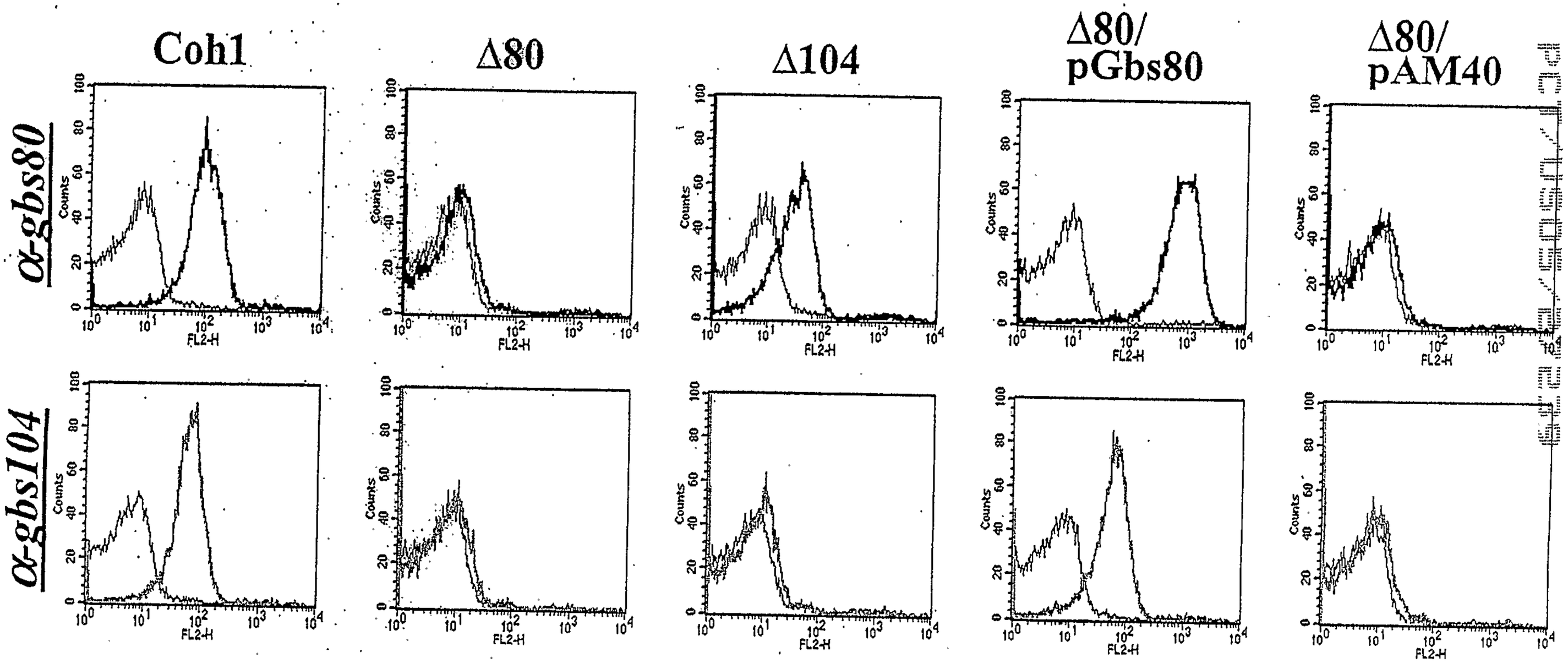


Figure 9: sortases 647 & 648 play a semi-redundant role in surface exposure of gbs80 and gbs104

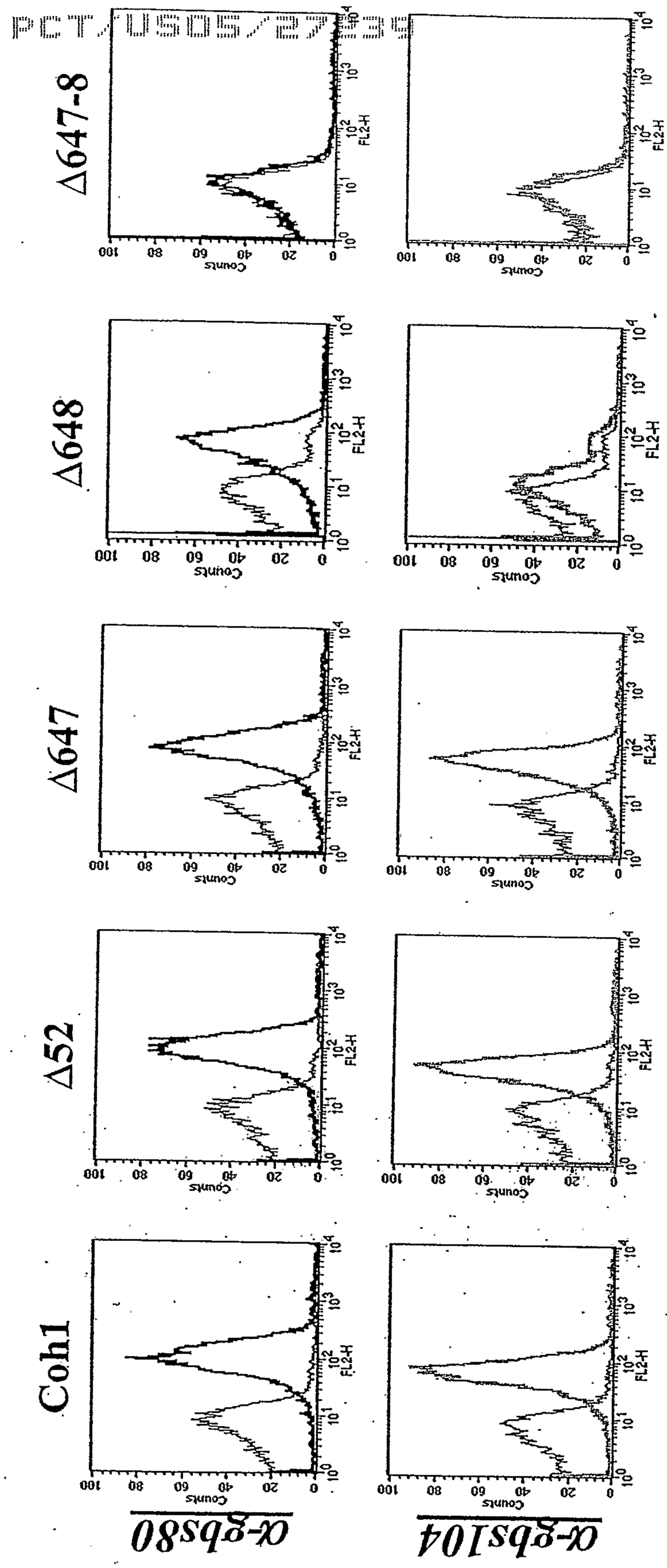
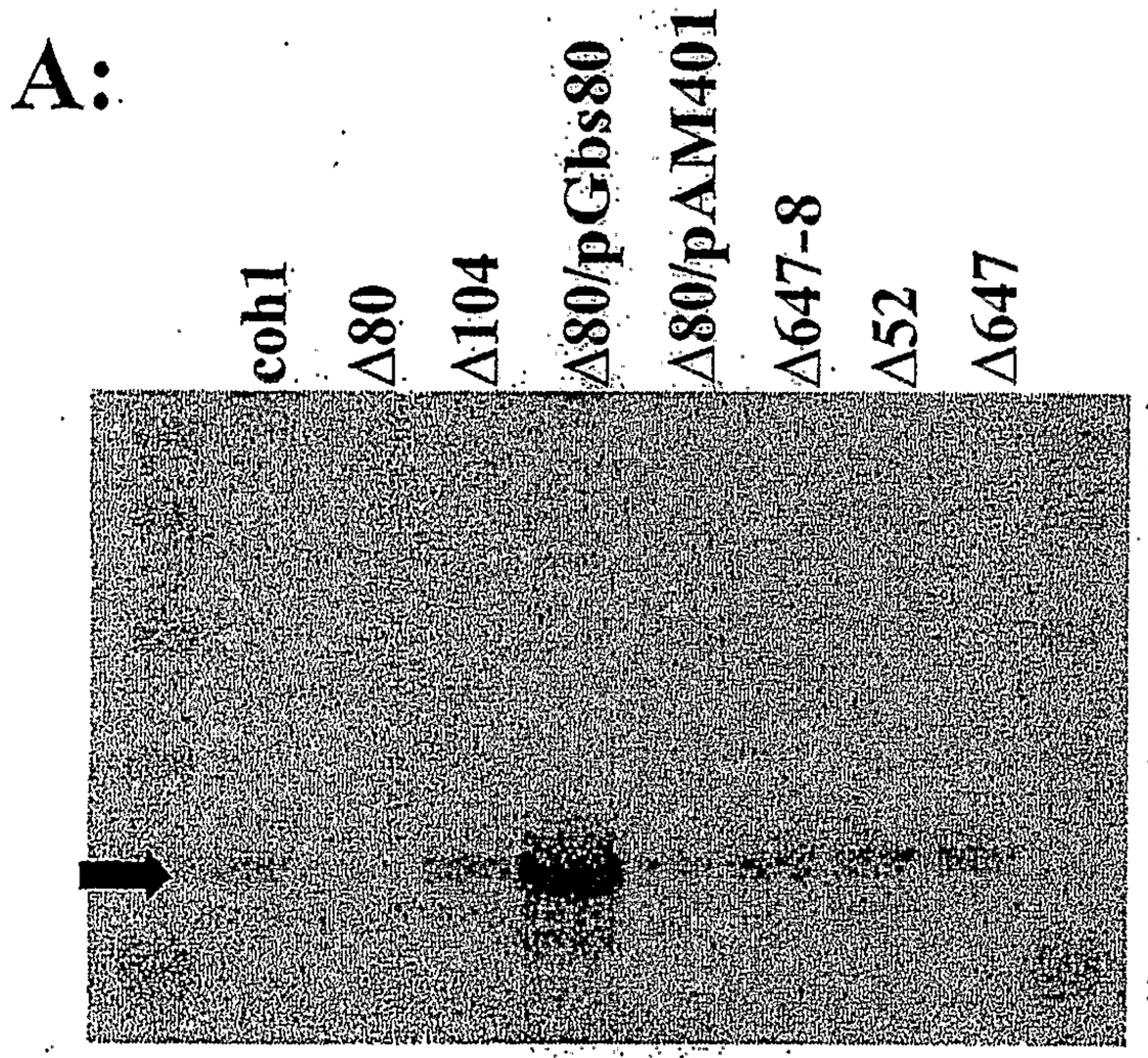
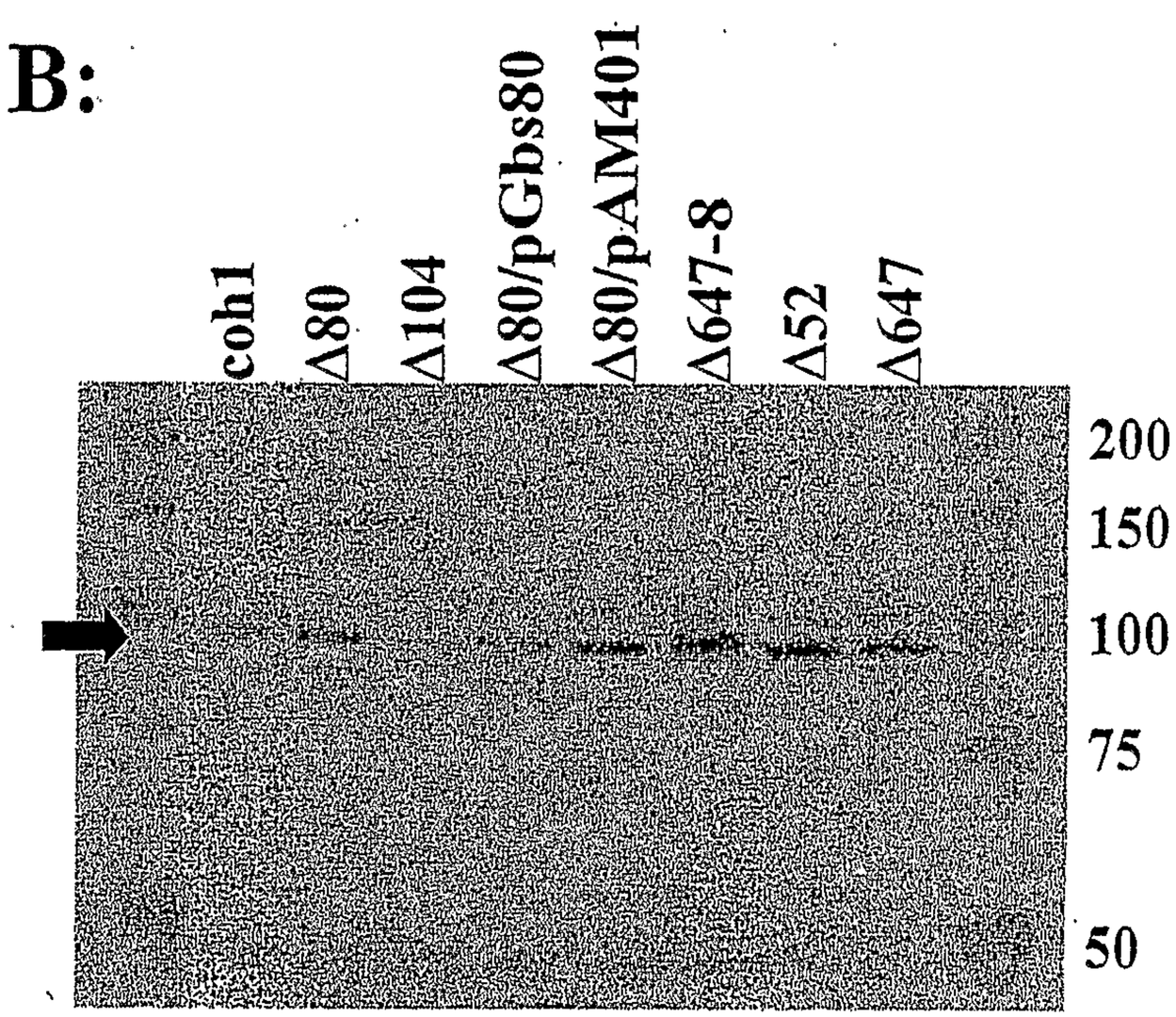


Figure 10: Western blot of mutant strains





WO 2006/078318 PCT/US2005/027239 CA 02575548 2007-01-29 10/487

FIGURE 11: Pre-embedding IEM staining of GBS 80

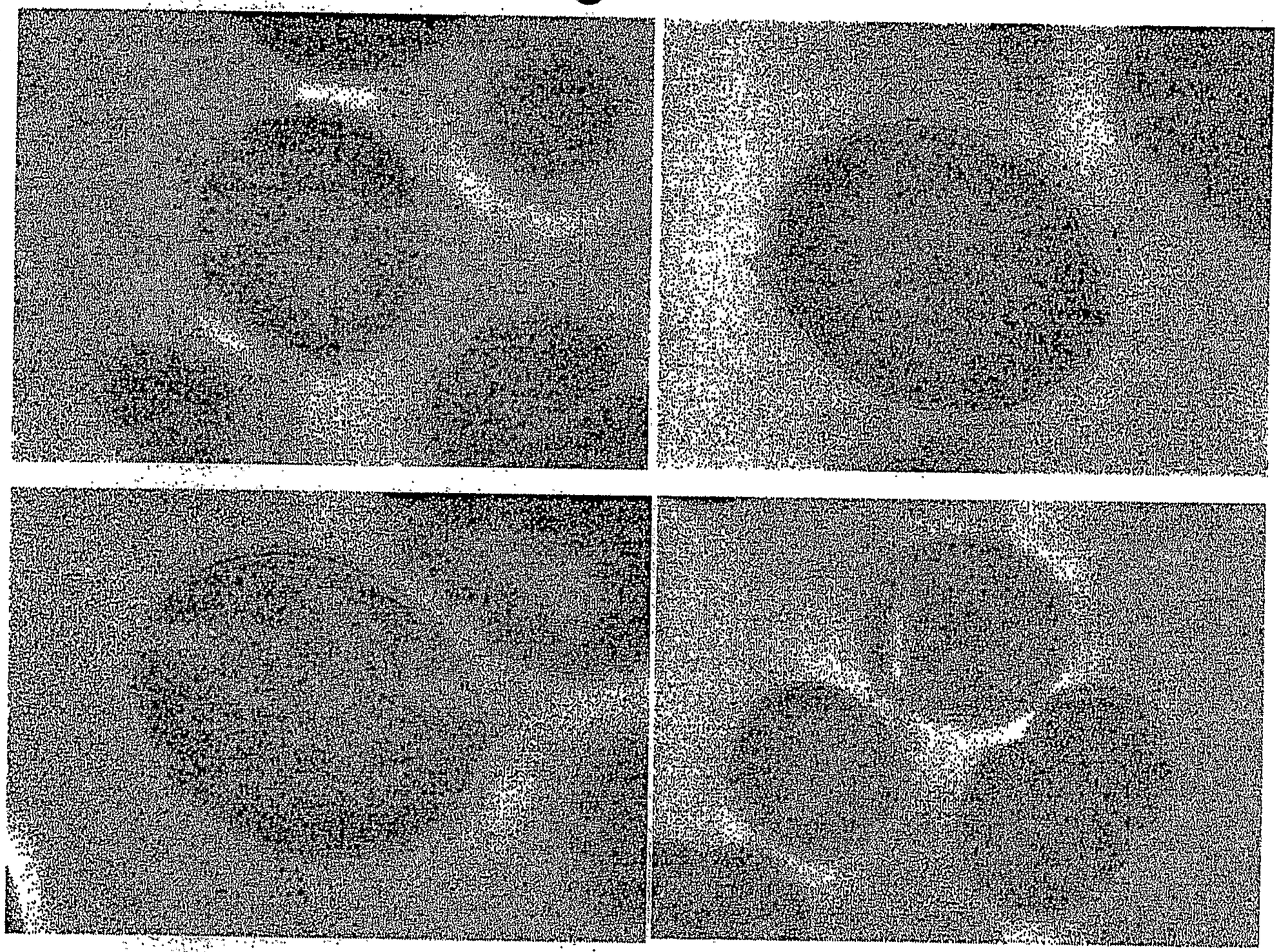
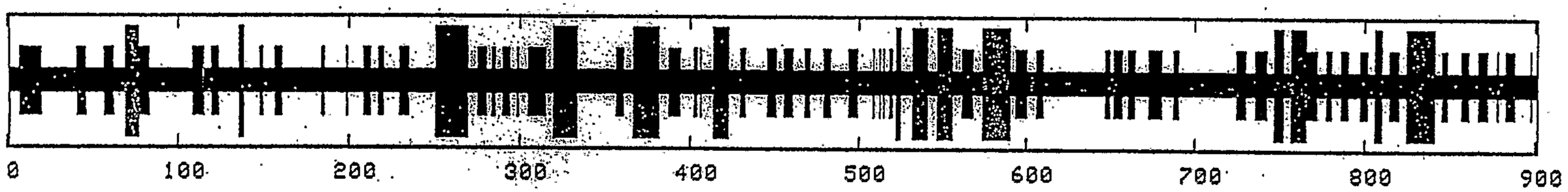


FIGURE 12: Predicted Secondary Structure for GBS 067

PHD SECONDARY STRUCTURE PREDICTION for GBS 067

```
          10        20        30        40        50        60        70
           |         |         |         |         |         |         |
MRKYQKFSKILTLSLFCLSQIPLNTNVLGESTVPENGAKGKLVVKKTDDQNKPLSKATFVLKTTAHPESK
Ccccccceeeee EEecceeeEccccccccccccCCCCCCcceEEEEeCCCCCCCCCcEEEEEecCCCch
IEKVTAELTGEATFDNLIPGDYTLSEETAPEGYKKTNQTWQVKVESNGKTTIQNSGDKNSTIGQNQEELD
hHHHhhheeEeeeecCCCCCCCcCcCCCCccccccCeEEEEEEeCCcEEEEeCCCCCCCCCCchHHhc
KQYPPTGIYEDTKESYKLEHVKGSVPNGKSEAKAVNPYSSEGEHIREIPEGTLSKRISEVGDLAHNKYKI
CCCCCCceeecCccccceEEeccccCCCCccccCCGCCCcccceeccCCCcccccccecccccCCcEE
ELTVSGKTIVKPVDKQKPLDVVFVLDNSNSMNNDGFNFQRHNKAKKAAEALGTAVKDILGANSDNRVALV
EEEccCcEEEecCCCCCCCeEEEEEcCCCCCCCCCCCCCChhHHHHHHHHHHHHHHHHcCcCCeEEEE
TYGSDIFDGRSVDVVKGFKEDDKYYGLQTKFTIQTENYSHKQLTNNAEEIIKRIPTEAPKAKWGSTTNGL
EcccceecCCCeeEEecccecCCCcceeeEEEEEEecCCchhHHHHHHHHHHHhcCCCCCCCCCCCCCCC
TPEQQKEYYLSKVGETFTMKAFMEADDILSQVNRNSQKIIVHVTDGVPTRSYAINNFKLGASYESQFEQM
CccccceEEEecccccchhHHHHHHHHHHHHHHhccCCCeEEEEEeCCCCCcccceeccecCCCCchHHHHHH
KKNGYLNKSNFLLTDKPEDIKGNGESYFLFPLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVKEH
HHhCCccccceeeecCCCCecCCCCceEEeeecCcCeeEEEecCccccceeeccCCCCCCeEEeCCCcCCC
GTPTKLYINSLKQKNYDIFNFGIDISGFRQVYNEEYKKNQDGTFQKLKEEAFKLSDGEITELMRSFSSKP
CCCceEeEEccccccccccecccecccceeCchHHhcccCCcHHHHHHHHHhcccCchHHHHHHHcCCC
EYYTPIVTSADTSNNEILSKIQQQFETILTKENSIVNGTIEDPMGDKINLQLGNGQTLQPSDYTLQGNDG
cceeEEEeecCCCcHHHHHHHHHHHHHHHHhhccceeeeeEecCCCCeeEEecCCCccccCCCCccccCCC
SVMKDGIATGGPNNDGGILKGVKLEYIGNKLYVRGLNLGEGQKVTLTYDVKLDDSFISNKFYDTNGRTTL
ccccccCCCCCCCCCCceeeceEEEEeCCcEEEEEccCCCCCEEEEEEEeecCCCcceeeecCCCCCccC
NPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKLLLKGATFELQEFNEDYKLYL
CCCCCCCCCCCCCCCCCCCcccccCCCeEEEEeCCCCCeEEEEEeCCCChhhHHccCchHHHHhchhhceee
PIKNNNSKVVTGENGKISYKDLKDGKYQLIEAVSPEDYQKITNKPILTFEVVKGSIKNIIAVNKQISEYH
eecCCCCcEEEccCCcEeEEecccCCcEEEEECCCchhHhhCCcCeEEEEecccchhhhhhhhHHHHHHH
EEGDKHLITNTHIPPKGIIPMTGGKGILSFILIGGAMMSIAGGIYIWKRYKKSSDMSIKKD
HhCCceEEeCCCCCCCCeEeecCCCccEEEEecccccceccCCeeEEEccCCCCCCcecCC
Sequence length :   901
PHD :
   Alpha helix     (Hh) :   148 is  16.43%     3_10 helix      (Gg) :     0 is   0.00%
   Pi helix        (Ii) :     0 is   0.00%     Beta bridge     (Bb) :     0 is   0.00%
   Extended strand (Ee) :   243 is  26.97%     Beta turn       (Tt) :     0 is   0.00%
   Bend region     (Ss) :     0 is   0.00%     Random coil     (Cc) :   510 is  56.60%
   Ambigous states (?)  :     0 is   0.00%     Other states         :     0 is   0.00%
```



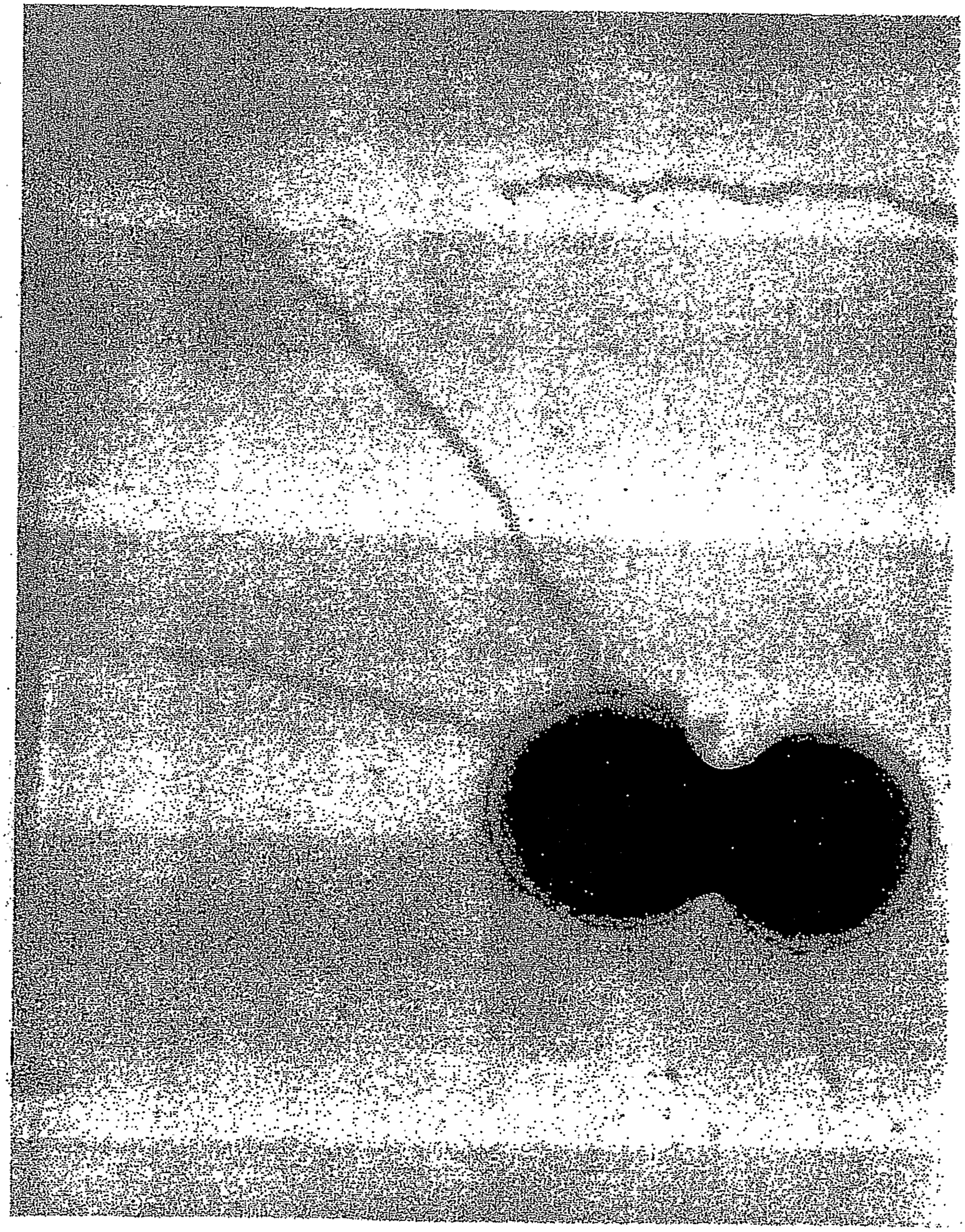


Figure 13

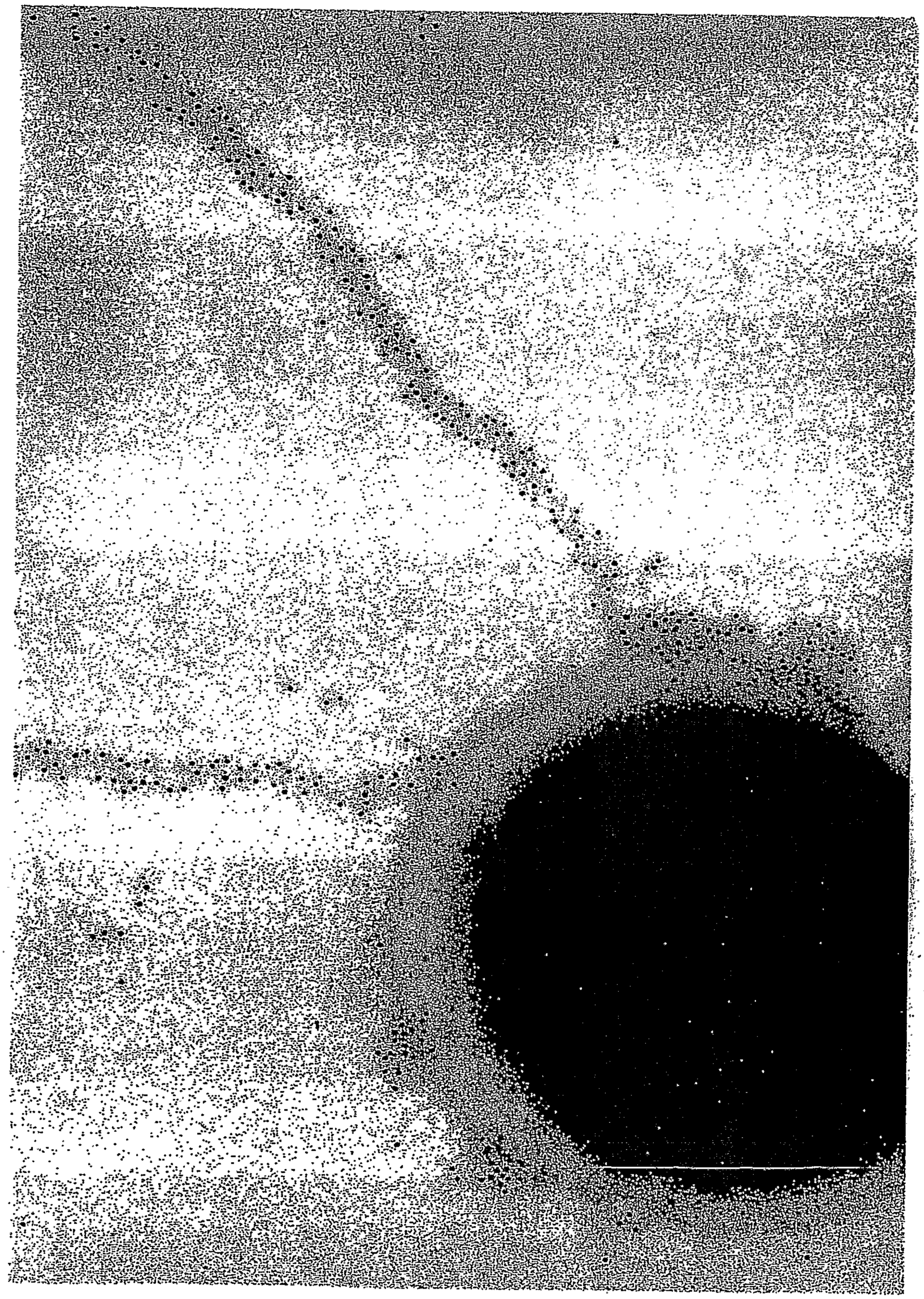


Figure 14

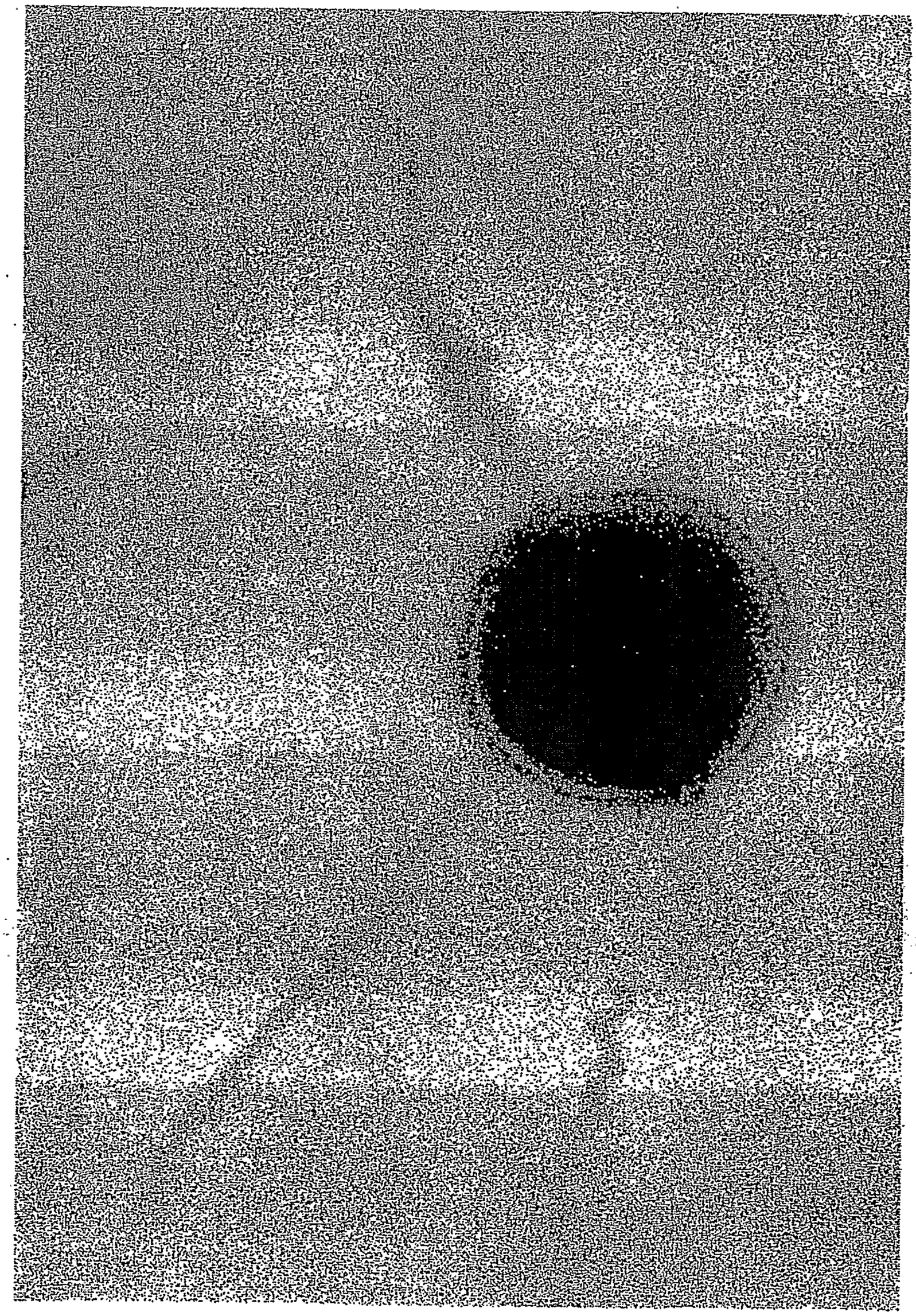


Figure 15



Figure 16

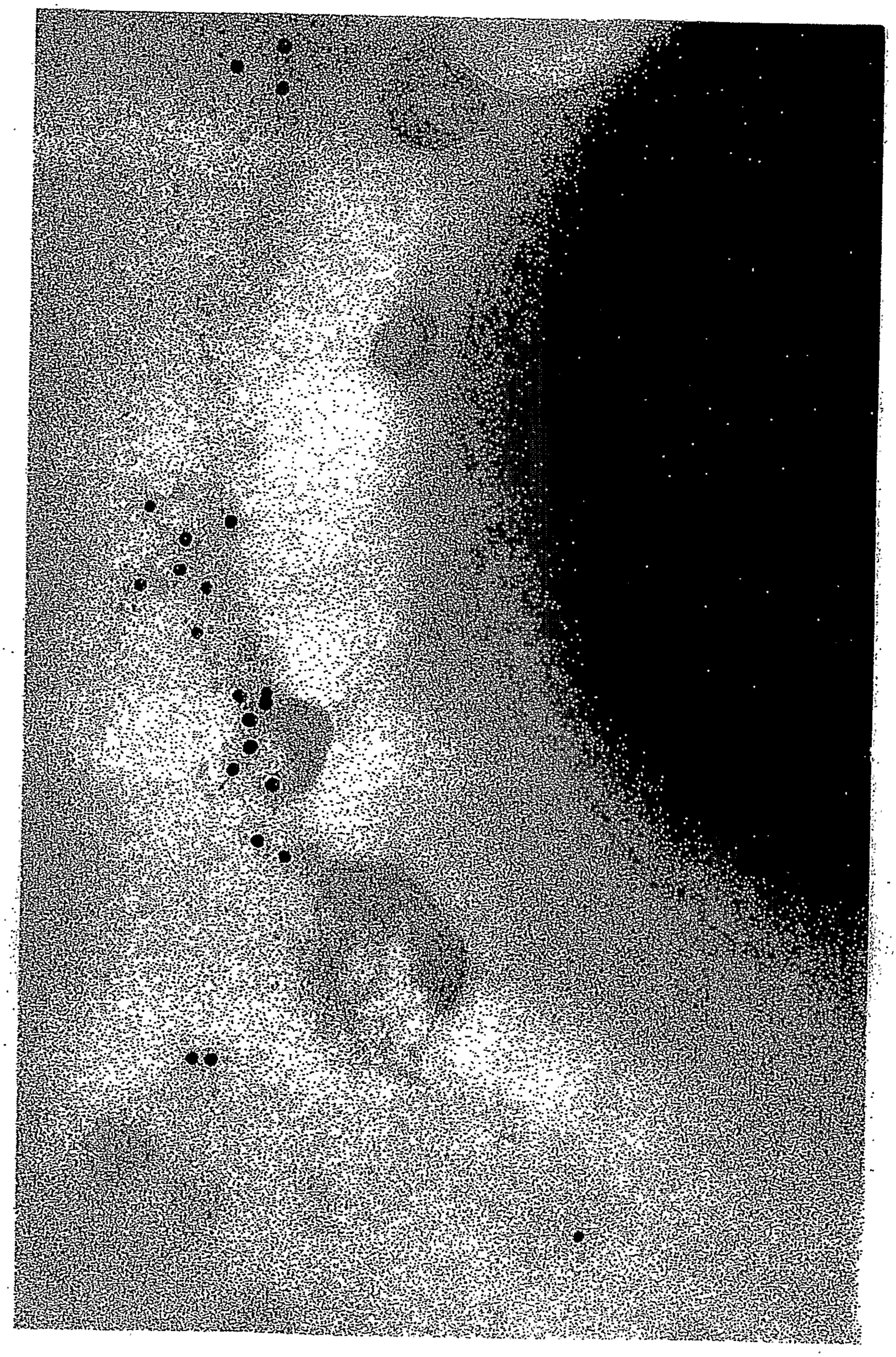


Figure 17

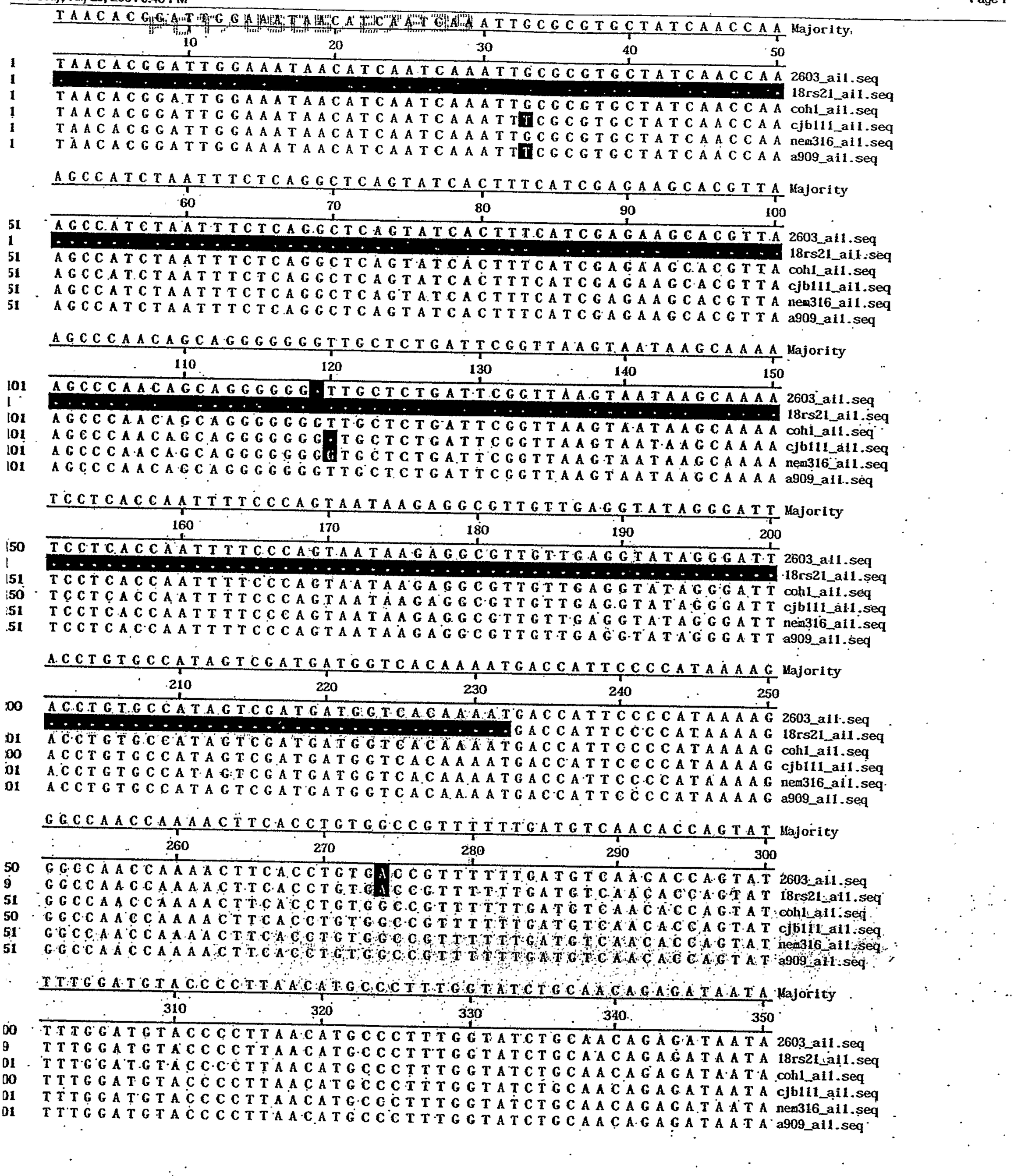

TAACACGGATTGGAAATAACATCAATCAAATTGCGCGTGCTATCAACCAA Majority
10 20 30 40 50
1 TAACACGGATTGGAAATAACATCAATCAAATTGCGCGTGCTATCAACCAA 2603_all.seq
1 18rs21_all.seq
1 TAACACGGATTGGAAATAACATCAATCAAATTGCGCGTGCTATCAACCAA coh1_all.seq
1 TAACACGGATTGGAAATAACATCAATCAAATTTCGCGTGCTATCAACCAA cjb111_all.seq
1 TAACACGGATTGGAAATAACATCAATCAAATTGCGCGTGCTATCAACCAA nem316_all.seq
1 TAACACGGATTGGAAATAACATCAATCAAATTTCGCGTGCTATCAACCAA a909_all.seq
AGCCATCTAATTTCTCAGGCTCAGTATCACTTTCATCGAGAAGCACGTTA Majority
60 70 80 90 100
51 AGCCATCTAATTTCTCAGGCTCAGTATCACTTTCATCGAGAAGCACGTTA 2603_all.seq
1 18rs21_all.seq
51 AGCCATCTAATTTCTCAGGCTCAGTATCACTTTCATCGAGAAGCACGTTA coh1_all.seq
51 AGCCATCTAATTTCTCAGGCTCAGTATCACTTTCATCGAGAAGCACGTTA cjb111_all.seq
51 AGCCATCTAATTTCTCAGGCTCAGTATCACTTTCATCGAGAAGCACGTTA nem316_all.seq
51 AGCCATCTAATTTCTCAGGCTCAGTATCACTTTCATCGAGAAGCACGTTA a909_all.seq
AGCCCAACAGCAGGGGGGGTTGCTCTGATTCGGTTAAGTAATAAGCAAAA Majority
110 120 130 140 150
101 AGCCCAACAGCAGGGGGGTTGCTCTGATTCGGTTAAGTAATAAGCAAAA 2603_all.seq
1 18rs21_all.seq
101 AGCCCAACAGCAGGGGGGGTTGCTCTGATTCGGTTAAGTAATAAGCAAAA coh1_all.seq
101 AGCCCAACAGCAGGGGGGG-TGCTCTGATTCGGTTAAGTAATAAGCAAAA cjb111_all.seq
101 AGCCCAACAGCAGGGGGGGGTGCTCTGATTCGGTTAAGTAATAAGCAAAA nem316_all.seq
101 AGCCCAACAGCAGGGGGGGGTTGCTCTGATTCGGTTAAGTAATAAGCAAAA a909_all.seq
TCCTCACCAATTTTCCCAGTAATAAGAGGCGTTGTTGAGGTATAGGGATT Majority
160 170 180 190 200
150 TCCTCACCAATTTTCCCAGTAATAAGAGGCGTTGTTGAGGTATAGGGATT 2603_all.seq
1 18rs21_all.seq
151 TCCTCACCAATTTTCCCAGTAATAAGAGGCGTTGTTGAGGTATAGGGATT coh1_all.seq
150 TCCTCACCAATTTTCCCAGTAATAAGAGGCGTTGTTGAGGTATAGGGATT cjb111_all.seq
151 TCCTCACCAATTTTCCCAGTAATAAGAGGCGTTGTTGAGGTATAGGGATT nem316_all.seq
151 TCCTCACCAATTTTCCCAGTAATAAGAGGCGTTGTTGAGGTATAGGGATT a909_all.seq
ACCTGTGCCATAGTCGATGATGGTCACAAAATGACCATTCCCCATAAAAG Majority
210 220 230 240 250
200 ACCTGTGCCATAGTCGATGATGGTCACAAAATGACCATTCCCCATAAAAG 2603_all.seq
GACCATTCCCCATAAAAG 18rs21_all.seq
201 ACCTGTGCCATAGTCGATGATGGTCACAAAATGACCATTCCCCATAAAAG coh1_all.seq
200 ACCTGTGCCATAGTCGATGATGGTCACAAAATGACCATTCCCCATAAAAG cjb111_all.seq
201 ACCTGTGCCATAGTCGATGATGGTCACAAAATGACCATTCCCCATAAAAG nem316_all.seq
201 ACCTGTGCCATAGTCGATGATGGTCACAAAATGACCATTCCCCATAAAAG a909_all.seq
GGCCAACCAAAACTTCACCTGTGGCCGTTTTTTGATGTCAACACCAGTAT Majority
260 270 280 290 300
250 GGCCAACCAAAACTTCACCTGTGACCGTTTTTTGATGTCAACACCAGTAT 2603_all.seq
9 GGCCAACCAAAACTTCACCTGTGACCGTTTTTTGATGTCAACACCAGTAT 18rs21_all.seq
51 GGCCAACCAAAACTTCACCTGTGGCCGTTTTTTGATGTCAACACCAGTAT coh1_all.seq
50 GGCCAACCAAAACTTCACCTGTGGCCGTTTTTTGATGTCAACACCAGTAT cjb111_all.seq
51 GGCCAACCAAAACTTCACCTGTGGCCGTTTTTTGATGTCAACACCAGTAT nem316_all.seq
51 GGCCAACCAAAACTTCACCTGTGGCCGTTTTTTGATGTCAACACCAGTAT a909_all.seq
TTTGGATGTACCCCTTAACATGCCCTTTGGTATCTGCAACAGAGATAATA Majority
310 320 330 340 350
300 TTTGGATGTACCCCTTAACATGCCCTTTGGTATCTGCAACAGAGATAATA 2603_all.seq
9 TTTGGATGTACCCCTTAACATGCCCTTTGGTATCTGCAACAGAGATAATA 18rs21_all.seq
01 TTTGGATGTACCCCTTAACATGCCCTTTGGTATCTGCAACAGAGATAATA coh1_all.seq
00 TTTGGATGTACCCCTTAACATGCCCTTTGGTATCTGCAACAGAGATAATA cjb111_all.seq
01 TTTGGATGTACCCCTTAACATGCCCTTTGGTATCTGCAACAGAGATAATA nem316_all.seq
01 TTTGGATGTACCCCTTAACATGCCCTTTGGTATCTGCAACAGAGATAATA a909_all.seq

Figure 18

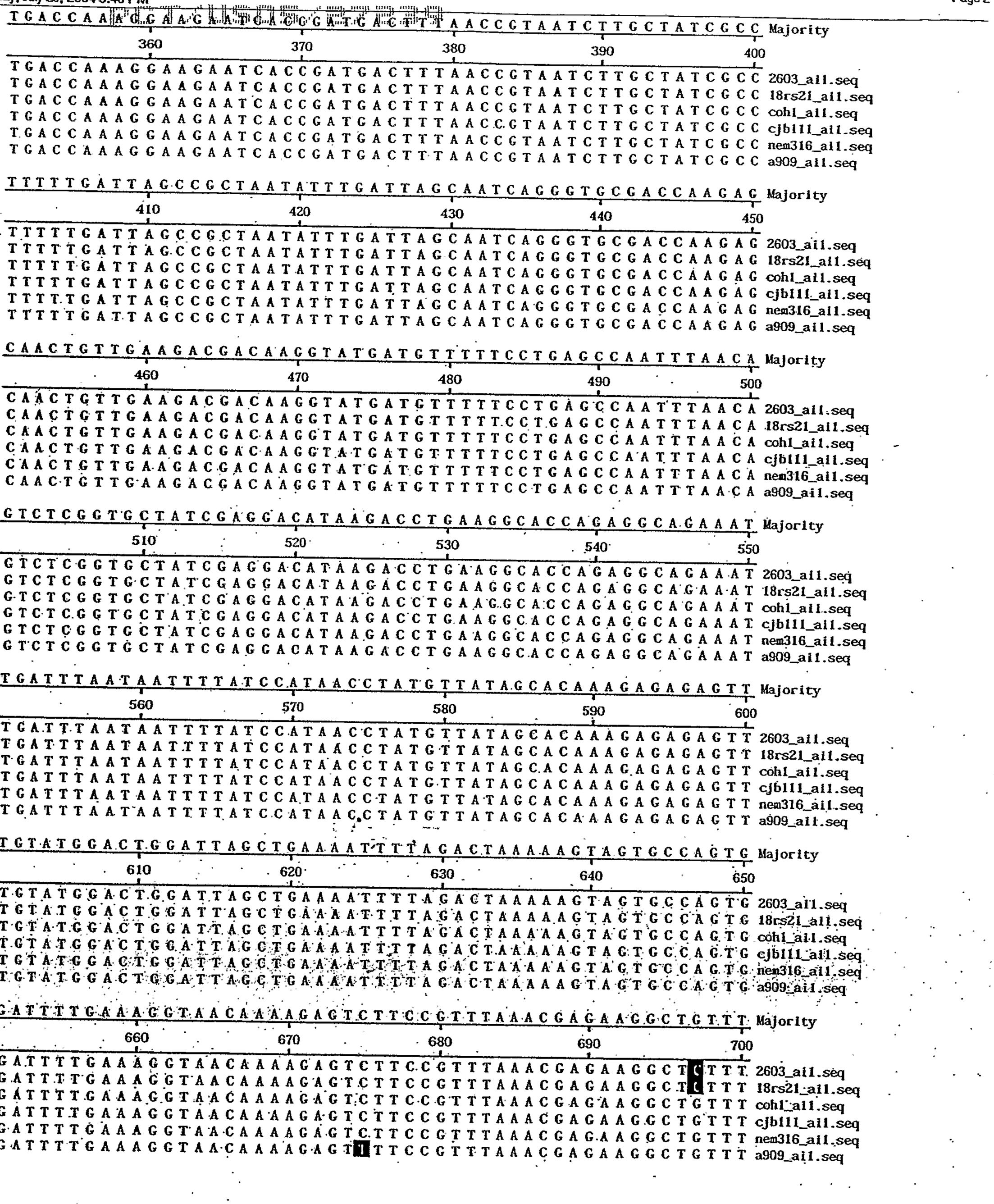

TGACCAAACGAACGAATCACCGATGACTTTAACCGTAATCTTGCTATCGCC Majority
360 370 380 390 400
350 TGACCAAAGGAAGAATCACCGATGACTTTAACCGTAATCTTGCTATCGCC 2603_all.seq
319 TGACCAAAGGAAGAATCACCGATGACTTTAACCGTAATCTTGCTATCGCC 18rs21_all.seq
351 TGACCAAAGGAAGAATCACCGATGACTTTAACCGTAATCTTGCTATCGCC coh1_all.seq
350 TGACCAAAGGAAGAATCACCGATGACTTTAACCGTAATCTTGCTATCGCC cjb111_all.seq
351 TGACCAAAGGAAGAATCACCGATGACTTTAACCGTAATCTTGCTATCGCC nem316_all.seq
351 TGACCAAAGGAAGAATCACCGATGACTTTAACCGTAATCTTGCTATCGCC a909_all.seq
TTTTTGATTAGCCGCTAATATTTGATTAGCAATCAGGGTGCCGACCAAGAG Majority
410 420 430 440 450
400 TTTTTGATTAGCCGCTAATATTTGATTAGCAATCAGGGTGCCGACCAAGAG 2603_all.seq
369 TTTTTGATTAGCCGCTAATATTTGATTAGCAATCAGGGTGCCGACCAAGAG 18rs21_all.seq
401 TTTTTGATTAGCCGCTAATATTTGATTAGCAATCAGGGTGCCGACCAAGAG coh1_all.seq
400 TTTTTGATTAGCCGCTAATATTTGATTAGCAATCAGGGTGCCGACCAAGAG cjb111_all.seq
401 TTTTTGATTAGCCGCTAATATTTGATTAGCAATCAGGGTGCCGACCAAGAG nem316_all.seq
401 TTTTTGATTAGCCGCTAATATTTGATTAGCAATCAGGGTGCCGACCAAGAG a909_all.seq
CAACTGTTGAAGACGACAAGGTATGATGTTTTTCCTGAGCCAATTTAACA Majority
460 470 480 490 500
450 CAACTGTTGAAGACGACAAGGTATGATGTTTTTCCTGAGCCAATTTAACA 2603_all.seq
419 CAACTGTTGAAGACGACAAGGTATGATGTTTTTCCTGAGCCAATTTAACA 18rs21_all.seq
451 CAACTGTTGAAGACGACAAGGTATGATGTTTTTCCTGAGCCAATTTAACA coh1_all.seq
450 CAACTGTTGAAGACGACAAGGTATGATGTTTTTCCTGAGCCAATTTAACA cjb111_all.seq
451 CAACTGTTGAAGACGACAAGGTATGATGTTTTTCCTGAGCCAATTTAACA nem316_all.seq
451 CAACTGTTGAAGACGACAAGGTATGATGTTTTTCCTGAGCCAATTTAACA a909_all.seq
GTCTCGGTGCTATCGAGGACATAAGACCTGAAGGCACCAGAGGCAGAAAT Majority
510 520 530 540 550
500 GTCTCGGTGCTATCGAGGACATAAGACCTGAAGGCACCAGAGGCAGAAAT 2603_all.seq
469 GTCTCGGTGCTATCGAGGACATAAGACCTGAAGGCACCAGAGGCAGAAAT 18rs21_all.seq
501 GTCTCGGTGCTATCGAGGACATAAGACCTGAAGGCACCAGAGGCAGAAAT coh1_all.seq
500 GTCTCGGTGCTATCGAGGACATAAGACCTGAAGGCACCAGAGGCAGAAAT cjb111_all.seq
501 GTCTCGGTGCTATCGAGGACATAAGACCTGAAGGCACCAGAGGCAGAAAT nem316_all.seq
501 GTCTCGGTGCTATCGAGGACATAAGACCTGAAGGCACCAGAGGCAGAAAT a909_all.seq
TGATTTAATAATTTTATCCATAACCTATGTTATAGCACAAAGAGAGAGTT Majority
560 570 580 590 600
550 TGATTTAATAATTTTATCCATAACCTATGTTATAGCACAAAGAGAGAGTT 2603_all.seq
519 TGATTTAATAATTTTATCCATAACCTATGTTATAGCACAAAGAGAGAGTT 18rs21_all.seq
551 TGATTTAATAATTTTATCCATAACCTATGTTATAGCACAAAGAGAGAGTT coh1_all.seq
550 TGATTTAATAATTTTATCCATAACCTATGTTATAGCACAAAGAGAGAGTT cjb111_all.seq
551 TGATTTAATAATTTTATCCATAACCTATGTTATAGCACAAAGAGAGAGTT nem316_all.seq
551 TGATTTAATAATTTTATCCATAACCTATGTTATAGCACAAAGAGAGAGTT a909_all.seq
TGTATGGACTGGATTAGCTGAAAATTTTAGACTAAAAAGTAGTGCCAGTG Majority
610 620 630 640 650
600 TGTATGGACTGGATTAGCTGAAAATTTTAGACTAAAAAGTAGTGCCAGTG 2603_all.seq
569 TGTATGGACTGGATTAGCTGAAAATTTTAGACTAAAAAGTAGTGCCAGTG 18rs21_all.seq
601 TGTATGGACTGGATTAGCTGAAAATTTTAGACTAAAAAGTAGTGCCAGTG coh1_all.seq
600 TGTATGGACTGGATTAGCTGAAAATTTTAGACTAAAAAGTAGTGCCAGTG cjb111_all.seq
601 TGTATGGACTGGATTAGCTGAAAATTTTAGACTAAAAAGTAGTGCCAGTG nem316_all.seq
601 TGTATGGACTGGATTAGCTGAAAATTTTAGACTAAAAAGTAGTGCCAGTG a909_all.seq
GATTTTGAAAGGTAACAAAAGAGTCTTCCGTTTAAACGAGAAGGCTGTTT Majority
660 670 680 690 700
650 GATTTTGAAAGGTAACAAAAGAGTCTTCCGTTTAAACGAGAAGGCTGTTT 2603_all.seq
619 GATTTTGAAAGGTAACAAAAGAGTCTTCCGTTTAAACGAGAAGGCTGTTT 18rs21_all.seq
651 GATTTTGAAAGGTAACAAAAGAGTCTTCCGTTTAAACGAGAAGGCTGTTT coh1_all.seq
650 GATTTTGAAAGGTAACAAAAGAGTCTTCCGTTTAAACGAGAAGGCTGTTT cjb111_all.seq
651 GATTTTGAAAGGTAACAAAAGAGTTTTCCGTTTAAACGAGAAGGCTGTTT nem316_all.seq
651 GATTTTGAAAGGTAACAAAAGAGTCTTCCGTTTAAACGAGAAGGCTGTTT a909_all.seq

FIGURE 18 A

```
   TTAGTCGTTTCAGATGAAGGCAAAAGGACGATGAAATAATTCCGTACCTT Majority
          710       720       730       740       750
00 TTAGTCGTTTCAGATGAAGGCAAAAGGACGATGAAATAATTCCGTACCTT 2603_all.seq
69 TTAGTCGTTTCAGATGAAGGCAAAAGGACGATGAAATAATTCCGTACCTT 18rs21_all.seq
01 TTAGTCGTTTCAGATGAAGGCAAAAGGACGATGAAATAATTCCGTACCTT cohl_all.seq
00 TTAGTCGTTTCAGATGAAGGCAAAAGGACGATGAAATAATTCCGTACCTT cjb111_all.seq
01 TTAGTCGTTTCAGATGAAGGCAAAAGGACGATGAAATAATTCCGTACCTT nem316_all.seq
01 TTAGTCGTTTCAGATGAAGGCAAAAGGACGATGAAATAATTCCGTACCTT a909_all.seq

   CATGGATTGCTATGTTACTGGCATGAGGTCTCACGATATTTAGTAAGATA Majority
          760       770       780       790       800
50 CATGGATTGCTATGTTACTGGCATGAGGTCTCACGATATTTAGTAAGATA 2603_all.seq
19 CATGGATTGCTATGTTACTGGCATGAGGTCTCACGATATTTAGTAAGATA 18rs21_all.seq
51 CATGGATTGCTATGTTACTGGCATGAGGTCTCACGATATTTAGTAAGATA cohl_all.seq
50 CATGGATTGCTATGTTACTGGCATGAGGTCTCACGATATTTAGTAAGATA cjb111_all.seq
51 CATGGATTGCTATGTTACTGGCATGAGGTCTCACGATATTTAGTAAGATA nem316_all.seq
51 CATGGATTGCTATGTTACTGGCATGAGGTCTCACGATATTTAGTAAGATA a909_all.seq

   TTCGTTTGAAGATATTCCCACGTATTTTTTAAAGGTTTTAAGAAAATGTG Majority
          810       820       830       840       850
00 TTCGTTTGAAGATATTCCCACGTATTTTTTAAAGGTTTTAAGAAAATGTG 2603_all.seq
69 TTCGTTTGAAGATATTCCCACGTATTTTTTAAAGGTTTTAAGAAAATGTG 18rs21_all.seq
01 TTCGTTTGAAGATATTCCCACGTATTTTTTAAAGGTTTTAAGAAAATGTG cohl_all.seq
00 TTCGTTTGAAGATATTCCCACGTATTTTTTAAAGGTTTTAAGAAAATGTG cjb111_all.seq
01 TTCGTTTGAAGATATTCCCACGTATTTTTTAAAGGTTTTAAGAAAATGTG nem316_all.seq
01 TTCGTTTGAAGATATTCCCACGTATTTTTTAAAGGTTTTAAGAAAATGTG a909_all.seq

   TGGTGTCGTAAAAATGTAATAATTTCGCTACTTCCCCAATCGGTACCCCT Majority
          860       870       880       890       900
50 TGGTGTCGTAAAAATGTAATAATTTCGCTACTTCCCCAATCGGTACCCCT 2603_all.seq
19 TGGTGTCGTAAAAATGTAATAATTTCGCTACTTCCCCAATCGGTACCCCT 18rs21_all.seq
51 TGGTGTCGTAAAAATGTAATAATTTCGCTACTTCCCCAATCGGTACCCCT cohl_all.seq
50 TGGTGTCGTAAAAATGTAATAATTTCGCTACTTCCCCAATCGGTACCCCT cjb111_all.seq
51 TGGTGTCGTAAAAATGTAAT/AATTTCGCTACTTCCCCAATCGGTACCCCT nem316_all.seq
51 TGGTGTCGTAAAAATGTAATAATTTCGCTACTTCCCCAATCGGTACCCCT a909_all.seq

   CTTTGCAGTAGTAATTGTCCCTCCTTAATTTTTGCCTTTAGAATATAACT Majority
          910       920       930       940       950
)0 CTTTGCAGTAGTAATTGTCCCTCCTTAATTTTTGCCTTTAGAATATAACT 2603_all.seq
59 CTTTGCAGTAGTAATTGTCCCTCCTTAATTTTTGCCTTTAGAATATAACT 18rs21_all.seq
)1 CTTTGCAGTAGTAATTGTCCCTCCTTAATTTTTGCCTTTAGAATATAACT cohl_all.seq
)0 CTTTGCAGTAGTAATTGTCCCTCCTTAATTTTTGCCTTTAGAATATAACT cjb111_all.seq
)1 CTTTGCAGTAGTAATTGTCCCTCCTTAATTTTTGCCTTTAGAATATAACT nem316_all.seq
)1 CTTTGCAGTAGTAATTGTCCCTCCTTAATTTTTGCCTTTAGAATATAACT a909_all.seq

   TTGCAAGGAAATGTCAGAGTATTTTTTAAAAACTGAGCGTAAGTGGGAAT Majority
          960       970       980       990      1000
30 TTGCAAGGAAATGTCAGAGTATTTTTTAAAAACTGAGCGTAAGTGGGAAT 2603_all.seq
.9 TTGCAAGGAAATGTCAGAGTATTTTTTAAAAACTGAGCGTAAGTGGGAAT 18rs21_all.seq
31 TTGCAAGGAAATGTCAGAGTATTTTTTAAAAACTGAGCGTAAGTGGGAAT cohl_all.seq
30 TTGCAAGGAAATGTCAGAGTATTTTTTAAAAACCGAGCGTAAGTGGGAAT cjb111_all.seq
31 TTGCAAGGAAATGTCAGAGTATTTTTTAAAAACCGAGCGTAAGTGGGAAT nem316_all.seq
31 TTGCAAGGAAATGTCAGAGTATTTTTTAAAAACCGAGCGTAAGTGGGAAT a909_all.seq

   CTGAGATATATAGGTAGTTGGCAATATCAGATACTTTGAGTTTGGAGTAG Majority
         1010      1020      1030      1040      1050
)00 CTGAGATATATAGGTAGTTGGCAATATCAGATACTTTGAGTTTGGAGTAG 2603_all.seq
39 CTGAGATATATAGGTAGTTGGCAATATCAGATACTTTGAGTTTGGAGTAG 18rs21_all.seq
)01 CTGAGATATATAGGTAGTTGGCAATATCAGATACTTTGAGTTTGGAGTAG cohl_all.seq
)00 TTGAGATATATAGGTAGTTGGCAATATCAGATACTTTGAGTTTGGAGTAG cjb111_all.seq
)01 CTGAGATATATAGGTAGTTGGCAATATCAGATACTTTGAGTTTGGAGTAG nem316_all.seq
)01 CTGAGATATATAGGTAGTTGGCAATATCAGATACTTTGAGTTTGGAGTAG a909_all.seq
```

FIGURE 18 B

```
          AGGTGGTTGTCCACATAATGGAGAATACTATTGTACATTTGCTGCTTGTC Majority
                1060      1070      1080      1090      1100
1050      AGGTGGTTGTCCACATAATGGAGAATACTATTGTACATTTGCTGCTTGTC 2603_all.seq
819       AGGTGGTTGTCCACATAATGGAGAATACTATTGTACATTTGCTGCTTGTC 18rs21_all.seq
1051      AGGTGGTTGTCCACATAATGGAGAATACTATTGTACATTTGCTGCTTGTC coh1_all.seq
1050      AGGTGGTTGTCCACATAATGGAGAATACTATTGTACATTTGCTGCTTGTC cjb111_all.seq
1051      AGGTGGTTGTCCACATAATGGAGAATACTATTGTACATTTGCTGCTTGTC nem316_all.seq
1051      AGGTGGTTGTCCACATAATGGAGAATACTATTGTACATTTGCTGCTTGTC a909_all.seq

          AGAGATGCTCTTATTGGTTAAGGATTCTGAAAAATCAATAAGAGCTGCAC Majority
                1110      1120      1130      1140      1150
1100      AGAGATGCTCTTATTGGTTAAGGATTCTGAAAAATCAATAAGAGCTGCAC 2603_all.seq
869       AGAGATGCTCTTATTGGTTAAGGATTCTGAAAAATCAATAAGAGCTGCAC 18rs21_all.seq
1101      AGAGATGCTCTTATTGGTTAAGGATTCTGAAAAATCAATAAGAGCTGCAC coh1_all.seq
1100      AGAGATGCTCTTATTGGTTAAGGATTCTGAAAAATCAATAAGAGCTGCAC cjb111_all.seq
1101      AGAGATGCTCTTATTGGTTAAGGATTCTGAAAAATCAATAAGAGCTGCAC nem316_all.seq
1101      AGAGATGCTCTTATTGGTTAAGGATTCTGAAAAATCAATAAGAGCTGCAC a909_all.seq

          AGCGAATTCTTGAAACATCAATAAGATCAGGAGCCTCTTCGTTTAAAGCC Majority
                1160      1170      1180      1190      1200
1150      AGCGAATTCTTGAAACATCAATAAGATCAGGAGCCTCTTCGTTTAAAGCC 2603_all.seq
919       AGCGAATTCTTGAAACATCAATAAGATCAGGAGCCTCTTCGTTTAAAGCC 18rs21_all.seq
1151      AGCGAATTCTTGAAACATCAATAAGATCAGGAGCCTCTTCGTTTAAAGCC coh1_all.seq
1150      AGCGAATTCTTGAAACATCAATAAGATCAGGAGCCTCTTCGTTTAAAGCC cjb111_all.seq
1151      AGCGAATTCTTGAAACATCAATAAGATCAGGAGCCTATTCGTTTAAAGCC nem316_all.seq
1151      AGCGAATTCTTGAAACATCAATAAGATCAGGAGCCTCTTCGTTTAAAGCC a909_all.seq

          ATATAGTGCTTTACCAGCGCATAACTTTTAGCCACATCAGTATTTTCCTC Majority
                1210      1220      1230      1240      1250
1200      ATATAGTGCTTTACCAGCGCATAACTTTTAGCCACATCAGTATTTTCCTC 2603_all.seq
969       ATATAGTGCTTTACCAGCGCATAACTTTTAGCCACATCAGTATTTTCCTC 18rs21_all.seq
1201      ATATAGTGCTTTACCAGCGCATAACTTTTAGCCACATCAGTATTTTCCTC coh1_all.seq
1200      ATATAGTGCTTTACCAGCGCATAACTTTTAGCCACATCAGTATTTTTCTC cjb111_all.seq
1201      ATATAGTGCTTTACCAGCGCATAACTTTTAGCCACATCAGTATTTTCCTC nem316_all.seq
1201      ATATAGTGCTTTACCAGCGCATAACTTTTAGCCACATCAGTATTTTCCTC a909_all.seq

          GAAACTTAATTCTAGTAATTTTGTTAAGTAAACAACAGTTAAGTTCTTTT Majority
                1260      1270      1280      1290      1300
1250      GAAACTTAATTCTAGTAATTTTGTTAAGTAAACAACAGTTAAGTTCTTTT 2603_all.seq
1019      GAAACTTAATTCTAGTAATTTTGTTAAGTAAACAACAGTTAAGTTCTTTT 18rs21_all.seq
1251      GAAACTTAATTCTAGTAATTTTGTTAAGTAAACAACAGTTAAGTTCTTTT coh1_all.seq
1250      GAAACTTAATTCTAGTAATTTTGTTAAGTAAACAACAGTTAAGTTCTTTT cjb111_all.seq
1251      GAAACTTAATTCTAGTAATTTTGTTAAGTAAACAACAGTTAAGTTCTTTT nem316_all.seq
1251      GAAACTTAATTCTAGTAATTTTGTTAAGTAAACAACAGTTAAGTTCTTTT a909_all.seq

          CAGCTCTTAGGGCAGGGATTGAAGATGAGGTAACACTGGATGATGGGAGG Majority
                1310      1320      1330      1340      1350
1300      CAGCTCTTAGGGCAGGGATTGAAGATGAGGTAACACTGGATGATGGGAGG 2603_all.seq
1069      CAGCTCTTAGGGCAGGGATTGAAGATGAGGTAACACTGGATGATGGGAGG 18rs21_all.seq
1301      CAGCTCTTAGGGCAGGGATTGAAGATGAGGTAACACTGGATGATGGGAGG coh1_all.seq
1300      CAGCTCTTAGGGCAGGGATTGAAGATGAGGTAACACTGGATGATGGGAGG cjb111_all.seq
1301      CAGCTCTTAGGGCAGGGATTGAAGATGAGGTAACACTGGATGATGGGAGG nem316_all.seq
1301      CAGCTCTTAGGGCAGGGATTGAAGATGAGGTAACACTGGATGATGGGAGG a909_all.seq

          CGATTAATTTCTTGCTTTAACAGTTGAGTGTTACCCAGCTTAACGAGATC Majority
                1360      1370      1380      1390      1400
1350      CGATTAATTTCTTGCTTTAACAGTTGAGTGTTACCCAGCTTAACGAGATC 2603_all.seq
1119      CGATTAATTTCTTGCTTTAACAGTTGAGTGTTACCCAGCTTAACGAGATC 18rs21_all.seq
1351      CGATTAATTTCTTGCTTTAACAGTTGAGTGTTACCCAGCTTAACGAGATC coh1_all.seq
1350      CGATTAATTTCTTGCTTTAACAGTTGAGTGTTACCCAGCTTAACGAGATC cjb111_all.seq
1351      CGATTAATTTCTTGCTTTAACAGTTGAGTGTTACCCAGCTTAACGAGATC nem316_all.seq
1351      CGATTAATTTCTTGCTTTAACAGTTGAGTGTTACCCAGCTTAACGAGATC a909_all.seq
```

FIGURE 18 C

```
      AATAATGTGATTGAGATGGTTTAAAACAGTGGGTAACTGAAAAGAGTTTT Majority
                1410      1420      1430      1440      1450
400   AATAATGTGATTGAGATGGTTTAAAACAGTGGGTAACTGAAAAGAGTTTT 2603_all.seq
169   AATAATGTCATTGAGATGGTTTAAAACAGTGGGTAACTGAAAAGAGTTTT 18rs21_all.seq
401   AATAATGTGATTGAGATGGTTTAAAACAGTGGGTAACTGAAAAGAGTTTT coh1_all.seq
400   AATAATGTGATTGAGATGGTTTAAAACAGTGGGTAACTGAAAAGAGTTTT cjb111_all.seq
401   AATAATGTCATTGAGATGGTTTAAAACAGTGGGTAACTGAAAAGAGTTTT nem316_all.seq
401   AATAATGTGATTGAGATGGTTTAAAACAGTCGGTAACTGAAAAGAGTTTT a909_all.seq

      TCTTAGTATGTTTTAGGTGAAGAACAATATCAGGATCCGCAACAATCTGT Majority
                1460      1470      1480      1490      1500
450   TCTTAGTATGTTTTAGGTGAAGAACAATATCAGGATCCGCAACAATCTGT 2603_all.seq
219   TCTTAGTATGTTTTAGGTGAAGAACAATATCAGGATCCGCAACAATCTGT 18rs21_all.seq
451   TCTTAGTATGTTTTAGGTGAAGAACAATATCAGGATCCGCAACAATCTGT coh1_all.seq
450   TCTTAGTATGTTTTAGGTGAAGAACAATATCAGGATCCGCAACAATCTGT cjb111_all.seq
451   TCTTAGTATGTTTTAGGTGAAGAACAATATCAGGATCCGCAACAATCTGT nem316_all.seq
451   TCTTAGTATGTTTTAGGTGAAGAACAATATCAGGATCCGCAACAATCTGT a909_all.seq

      TCTGACTCTTCTAATAAATGATTGATGGCTTGTTGGCAACTAGCCTCAAA Majority
                1510      1520      1530      1540      1550
500   TCTGACTCTTCTAATAAATGATTGATGACTTGTTGGCAACTAGCCTCAAA 2603_all.seq
269   TCTGACTCTTCTAATAAATGATTGATGACTTGTTGGCAACTAGCCTCAAA 18rs21_all.seq
501   TCTGACTCTTCTAATAAATGATTGATGACTTGTTGGCAACTAGCCTCAAA coh1_all.seq
500   TCTGACTCTTCTAATAAATGATTGATGGCTTGTTGGCAACTAGCCTCAAA cjb111_all.seq
501   TCTGACTCTTCTAATAAATGATTGATGGCTTGTTGGCAACTAGCCTCAAA nem316_all.seq
501   TCTGACTCTTCTAATAAATGATTGATGGCTTGTTGGCAACTAGCCTCAAA a909_all.seq

      CTGTGTTTGGAAAAAGGCATCGATAGACACAAGAAGACTACGTATACTGG Majority
                1560      1570      1580      1590      1600
550   CTGTGTTTGGAAAAAGGCATCGATAGACACAAGAAGACTACGTATACTGG 2603_all.seq
319   CTGTGTTTGGAAAAAGGCATCGATAGACACAAGAAGACTACGTATACTGG 18rs21_all.seq
551   CTGTGTTTGGAAAAAGGCATCGATAGACACAAGAAGACTACGTATACTGG coh1_all.seq
550   CTGTGTTTGGAAAAAGGCATCGATAGACACAAGAAGACTACGTATACTGG cjb111_all.seq
551   CTGTGTTTGGAAAAAGGCATCGATAGACACAAGAAGACTACGTATACTGG nem316_all.seq
551   CTGTGTTTGGAAAAAGGCATCGATAGACACAAGAAGACTACGTATACTGG a909_all.seq

      TAGTAGGAAAACAAGGGACAAGCTTTATATAGGATAAGATTTCTTTTTTA Majority
                1610      1620      1630      1640      1650
600   TAGTAGGAAAACAAGGGACAAGCTTTATATAGGATAAGATTTCTTTTTTA 2603_all.seq
569   TAGTAGGAAAACAAGGGACAAGCTTTATATAGGATAAGATTTCTTTTTTA 18rs21_all.seq
601   TAGTAGGAAAACAAGGGACAAGCTTTATATAGGATAAGATTTCTTTTTTA coh1_all.seq
600   TAGTAGGAAAACAAGGGACAAGCTTTATATAGGATAAGATTTCTTTTTTA cjb111_all.seq
601   TAGTAGGAAAACAAGGGACAAGCTTTATATAGGATAAGATTTCTTTTTTA nem316_all.seq
601   TAGTAGGAAAACAAGGGACAAGCTTTATATAGGATAAGATTTCTTTTTTA a909_all.seq

      CTACGATGAGAAAATTGTTCTAGAAAGCGACTGGATAACTGTTCTTGCCT Majority
                1660      1670      1680      1690      1700
650   TTACGATGAGAAAATTGTTCTAGAAAGCGACTGGATAACTGTTCTTGCCT 2603_all.seq
619   CTACGATGAGAAAATTGTTCTAGAAAGCGACTGGATAACTGTTCTTGCCT 18rs21_all.seq
651   CTACGATGAGAAAATTGTTCTAGAAAGCGACTGGATAACTGTTCTTGCCT coh1_all.seq
650   CTACGATGAGAAAATTGTTCTAGAAAGCGACTGGATAACTGTTCTTGCCT cjb111_all.seq
651   CTACGATGAGAAAATTGTTCTAGAAAGCGACTGGATAACTGTTCTTGCCT nem316_all.seq
651   CTACGATGAGAAAATTGTTCTAGAAAGCGACTGGATAACTGTTCTTGCCT a909_all.seq

      ATTGATATCAGGGCTATAGGGATAAAATGGTCCAATAGCAATAAGATATT Majority
                1710      1720      1730      1740      1750
700   ATTGATATCAGGGCTATAGGGATAAAATGGTCCAATAGCAATAAGATATT 2603_all.seq
669   ATTGATATCAGGGCTATAGGGATAAAATGGTCCAATAGCAATAAGATATT 18rs21_all.seq
701   ATTGATATCAGGGCTATAGGGATAAAATGGTCCAATAGCAATAAGATATT coh1_all.seq
700   ATTGATATCAGGGCTATAGGGATAAAATGGTCCAATAGCAATAAGATATT cjb111_all.seq
701   ATTGATATCAGGGCTATAGGGATAAAATGGTCCAATAGCAATAAGATATT nem316_all.seq
701   ATTGATATCAGGGCTATAGGGATAAAATGGTCCAATAGCAATAAGATATT a909_all.seq
```

FIGURE 18 D

```
     GACAGACAGGAAAAATTAAGAATGATTCTTCAAAAAGACCCTCATAAACA Majority
          1760      1770      1780      1790      1800
750  GACAGACAGGAAAAATTAAGAATGATTCTTCAAAAAGACCCTCATAAACA
519  GACAGACAGGAAAAATTAAGAATGATTCTTCAAAAAGACCCTCATAAACA 2603_ail.seq
751  GACAGACAGGAAAAATTAAGAATGATTCTTCAAAAAGACCCTCATAAACA 18rs21_ail.seq
750  GACAGACAGGAAAAATTAAGAATGATTCTTCAAAAAGACCCTCATAAACA coh1_ail.seq
751  GACAGACAGGAAAAATTAAGAATGATTCTTCAAAAAGACCCTCATAAACA cjb111_ail.seq
751  GACAGACAGGAAAAATTAAGAATGATTCTTCAAAAAGACCCTCATAAACA nem316_ail.seq
751  GACAGACAGGAAAAATTAAGAATGATTCTTCAAAAAGATCCTCATAAACA a909_ail.seq

     GTGATATCTTGGTTATAAGGGATAGCTAAATGTTTTAAAAACTGATAGTA Majority
          1810      1820      1830      1840      1850
800  GTGATATCTTGGTTATAAGGGATAGCTAAATGTTTTAAAAACTGATAGTA
569  GTGATATCTTGGTTATAAGGGATAGCTAAATGTTTTAAAAACTGATAGTA 2603_ail.seq
801  GTGATATCTTGGTTATAAGGGATAGCTAAATGTTTTAAAAACTGATAGTA 18rs21_ail.seq
800  GTGATATCTTGGTTATAAGGGATAGCTAAATGTTTTAAAAACTGATAGTA coh1_ail.seq
801  GTGATATCTTGGTTATAAGGGATAGCTAAATGTTTTAAAAACTGATAGTA cjb111_ail.seq
801  GTGATATCTTGGTTATAAGGGATAGCTAAATGTTTTAAAAACTGATAGTA nem316_ail.seq
801  GTGATATCTTGGTTATAAGGGATAGCTAAATGTTTTAAAAACTGATAGTA a909_ail.seq

     AGGCAACAGATAGTCTTCGTTACCATATAACTGAACGAGTTCCTTGTCTC Majority
          1860      1870      1880      1890      1900
850  AGGCAACAGATAGTCTTCGTTACCATATAACTGAACGAGTTCCTTGTCTC
619  AGGCAACAGATAGTCTTCGTTACCATATAACTGAACGAGTTCCTTGTCTC 2603_ail.seq
851  AGGCAACAGATAGTCTTCGTTACTATATAACTGAACGAGTTCCTTGTCTC 18rs21_ail.seq
850  AGGCAACAGATAGTCTTCGTTACCATATAACTGAACGAGTTCCTTGTCTC coh1_ail.seq
851  AGGCAACAGATAGTCTTCGTTACCATATAACTGAACGAGTTCCTTGTCTC cjb111_ail.seq
851  AGGCAACAGATAGTCTTCGTTACCATATAACTGAACGAGTTCCTTGTCTC nem316_ail.seq
851  AGGCAACAGATAGTCTTCGTTACCATATAACTGAACGAGTTCCTTGTCTC a909_ail.seq

     GTGACATGACTGAAATAGGTAGTTGAGATATGGTATGCAATGTTTGAACA Majority
          1910      1920      1930      1940      1950
900  GTGACATGACTGAAATAGGTAGTTGAGATATGGTATGCAATGTTTGAACA
669  GTGACATGACTGAAATAGGTAGTTGAGATATGGTATGCAATGTTTGAACA 2603_ail.seq
901  GTGACATGACTGAAATAGGTAGTTGAGATATGGTATGCAATGTTTGAACA 18rs21_ail.seq
900  GTGACATGACTGAAATAGGTAGTTGAGATATGGTATGCAATGTTTGAACA coh1_ail.seq
901  GTGACATGACTGAAATAGGTAGTTGAGATATGGTATGCAATGTTTGAACA cjb111_ail.seq
901  GTGACATGACTGAAATAGGTAGTTGAGATATGGTATGCAATGTTTGAACA nem316_ail.seq
901  GTGACATGACTGAAATAGGTAGTTGAGATATGGTATGCAATGTTTGAACA a909_ail.seq

     TGTTTAAAATCGAATGTAACCATTTGATAGACCGCCTTCATTATCATTTC Majority
          1960      1970      1980      1990      2000
950  TGTTTAAAATCGAATGTAACCATTTGATAGACCGCCTTCATTATCATTTC
719  TGTTTAAAATCGAATGTAACCATTTGATAGACCGCCTTCATTATCATTTC 2603_ail.seq
951  TGTTTAAAATCGAATGTAACCATTTGATAGACCGCCTTCATTATCATTTC 18rs21_ail.seq
950  TGTTTAAAATCGAATGTAACCATTTGATAGACCGCCTTCATTATCATTTC coh1_ail.seq
951  TGTTTAAAATCGAATGTAACCATTTGATAGACCGCCTTCATTATCATTTC cjb111_ail.seq
951  TGTTTAAAATCGAATGTAACCATTTGATAGACCGCCTTCATTATCATTTC nem316_ail.seq
951  TGTTTAAAATCGAATGTAACCATTTGATAGACCGCCTTCATTATCATTTC a909_ail.seq

     TAGAATTTTTCTTTAGGTTTGTAAAGACTACAAAATAAAATGATGAAAAC Majority
          2010      2020      2030      2040      2050
000  TAGAATTTTTCTTTAGGTTTGTAAAGACTACAAAATAAAATGATGAAAAC
769  TAGAATTTTTCTTTAGGTTTGTAAAGACTACAAAATAAAATGATGAAAAC 2603_ail.seq
001  TAGAATTTTTCTTTAGGTTTGTAAAGACTACAAAATAAAATGATGAAAAC 18rs21_ail.seq
000  TAGAATTTTTCTTTAGGTTTGTAAAGACTACAAAATAAAATGATGAAAAC coh1_ail.seq
001  TAGAATTTTTCTTTAGGTTTGTAAAGACTACAAAATAAAATGATGAAAAC cjb111_ail.seq
001  TAGAATTTTTCTTTAGGTTTGTAAAGACTACAAAATAAAATGATGAAAAC nem316_ail.seq
001  TAGAATTTTTCTTTAGGTTTGTAAAGACTACAAAATAAAATGATGAAAAC a909_ail.seq

     AACTATCTTGTGGATACACTAAAAAGACACGCTAATTAGCAAACTCTCTC Majority
          2060      2070      2080      2090      2100
050  AACTATCTTGTGGATACACTAAAAAGACACGCTAATTAGCAAACTCTCTC
819  AACTATCTTGTGGATACACTAAAAAGACACGCTAATTAGCAAACTCTCTC 2603_ail.seq
051  AACTATCTTGTGGATACACTAAAAAGACACGCTAATTAGCAAACTCTCTC 18rs21_ail.seq
050  AACTATCTTGTGGATACACTAAAAAGACACGCTAATTAGCAAACTCTCTC coh1_ail.seq
051  AACTATCTTGTGGATACACTAAAAAGACACGCTAATTAGCAAACTCTCTC cjb111_ail.seq
051  AACTATCTTGTGGATACACTAAAAAGACACGCTAATTAGCAAACTCTCTC nem316_ail.seq
051  AACTATCTTGTGGATACACTAAAAAGACACGCTAATTAGCAAACTCTCTC a909_ail.seq
```

FIGURE 18 E

```
                TTCATCATCTCTCACCATTATTATACTACTATTTATATGACAAATAAAGG Majority
                        2110        2120        2130        2140        2150
2100 TTCATCATCTCTCACCATTATTATACTACTATTTATATGACAAATAAAGG 2603_all.seq
1869 TTCATCATCTCTCACCATTATTATACTACTATTTATATGACAAATAAAGG 18rs21_all.seq
2101 TTCATCATCTCTCACCATTATTATACTACTATTTATATGACAAATAAAGG coh1_all.seq
2100 TTCATCATCTCTCACCATTATTATACTACTATTTATATGACAAATAAAGG cjb111_all.seq
2101 TTCATCATCTCTCACCATTATTATACTACTATTTATATGACAAATAAAGG nem316_all.seq
2101 TTCATCATCTCTCACCATTATTATACTACTATTTATATGACAAATAAAGG a909_all.seq

                TGATTTTGTTAAAAATATAACTTTGAAAATCCACATATATTTTTAATCTT Majority
                        2160        2170        2180        2190        2200
2150 TGATTTTGTTAAAA TATAACTTTGAAAATCCACATATATTTTTAATCTT 2603_all.seq
1919 TGATTTTGTTAAAA TATAACTTTGAAAATCCACATATATTTTTAATCTT 18rs21_all.seq
2151 TTATTTTGTTAAAA TATAACTTTGAAAATCCACATATATTTTTAATCTT coh1_all.seq
2150 TGATTTTGTTAAAAATATAACTTTGAAAATCCACATATATTTTTAATCTT cjb111_all.seq
2151 TGATTTTGTTAAAAATATAACTTTGAAAATCCACATATATTTTTAATCTT nem316_all.seq
2151 TGATTTTGTTAAAAATATAACTTTGAAAATCCACATATATTTTTAATCTT a909_all.seq

                CCGTCTGAAAAAA-TAAATAAAAATAGTAAAAATAAACACGAATTTAAAA Majority
                        2210        2220        2230        2240        2250
2199 CCGTCTGAAAAAA-TAAATAAAAATAGTAAAAATAAACACGAATTTAAAA 2603_all.seq
1968 CCGTCTGAAAAAA-TAAATAAAAATAGTAAAAATAAACACGAATTTAAAA 18rs21_all.seq
2200 CCGTCTGAAAAAA-TAAATAAAAATAGTAAAAATAAACACGAATTTAAAA coh1_all.seq
2200 CCGTCTGAAAAAA-TAAATAAAAATAGTAAAAATAAACACGAATTTAAAA cjb111_all.seq
2201 CCGTCTGAAAAAAATAAATAAAAATAGTAAAAATAAACACGAATTTAAAA nem316_all.seq
2201 CCGTCTGAAAAAA-TAAATAAAAATAGTAAAAATAAACACGAATTTAAAA a909_all.seq

                TAAGCAAATTTTTTAAGAAAATCTGTGCTAAACTTTAATAGTTTTGTGCT Majority
                        2260        2270        2280        2290        2300
2248 TAAGCAAATTTTTTAAGAAAATCTGTGCTAAACTTTAATAGTTTTGTGCT 2603_all.seq
2017 TAAGCAAATTTTTTAAGAAAATCTGTGCTAAACTTTAATAGTTTTGTGCT 18rs21_all.seq
2249 TAAGCAAATTTTTTAAGAAAATCTGTGCTAAACTTTAATAGTTTTGTGCT coh1_all.seq
2249 TAAGCAAATTTTTTAAGAAAATCTGTGCTAAACTTTAATAGTTTTGTGCT cjb111_all.seq
2251 TAAGCAAATTTTTTAAGAAAATCTGTGCTAAACTTTAATAGTTTTGTGCT nem316_all.seq
2250 TAAGCAAATTTTTTAAGAAAATCTGTGCTAAACTTTAATAGTTTTGTGCT a909_all.seq

                TAATAATAATCAGCACTTACAAAGAACAAAGGGAAAAGCGAGGAGAGAAC Majority
                        2310        2320        2330        2340        2350
298 TAATAATAATCAGCACTTACAAAGAACAAAGGGAAAAGCGAGGAGAGAAC 2603_all.seq
067 TAATAATAATCAGCACTTACAAAGAACAAAGGGAAAAGCGAGGAGAGAAC 18rs21_all.seq
299 TAATAATAATCAGCACTTACAAAGAACAAAGGGAAAAGCGAGGAGAGAAC coh1_all.seq
299 TAATAATAATCAGCACTTACAAAGAACAAAGGGAAAAGCGAGGAGAGAAC cjb111_all.seq
301 TAATAATAATCAGCACTTACAAAGAACAAAGGGAAAAGCGAGGAGAGAAC nem316_all.seq
300 TAATAATAATCAGCACTTACAAAGAACAAAGGGAAAAGCGAGGAGAGAAC a909_all.seq

                TTTTAATGAAATTATCGAAGAAGTTATTGTTTTCGGCTGCTGTTTTAACA Majority
                        2360        2370        2380        2390        2400
348 TTTTAATGAAATTATCGAAGAAGTTATTGTTTTCGGCTGCTGTTTTAACA 2603_all.seq
117 TTTTAATGAAATTATCGAAGAAGTTATTGTTTTCGGCTGCTGTTTTAACA 18rs21_all.seq
349 TTTTAATGAAATTATCGAAGAAGTTATTGTTTTCGGCTGCTGTTTTAACA coh1_all.seq
349 TTTTAATGAAATTATCGAAGAAGTTATTGTTTTCGGCTGCTGTTTTAACA cjb111_all.seq
351 TTTTAATGAAATTATCGAAGAAGTTATTGTTTTCGGCTGCTGTTTTAACA nem316_all.seq
350 TTTTAATGAAATTATCGAAGAAGTTATTGTTTTCGGCTGCTGTTTTAACA a909_all.seq

                ATGGTGGCGGGGTCAACTGTTGAACCAGTAGCTCAGTTTGCCGACTGGAAT Majority
                        2410        2420        2430        2440        2450
398 ATGGTGGCGGGGTCAACTGTTGAACCAGTAGCTCAGTTTGCCGACTGGAAT 2603_all.seq
167 ATGGTGGCGGGGTCAACTGTTGAACCAGTAGCTCAGTTTGCCGACTGGAAT 18rs21_all.seq
399 ATGGTGGCGGGGTCAACTGTTGAACCAGTAGCTCAGTTTGCCGACTGGAAT coh1_all.seq
399 ATAGTGGCGGGGTCAACTGTTGAACCAGTAGCTCAGTTTGCCGACTGGAAT cjb111_all.seq
401 ATGGTGGCGGGGTCAACTGTTGAACCAGTAGCTCAGTTTGCCGACTGGAAT nem316_all.seq
400 ATAGTGGCGGGGTCAACTGTTGAACCAGTAGCTCAGTTTGCCGACTGGAAT a909_all.seq
```

FIGURE 18 F

```
                                                
```

```
CAAAAAGTAATGTGAGATACTTGTATGTAGAAGATTTAAAGAATTCACCT Majority
        2810      2820      2830      2840      2850
2798 CAAAAAGTAATGTGAGATACTTGTATGTAGAAGATTTAAAGAATTCACCT 2603_all.seq
2567 CAAAAAGTAATGTGAGATACTTGTATGTAGAAGATTTAAAGAATTCACCT 18rs21_all.seq
2799 CAAAAAGTAATGTGAGATACTTGTATGTAGAAGATTTAAAGAATTCACCT coh1_all.seq
2799 CAAAAAGTAATGTGAGATACTTGTATGTAGAAGATTTAAAGAATTCACCT cjb111_all.seq
2801 CAAAAAGTAATGTGAGATACTTGTATGTAGAAGATTTAAAGAATTCACCT nem316_all.seq
2800 CAAAAAGTAATGTGAGATACTTGTATGTAGAAGATTTAAAGAATTCACCT a909_all.seq

TCAAACATTACCAAAGCTTATGCTGTACCGTTTGTGTTGGAATTACCAGT Majority
        2860      2870      2880      2890      2900
2848 TCAAACATTACCAAAGCTTATGCTGTACCGTTTGTGTTGGAATTACCAGT 2603_all.seq
2617 TCAAACATTACCAAAGCTTATGCTGTACCGTTTGTGTTGGAATTACCAGT 18rs21_all.seq
2849 TCAAACATTACCAAAGCTTATGCTGTACCGTTTGTGTTGGAATTACCAGT coh1_all.seq
2849 TCAAACATTACCAAAGCTTATGCTGTACCGTTTGTGTTGGAATTACCAGT cjb111_all.seq
2851 TCAAACATTACCAAAGCTTATGCTGTACCGTTTGTGTTGGAATTACCAGT nem316_all.seq
2850 TCAAACATTACCAAAGCTTATGCTGTACCGTTTGTGTTGGAATTACCAGT a909_all.seq

TGCTAACTCTACAGGTACAGGTTTCCTTTCTGAAATTAATATTTACCCTA Majority
        2910      2920      2930      2940      2950
2898 TGCTAACTCTACAGGTACAGGTTTCCTTTCTGAAATTAATATTTACCCTA 2603_all.seq
2667 TGCTAACTCTACAGGTACAGGTTTCCTTTCTGAAATTAATATTTACCCTA 18rs21_all.seq
2899 TGCTAACTCTACAGGTACAGGTTTCCTTTCTGAAATTAATATTTACCCTA coh1_all.seq
2899 TGCTAACTCTACAGGTACAGGTTTCCTTTCTGAAATTAATATTTACCCTA cjb111_all.seq
2901 TGCTAACTCTACAGGTACAGGTTTCCTTTCTGAAATTAATATTTACCCTA nem316_all.seq
2900 TGCTAACTCTACAGGTACAGGTTTCCTTTCTGAAATTAATATTTACCCTA a909_all.seq

AAAACGTTGTAACTGATGAACCAAAAAACAGATAAAGATGTTAAAAAATTA Majority
        2960      2970      2980      2990      3000
2948 AAAACGTTGTAACTGATGAACCAAAAAACAGATAAAGATGTTAAAAAATTA 2603_all.seq
2717 AAAACGTTGTAACTGATGAACCAAAAAACAGATAAAGATGTTAAATAATTA 18rs21_all.seq
2949 AAAACGTTGTAACTGATGAACCAAAAAACAGATAAAGATGTTAAAAAATTA coh1_all.seq
2949 AAAACGTTGTAACTGATGAACCAAAAAACAGATAAAGATGTTAAAAAATTA cjb111_all.seq
2951 AAAACGTTGTAACTGATGAACCAAAAAACAGATAAAGATGTTAAAAAATTA nem316_all.seq
2950 AAAACGTTGTAACTGATGAACCAAAAAACAGATAAAGATGTTAAAAAATTA a909_all.seq

GGTCAGGACGATGCAGGTTATACGATTGGTGAAGAATTCAAATGGTTCTT Majority
        3010      3020      3030      3040      3050
998 GGTCAGGACGATGCAGGTTATACGATTGGTGAAGAATTCAAATGGTTCTT 2603_all.seq
767 GGTCAGGACGATGCAGGTTATACGATTGGTGAAGAATTCAAATGGTTCTT 18rs21_all.seq
999 GGTCAGGACGATGCAGGTTATACGATTGGTGAAGAATTCAAATGGTTCTT coh1_all.seq
999 GGTCAGGACGATGCAGGTTATACGATTGGTGAAGAATTCAAATGGTTCTT cjb111_all.seq
001 GGTCAGGACGATGCAGGTTATACGATTGGTGAAGAATTCAAATGGTTCTT nem316_all.seq
000 GGTCAGGACGATGCAGGTTATACGATTGGTGAAGAATTCAAATGGTTCTT a909_all.seq

GAAATCTACAATCCCTGCCAATTTAGGTGACTATGAAAAATTTGAAATTA Majority
        3060      3070      3080      3090      3100
048 GAAATCTACAATCCCTGCCAATTTAGGTGACTATGAAAAATTTGAAATTA 2603_all.seq
817 GAAATCTACAATCCCTGCCAATTTAGGTGACTATGAAAAATTTGAAATTA 18rs21_all.seq
049 GAAATCTACAATCCCTGCCAATTTAGGTGACTATGAAAAATTTGAAATTA coh1_all.seq
049 GAAATCTACAATCCCTGCCAATTTAGGTGACTATGAAAAATTTGAAATTA cjb111_all.seq
051 GAAATCTACAATCCCTGCCAATTTAGGTGACTATGAAAAATTTGAAATTA nem316_all.seq
050 GAAATCTACAATCCCTGCCAATTTAGGTGACTATGAAAAATTTGAAATTA a909_all.seq

CTGATAAATTTGCAGATGGCTTGACTTATAAATCTGTTGGAAAAATCAAG Majority
        3110      3120      3130      3140      3150
098 CTGATAAATTTGCAGATGGCTTGACTTATAAATCTGTTGGAAAAATCAAG 2603_all.seq
067 CTGATAAATTTGCAGATGGCTTGACTTATAAATCTGTTGGAAAAATCAAG 18rs21_all.seq
099 CTGATAAATTTGCAGATGGCTTGACTTATAAATCTGTTGGAAAAATCAAG coh1_all.seq
099 CTGATAAATTTGCAGATGGCTTGACTTATAAATCTGTTGGAAAAATCAAG cjb111_all.seq
101 CTGATAAATTTGCAGATGGCTTGACTTATAAATCTGTTGGAAAAATCAAG nem316_all.seq
100 CTGATAAATTTGCAGATGGCTTGACTTATAAATCTGTTGGAAAAATCAAG a909_all.seq
```

FIGURE 18 H

```
                ATTGGTTCGAAAACACTGAATAGAGATGAGCACTACACTATTGATGAACC Majority
                    3160      3170      3180      3190      3200
148 ATTGGTTCGAAAACACTGAATAGAGATGAGCACTACACTATTGATGAACC 2603_all.seq
917 ATTGGTTCGAAAACACTGAATAGAGATGAGCACTACACTATTGATGAACC 18rs21_all.seq
149 ATTGGTTCGAAAACACTGAATAGAGATGAGCACTACACTATTGATGAACC coh1_all.seq
149 ATTGGTTCGAAAACACTGAATAGAGATGAGCACTACACTATTGATGAACC cjb111_all.seq
151 ATTGGTTCGAAAACACTGAATAGAGATGAGCACTACACTATTGATGAACC nem316_all.seq
150 ATTGGTTCGAAAACACTGAATAGAGATGAGCACTACACTATTGATGAACC a909_all.seq

                AACAGTTGATAACCAAAATACATTAAAAATTACGTTTAAACCAGAGAAAT Majority
                    3210      3220      3230      3240      3250
198 AACAGTTGATAACCAAAATACATTAAAAATTACGTTTAAACCAGAGAAAT 2603_all.seq
967 AACAGTTGATAACCAAAATACATTAAAAATTACGTTTAAACCAGAGAAAT 18rs21_all.seq
199 AACAGTTGATAACCAAAATACATTAAAAATTACGTTTAAACCAGAGAAAT coh1_all.seq
199 AACAGTTGATAACCAAAATACATTAAAAATTACGTTTAAACCAGAGAAAT cjb111_all.seq
201 AACAGTTGATAACCAAAATACATTAAAAATTACGTTTAAACCAGAGAAAT nem316_all.seq
200 AACAGTTGATAACCAAAATACATTAAAAATTACGTTTAAACCAGAGAAAT a909_all.seq

                TTAAAGAAATTGCTGAGCTACTTAAAGGAATGACCCTTGTTAAAAATCAA Majority
                    3260      3270      3280      3290      3300
248 TTAAAGAAATTGCTGAGCTACTTAAAGGAATGACCCTTGTTAAAAATCAA 2603_all.seq
017 TTAAAGAAATTGCTGAGCTACTTAAAGGAATGACCCTTGTTAAAAATCAA 18rs21_all.seq
249 TTAAAGAAATTGCTGAGCTACTTAAAGGAATGACCCTTGTTAAAAATCAA coh1_all.seq
249 TTAAAGAAATTGCTGAGCTACTTAAAGGAATGACCCTTGTTAAAAATCAA cjb111_all.seq
251 TTAAAGAAATTGCTGAGCTACTTAAAGGAATGACCCTTGTTAAAAATCAA nem316_all.seq
250 TTAAAGAAATTGCTGAGCTACTTAAAGGAATGACCCTTGTTAAAAATCAA a909_all.seq

                GATGCTCTTGATAAAGCTACTGCAAATACAGATGATGCGGCATTTTTGGA Majority
                    3310      3320      3330      3340      3350
298 GATGCTCTTGATAAAGCTACTGCAAATACAGATGATGCGGCATTTTTGGA 2603_all.seq
967 GATGCTCTTGATAAAGCTACTGCAAATACAGATGATGCGGCATTTTTGGA 18rs21_all.seq
299 GATGCTCTTGATAAAGCTACTGCAAATACAGATGATGCGGCATTTTTGGA coh1_all.seq
299 GATGCTCTTGATAAAGCTACTGCAAATACAGATGATGCGGCATTTTTGGA cjb111_all.seq
301 GATGCTCTTGATAAAGCTACTGCAAATACAGATGATGCGGCATTTTTGGA nem316_all.seq
300 GATGCTCTTGATAAAGCTACTGCAAATACAGATGATGCGGCATTTTTGGA a909_all.seq

                AATTCCAGTTGCATCAACTATTAATGAAAAAGCAGTTTTAGGAAAAGCAA Majority
                    3360      3370      3380      3390      3400
348 AATTCCAGTTGCATCAACTATTAATGAAAAAGCAGTTTTAGGAAAAGCAA 2603_all.seq
317 AATTCCAGTTGCATCAACTATTAATGAAAAAGCAGTTTTAGGAAAAGCAA 18rs21_all.seq
349 AATTCCAGTTGCATCAACTATTAATGAAAAAGCAGTTTTAGGAAAAGCAA coh1_all.seq
349 AATTCCAGTTGCATCAACTATTAATGAAAAAGCAGTTTTAGGAAAAGCAA cjb111_all.seq
351 AATTCCAGTTGCATCAACTATTAATGAAAAAGCAGTTTTAGGAAAAGCAA nem316_all.seq
350 AATTCCAGTTGCATCAACTATTAATGAAAAAGCAGTTTTAGGAAAAGCAA a909_all.seq

                TTGAAAATACTTTTGAACTTCAATATGACCATACTCCTGATAAAGCTGAC Majority
                    3410      3420      3430      3440      3450
398 TTGAAAATACTTTTGAACTTCAATATGACCATACTCCTGATAAAGCTGAC 2603_all.seq
367 TTGAAAATACTTTTGAACTTCAATATGACCATACTCCTGATAAAGCTGAC 18rs21_all.seq
399 TTGAAAATACTTTTGAACTTCAATATGACCATACTCCTGATAAAGCTGAC coh1_all.seq
399 TTGAAAATACTTTTGAACTTCAATATGACCATACTCCTGATAAAGCTGAC cjb111_all.seq
401 TTGAAAATACTTTTGAACTTCAATATGACCATACTCCTGATAAAGCTGAC nem316_all.seq
400 TTGAAAATACTTTTGAACTTCAATATGACCATACTCCTGATAAAGCTGAC a909_all.seq

                AATCCAAAACCATCTAATCCTCCAAGAAAACCAGAAGTTCATACTGGTGG Majority
                    3460      3470      3480      3490      3500
448 AATCCAAAACCATCTAATCCTCCAAGAAAACCAGAAGTTCATACTGGTGG 2603_all.seq
417 AATCCAAAACCATCTAATCCTCCAAGAAAACCAGAAGTTCATACTGGTGG 18rs21_all.seq
449 AATCCAAAACCATCTAATCCTCCAAGAAAACCAGAAGTTCATACTGGTGG coh1_all.seq
449 AATCCAAAACCATCTAATCCTCCAAGAAAACCAGAAGTTCATACTGGTGG cjb111_all.seq
451 AATCCAAAACCATCTAATCCTCCAAGAAAACCAGAAGTTCATACTGGTGG nem316_all.seq
450 AATCCAAAACCATCTAATCCTCCAAGAAAACCAGAAGTTCATACTGGTGG a909_all.seq
```

FIGURE 18 I

```
                GAAACGATTGTAAAGAAAGACTCAACAGAAACACAAACACTAGGTGGTG Majority
                    3510        3520        3530        3540        3550
3498  GAAACGATTTGTAAAGAAAGACTCAACAGAAACACAAACACTAGGTGGTG 2603_ail.seq
3267  GAAACGATTTGTAAAGAAAGACTCAACAGAAACACAAACACTAGGTGGTG 18rs21_ail.seq
3499  GAAACGATTTGTAAAGAAAGACTCAACAGAAACACAAACACTAGGTGGTG coh1_ail.seq
3499  GAAACGATTTGTAAAGAAAGACTCAACAGAAACACAAACACTAGGTGGTG cjb111_ail.seq
3501  GAAACGATTTGTAAAGAAAGACTCAACAGAAACACAAACACTAGGTGGTG nem316_ail.seq
3500  GAAACGATTTGTAAAGAAAGACTCAACAGAAACACAAACACTAGGTGGTG a909_ail.seq

                CTGAGTTTGATTTGTTGGCTTCTGATGGGACAGCAGTAAAATGGACAGAT Majority
                    3560        3570        3580        3590        3600
3548  CTGAGTTTGATTTGTTGGCTTCTGATGGGACAGCAGTAAAATGGACAGAT 2603_ail.seq
3317  CTGAGTTTGATTTGTTGGCTTCTGATGGGACAGCAGTAAAATGGACAGAT 18rs21_ail.seq
3549  CTGAGTTTGATTTGTTGGCTTCTGATGGGACAGCAGTAAAATGGACAGAT coh1_ail.seq
3549  CTGAGTTTGATTTGTTGGCTTCTGATGGGACAGCAGTAAAATGGACAGAT cjb111_ail.seq
3551  CTGAGTTTGATTTGTTGGCTTCTGATGGGACAGCAGTAAAATGGACAGAT nem316_ail.seq
3550  CTGAGTTTGATTTGTTGGCTTCTGATGGGACAGCAGTAAAATGGACAGAT a909_ail.seq

                GCTCTTATTAAAGCGAATACTAATAAAAACTATATTGCTGGAGAAGCTGT Majority
                    3610        3620        3630        3640        3650
598   GCTCTTATTAAAGCGAATACTAATAAAAACTATATTGCTGGAGAAGCTGT 2603_ail.seq
367   GCTCTTATTAAAGCGAATACTAATAAAAACTATATTGCTGGAGAAGCTGT 18rs21_ail.seq
599   GCTCTTATTAAAGCGAATACTAATAAAAACTATATTGCTGGAGAAGCTGT coh1_ail.seq
599   GCTCTTATTAAAGCGAATACTAATAAAAACTATATTGCTGGAGAAGCTGT cjb111_ail.seq
601   GCTCTTATTAAAGCGAATACTAATAAAAACTATATTGCTGGAGAAGCTGT nem316_ail.seq
600   GCTCTTATTAAAGCGAATACTAATAAAAACTATATTGCTGGAGAAGCTGT a909_ail.seq

                TACTGGGCAACCAATCAAATTGAAATCACATACAGACGGTACGTTTGAGA Majority
                    3660        3670        3680        3690        3700
648   TACTGGGCAACCAATCAAATTGAAATCACATACAGACGGTACGTTTGAGA 2603_ail.seq
417   TACTGGGCAACCAATCAAATTGAAATCACATACAGACGGTACGTTTGAGA 18rs21_ail.seq
649   TACTGGGCAACCAATCAAATTGAAATCACATACAGACGGTACGTTTGAGA coh1_ail.seq
649   TACTGGGCAACCAATCAAATTGAAATCACATACAGACGGTACGTTTGAGA cjb111_ail.seq
651   TACTGGGCAACCAATCAAATTGAAATCACATACAGACGGTACGTTTGAGA nem316_ail.seq
650   TACTGGGCAACCAATCAAATTGAAATCACATACAGACGGTACGTTTGAGA a909_ail.seq

                TTAAAGGTTTGGCTTATGCAGTTGATGCGAATGCAGAGGGTACAGCAGTA Majority
                    3710        3720        3730        3740        3750
698   TTAAAGGTTTGGCTTATGCAGTTGATGCGAATGCAGAGGGTACAGCAGTA 2603_ail.seq
467   TTAAAGGTTTGGCTTATGCAGTTGATGCGAATGCAGAGGGTACAGCAGTA 18rs21_ail.seq
699   TTAAAGGTTTGGCTTATGCAGTTGATGCGAATGCAGAGGGTACAGCAGTA coh1_ail.seq
699   TTAAAGGTTTGGCTTATGCAGTTGATGCGAATGCAGAGGGTACAGCAGTA cjb111_ail.seq
701   TTAAAGGTTTGGCTTATGCAGTTGATGCGAATGCAGAGGGTACAGCAGTA nem316_ail.seq
700   TTAAAGGTTTGGCTTATGCAGTTGATGCGAATGCAGAGGGTACAGCAGTA a909_ail.seq

                ACTTACAAATTAAAAGAAACAAAAGCACCAGAAGGTTATGTAATCCCTGA Majority
                    3760        3770        3780        3790        3800
748   ACTTACAAATTAAAAGAAACAAAAGCACCAGAAGGTTATGTAATCCCTGA 2603_ail.seq
517   ACTTACAAATTAAAAGAAACAAAAGCACCAGAAGGTTATGTAATCCCTGA 18rs21_ail.seq
749   ACTTACAAATTAAAAGAAACAAAAGCACCAGAAGGTTATGTAATCCCTGA coh1_ail.seq
749   ACTTACAAATTAAAAGAAACAAAAGCACCAGAAGGTTATGTAATCCCTGA cjb111_ail.seq
751   ACTTACAAATTAAAAGAAACAAAAGCACCAGAAGGTTATGTAATCCCTGA nem316_ail.seq
750   ACTTACAAATTAAAAGAAACAAAAGCACCAGAAGGTTATGTAATCCCTGA a909_ail.seq

                TAAAGAAATCGAGTTTACAGTATCACAAACATCTTATAATACAAAACCAA Majority
                    3810        3820        3830        3840        3850
798   TAAAGAAATCGAGTTTACAGTATCACAAACATCTTATAATACAAAACCAA 2603_ail.seq
567   TAAAGAAATCGAGTTTACAGTATCACAAACATCTTATAATACAAAACCAA 18rs21_ail.seq
799   TAAAGAAATCGAGTTTACAGTATCACAAACATCTTATAATACAAAACCAA coh1_ail.seq
799   TAAAGAAATCGAGTTTACAGTATCACAAACATCTTATAATACAAAACCAA cjb111_ail.seq
801   TAAAGAAATCGAGTTTACAGTATCACAAACATCTTATAATACAAAACCAA nem316_ail.seq
800   TAAAGAAATCGAGTTTACAGTATCACAAACATCTTATAATACAAAACCAA a909_ail.seq
```

FIGURE 18 J

```
                                                
```

```
                TCAGTTGACGATTGTTCATCTTGAAGCAAGGGATATTGATCGTCCAAATC Majority
                    4210      4220      4230      4240      4250
4198 TCAGTTGACGATTGTTCATCTTGAAGCAAGGGATATTGATCGTCCAAATC 2603_all.seq
3967 TCAGTTGACGATTGTTCATCTTGAAGCAAGGGATATTGATCGTCCAAATC 2603_all.seq
4199 TCAGTTGACGATTGTTCATCTTGAAGCAAGGGATATTGATCGTCCAAATC 18rs21_all.seq
4199 TCAGTTGACGATTGTTCATCTTGAAGCAAGGGATATTGATCGTCCAAATC cohl_all.seq
4201 TCAGTTGACGATTGTTCATCTTGAAGCAAGGGATATTGATCGTCCAAATC cjb111_all.seq
4200 TCAGTTGACGATTGTTCATCTTGAAGCAAGGGATATTGATCGTCCAAATC nem316_all.seq
4248 TCAGTTGACGATTGTTCATCTTGAAGCAAGGGATATTGATCGTCCAAATC a909_all.seq

                CACAGTTGGAGATTGCCCCTAAAGAAGGGACTCCAATTGAAGGAGTACTC Majority
                    4260      4270      4280      4290      4300
4248 CACAGTTGGAGATTGCCCCTAAAGAAGGGACTCCAATTGAAGGAGTACTC 2603_all.seq
4017 CACAGTTGGAGATTGCCCCTAAAGAAGGGACTCCAATTGAAGGAGTACTC 2603_all.seq
4249 CACAGTTGGAGATTGCCCCTAAAGAAGGGACTCCAATTGAAGGAGTACTC 18rs21_all.seq
4249 CACAGTTGGAGATTGCCCCTAAAGAAGGGACTCCAATTGAAGGAGTACTC cohl_all.seq
4251 CACAGTTGGAGATTGCCCCTAAAGAAGGGACTCCAATTGAAGGAGTACTC cjb111_all.seq
4250 CACAGTTGGAGATTGCCCCTAAAGAACGGACTCCAATTGAAGGAGTACTC nem316_all.seq
4298 CACAGTTGGAGATTGCCCCTAAAGAAGGGACTCCAATTGAAGGAGTACTC a909_all.seq

                TATCAGTTGTACCAATTAAAATCAACTGAAGATGGCGATTTGTTGGCACA Majority
                    4310      4320      4330      4340      4350
4298 TATCAGTTGTACCAATTAAAATCAACTGAAGATGGCGATTTGTTGGCACA 2603_all.seq
4067 TATCAGTTGTACCAATTAAAATCAACTGAAGATGGCGATTTGTTGGCACA 2603_all.seq
4299 TATCAGTTGTACCAATTAAAATCAACTGAAGATGGCGATTTGTTGGCACA 18rs21_all.seq
4299 TATCAGTTGTACCAATTAAAATCAACTGAAGATGGCGATTTGTTGGCACA cohl_all.seq
4301 TATCAGTTGTACCAATTAAAATCAACTGAAGATGGCGATTTGTTGGCACA cjb111_all.seq
4300 TATCAGTTGTACCAATTAAAATCAACTGAAGATGGCGATTTGTTGGCACA nem316_all.seq
4348 TATCAGTTGTACCAATTAAAATCAACTGAAGATGGCGATTTGTTGGCACA a909_all.seq

                TTGGAATTCCCTAACTATCACAGAATTGAAAAAACAGGCGCAGCAGGTTT Majority
                    4360      4370      4380      4390      4400
4348 TTGGAATTCCCTAACTATCACAGAATTGAAAAAACAGGCGCAGCAGGTTT 2603_all.seq
4117 TTGGAATTCCCTAACTATCACAGAATTGAAAAAACAGGCGCAGCAGGTTT 2603_all.seq
4349 TTGGAATTCCCTAACTATCACAGAATTGAAAAAACAGGCGCAGCAGGTTT 18rs21_all.seq
4349 TTGGAATTCCCTAACTATCACAGAATTGAAAAAACAGGCGCAGCAGGTTT cohl_all.seq
4351 TTGGAATTCCCTAACTATCACAGAATTGAAAAAACAGGCGCAGCAGGTTT cjb111_all.seq
4350 TTGGAATTCCCTAACTATCACAGAATTGAAAAAACAGGCGCAGCAGGTTT nem316_all.seq
4398 TTGGAATTCCCTAACTATCACAGAATTGAAAAAACAGGCGCAGCAGGTTT a909_all.seq

                TTGAAGCCACTACTAATCAACAAGGAAAGGCTACATTTAACCAACTACCA Majority
                    4410      4420      4430      4440      4450
4398 TTGAAGCCACTACTAATCAACAAGGAAAGGCTACATTTAACCAACTACCA 2603_all.seq
4167 TTGAAGCCACTACTAATCAACAAGGAAAGGCTACATTTAACCAACTACCA 2603_all.seq
4399 TTGAAGCCACTACTAATCAACAAGGAAAGGCTACATTTAACCAACTACCA 18rs21_all.seq
4399 TTGAAGCCACTACTAATCAACAAGGAAAGGCTACATTTAACCAACTACCA cohl_all.seq
4401 TTGAAGCCACTACTAATCAACAAGGAAAGGCTACATTTAACCAACTACCA cjb111_all.seq
4400 TTGAAGCCACTACTAATCAACAAGGAAAGGCTACATTTAACCAACTACCA nem316_all.seq
4448 TTGAAGCCACTACTAATCAACAAGGAAAGGCTACATTTAACCAACTACCA a909_all.seq

                GATGGAATTTATTATGGTCTGGCGGTTAAAGCCGGTGAAAAAAATCGTAA Majority
                    4460      4470      4480      4490      4500
4448 GATGGAATTTATTATGGTCTGGCGGTTAAAGCCGGTGAAAAAAATCGTAA 2603_all.seq
4217 GATGGAATTTATTATGGTCTGGCGGTTAAAGCCGGTGAAAAAAATCGTAA 2603_all.seq
4449 GATGGAATTTATTATGGTCTGGCGGTTAAAGCCGGTGAAAAAAATCGTAA 18rs21_all.seq
4449 GATGGAATTTATTATGGTCTGGCGGTTAAAGCCGGTGAAAAAAATCGTAA cohl_all.seq
4451 GATGGAATTTATTATGGTCTGGCGGTTAAAGCCGGTGAAAAAAATCGTAA cjb111_all.seq
4450 GATGGAATTTATTATGGTCTGGCGGTTAAAGCCGGTGAAAAAAATCGTAA nem316_all.seq
4498 GATGGAATTTATTATGGTCTGGCGGTTAAAGCCGGTGAAAAAAATCGTAA a909_all.seq

                TGTCTCAGCTTTCTTGGTTGACTTGTCTGAGGATAAAGTGATTTATCCTA Majority
                    4510      4520      4530      4540      4550
4498 TGTCTCAGCTTTCTTGGTTGACTTGTCTGAGGATAAAGTGATTTATCCTA 2603_all.seq
4267 TGTCTCAGCTTTCTTGGTTGACTTGTCTGAGGATAAAGTGATTTATCCTA 2603_all.seq
4499 TGTCTCAGCTTTCTTGGTTGACTTGTCTGAGGATAAAGTGATTTATCCTA 18rs21_all.seq
4499 TGTCTCAGCTTTCTTGGTTGACTTGTCTGAGGATAAAGTGATTTATCCTA cohl_all.seq
4501 TGTCTCAGCTTTCTTGGTTGACTTGTCTGAGGATAAAGTGATTTATCCTA cjb111_all.seq
4500 TGTCTCAGCTTTCTTGGTTGACTTGTCTGAGGATAAAGTGATTTATCCTA nem316_all.seq
4548 TGTCTCAGCTTTCTTGGTTGACTTGTCTGAGGATAAAGTGATTTATCCTA a909_all.seq
```

FIGURE 18 L

```
                A A A T C A T C T G G T C C A C A G G T G A G T T G G A C T T G C T T A A A G T T G G T G T G G A T Majority
                        4560        4570        4580        4590        4600
1548  A A A T C A T C T G G T C C A C A G G T G A G T T G G A C T T G C T T A A A G T T G G T G T G G A T 2603_ail.seq
1317  A A A T C A T C T G G T C C A C A G G T G A G T T G G A C T T G C T T A A A G T T G G T G T G G A T 18rs21_ail.seq
1549  A A A T C A T C T G G T C C A C A G G T G A G T T G G A C T T G C T T A A A G T T G G T G T G G A T cohl_ail.seq
1549  A A A T C A T C T G G T C C A C A G G T G A G T T G G A C T T G C T T A A A G T T G G T G T G G A T cjbl11_ail.seq
1551  A A A T C A T C T G G T C C A C A G G T G A G T T G G A C T T G C T T A A A G T T G G T G T G G A T nem316_ail.seq
1550  A A A T C A T C T G G T C C A C A G G T G A G T T G G A C T T G C T T A A A G T T G G T G T G G A T a909_ail.seq

                G G T G A T A C C A A A A A A C C A C T A G C A G G C G T T G T C T T T G A A C T T T A T G A A A A Majority
                        4610        4620        4630        4640        4650
1598  G G T G A T A C C A A A A A A C C A C T A G C A G G C G T T G T C T T T G A A C T T T A T G A A A A 2603_ail.seq
1367  G G T G A T A C C A A A A A A C C A C T A G C A G G C G T T G T C T T T G A A C T T T A T G A A A A 18rs21_ail.seq
1599  G G T G A T A C C A A A A A A C C A C T A G C A G G C G T T G T C T T T G A A C T T T A T G A A A A cohl_ail.seq
1599  G G T G A T A C C A A A A A A C C A C T A G C A G G C G T T G T C T T T G A A C T T T A T G A A A A cjbl11_ail.seq
1601  G G T G A T A C C A A A A A A C C A C T A G C A G G C G T T G T C T T T G A A C T T T A T G A A A A nem316_ail.seq
1600  G G T G A T A C C A A A A A A A C A C T A G C A G G C G T T G T C T T T G A A C T T T A T G A A A A a909_ail.seq

                G A A T G G T A G G A C T C C T A T T C G T G T G A A A A A T G G G G T G C A T T C T C A A G A T A Majority
                        4660        4670        4680        4690        4700
1648  G A A T G G T A G G A C T C C T A T T C G T G T G A A A A A T G G G G T G C A T T C T C A A G A T A 2603_ail.seq
1417  G A A T G G T A G G A C T C C T A T T C G T G T G A A A A A T G G G G T G C A T T C T C A A G A T A 18rs21_ail.seq
1649  G A A T G G T A G G A C T C C T A T T C G T G T G A A A A A T G G G G T G C A T T C T C A A G A T A cohl_ail.seq
1649  G A A T G G T A G G A C T C C T A T T C G T G T G A A A A A T G G G G T G C A T T C T C A A G A T A cjbl11_ail.seq
1651  G A A T G G T A G G A C T C C T A T T C G T G T G A A A A A T G G G G T G C A T T C T C A A G A T A nem316_ail.seq
1650  G A A T G G T A G G A C T C C T A T T C G T G T G A A A A A T G G G G T G C A T T C T C A A G A T A a909_ail.seq

                T T G A C G C T G C A A A A C A T T T A G A A A C A G A T T C A T C A G G G C A T A T C A G A A T T Majority
                        4710        4720        4730        4740        4750
1698  T T G A C G C T G C A A A A C A T T T A G A A A C A G A T T C A T C A G G G C A T A T C A G A A T T 2603_ail.seq
1467  T T G A C G C T G C A A A A C A T T T A G A A A C A G A T T C A T C A G G G C A T A T C A G A A T T 18rs21_ail.seq
1699  T T G A C G C T G C A A A A C A T T T A G A A A C A G A T T C A T C A G G G C A T A T C A G A A T T cohl_ail.seq
1699  T T G A C G C T G C A A A A C A T T T A G A A A C A G A T T C A T C A G G G C A T A T C A G A A T T cjbl11_ail.seq
1701  T T G A C G C T G C A A A A C A T T T A G A A A C A G A T T C A T C A G G G C A T A T C A G A A T T nem316_ail.seq
1700  T T G A C G C T G C A A A A C A T T T A G A A A C A G A T T C A T C A G G G C A T A T C A G A A T T a909_ail.seq

                T C C G G G C T C A T C C A T G G G G A C T A T G T C T T A A A A G A A A T C G A G A C A C A G T C Majority
                        4760        4770        4780        4790        4800
1748  T C C G G G C T C A T C C A T G G G G A C T A T G T C T T A A A A G A A A T C G A G A C A C A G T C 2603_ail.seq
1517  T C C G G G C T C A T C C A T G G G G A C T A T G T C T T A A A A G A A A T C G A G A C A C A G T C 18rs21_ail.seq
1749  T C C G G G C T C A T C C A T G G G G A C T A T G T C T T A A A A G A A A T C G A G A C A C A G T C cohl_ail.seq
1749  T C C G G G C T C A T C C A T G G G G A C T A T G T C T T A A A A G A A A T C G A G A C A C A G T C cjbl11_ail.seq
1751  T C C G G G C T C A T C C A T G G G G A C T A T G T C T T A A A A G A A A T C G A G A C A C A G T C nem316_ail.seq
1750  T C C G G G C T C A T C C A T G G G G A C T A T G T C T T A A A A G A A A T C G A G A C A C A G T C a909_ail.seq

                A G G A T A T C A G A T C G G A C A G G C A G A G A C T G C T G T G A C T A T T G A A A A A T C A A Majority
                        4810        4820        4830        4840        4850
1798  A G G A T A T C A G A T C G G A C A G G C A G A G A C T G C T G T G A C T A T T G A A A A A T C A A 2603_ail.seq
1567  A G G A T A T C A G A T C G G A C A G G C A G A G A C T G C T G T G A C T A T T G A A A A A T C A A 18rs21_ail.seq
1799  A G G A T A T C A G A T C G G A C A G G C A G A G A C T G C T G T G A C T A T T G A A A A A T C A A cohl_ail.seq
1799  A G G A T A T C A G A T C G G A C A G G C A G A G A C T G C T G T G A C T A T T G A A A A A T C A A cjbl11_ail.seq
1801  A G G A T A T C A G A T C G G A C A G G C A G A G A C T G C T G T G A C T A T T G A A A A A T C A A nem316_ail.seq
1800  A G G A T A T C A G A T C G G A C A G G C A G A G A C T G C T G T G A C T A T T G A A A A A T C A A a909_ail.seq

                A A A C A G T A A C A G T A A C G A T T G A A A A T A A A A A A G T T C C G A C A C C T A A A G T G Majority
                        4860        4870        4880        4890        4900
1848  A A A C A G T A A C A G T A A C G A T T G A A A A T A A A A A A G T T C C G A C A C C T A A A G T G 2603_ail.seq
1617  A A A C A G T A A C A G T A A C G A T T G A A A A T A A A A A A G T T C C G A C A C C T A A A G T G 18rs21_ail.seq
1849  A A A C A G T A A C A G T A A C G A T T G A A A A T A A A A A A G T T C C G A C A C C T A A A G T G cohl_ail.seq
1849  A A A C A G T A A C A G T A A C G A T T G A A A A T A A A A A A G T T C C G A C A C C T A A A G T G cjbl11_ail.seq
1851  A A A C A G T A A C A G T A A C G A T T G A A A A T A A A A A A G T T C C G A C A C C T A A A G T G nem316_ail.seq
1850  A A A C A G T A A C A G T A A C G A T T G A A A A T A A A A A A G T T C C G A C A C C T A A A G T G a909_ail.seq
```

FIGURE 18 M

```
CCATCTCGAGGAGGTCTTATTCCCAAAACAGGTGAGCAACAGGCAATGGC Majority
          4910      4920      4930      4940      4950
4898 CCATCTCGAGGAGGTCTTATTCCCAAAACAGGTGAGCAACAGGCAATGGC 2603_ail.seq
4667 CCATCTCGAGGAGGTCTTATTCCCAAAACAGGTGAGCAACAGGCAATGGC 18rs21_ail.seq
4899 CCATCTCGAGGAGGTCTTATTCCCAAAACAGGTGAGCAACAGGCAATGGC coh1_ail.seq
4899 CCATCTCGAGGAGGTCTTATTCCCAAAACAGGTGAGCAACAGGCAATGGC cjb111_ail.seq
4901 CCATCTCGAGGAGGTCTTATTCCCAAAACAGGTGAGCAACAGGCAATGGC nem316_ail.seq
4900 CCATCTCGAGGAGGTCTTATTCCCAAAACAGGTGAGCAACAGGCAATGGC a909_ail.seq

ACTTGTAATTATTGGTGGTATTTTAATTGCTTTAGCCTTACGATTACTAT Majority
          4960      4970      4980      4990      5000
4948 ACTTGTAATTATTGGTGGTATTTTAATTGCTTTAGCCTTACGATTACTAT 2603_ail.seq
4717 ACTTGTAATTATTGGTGGTATTTTAATTGCTTTAGCCTTACGATTACTAT 18rs21_ail.seq
4949 ACTTGTAATTATTGGTGGTATTTTAATTGCTTTAGCCTTACGATTACTAT coh1_ail.seq
4949 ACTTGTAATTATTGGTGGTATTTTAATTGCTTTAGCCTTACGATTACTAT cjb111_ail.seq
4951 ACTTGTAATTATTGGTGGTATTTTAATTGCTTTAGCCTTACGATTACTAT nem316_ail.seq
4950 ACTTGTAATTATTGGTGGTATTTTAATTGCTTTAGCCTTACGATTACTAT a909_ail.seq

CAAAACATCGGAAACATCAAAATAAGGATTAGCATGGGACAAAAATCAAA Majority
          5010      5020      5030      5040      5050
4998 CAAAACATCGGAAACATCAAAATAAGGATTAGCATGGGACAAAAATCAAA 2603_ail.seq
4767 CAAAACATCGGAAACATCAAAATAAGGATTAGCATGGGACAAAAATCAAA 18rs21_ail.seq
4999 CAAAACATCGGAAACATCAAAATAAGGATTAGCATGGGACAAAAATCAAA coh1_ail.seq
4999 CAAAACATCGGAAACATCAAAATAAGGATTAGCATGGGACAAAAATCAAA cjb111_ail.seq
5001 CAAAACATCGGAAACATCAAAATAAGGATTAGCATGGGACAAAAATCAAA nem316_ail.seq
5000 CAAAACATCGGAAACATCAAAATAAGGATTAGCATGGGACAAAAATCAAA a909_ail.seq

AATATCTCTAGCTACGAATATTCGTATATGGATTTTTCGTTTAATTTTCT Majority
          5060      5070      5080      5090      5100
5048 AATATCTCTAGCTACGAATATTCGTATATGGATTTTTCGTTTAATTTTCT 2603_ail.seq
4817 AATATCTCTAGCTACGAATATTCGTATATGGATTTTTCGTTTAATTTTCT 18rs21_ail.seq
5049 AATATCTCTAGCTACGAATATTCGTATATGGATTTTTCGTTTAATTTTCT coh1_ail.seq
5049 AATATCTCTAGCTACGAATATTCGTATATGGATTTTTCGTTTAATTTTCT cjb111_ail.seq
5051 AATATCTCTAGCTACGAATATTCGTATATGGATTTTTCGTTTAATTTTCT nem316_ail.seq
5050 AATATCTCTAGCTACGAATATTCGTATATGGATTTTTCGTTTAATTTTCT a909_ail.seq

TAGCGGGTTTCCTTGTTTTGGCATTTCCCATCGTTAGTCAGGTCATGTAC Majority
          5110      5120      5130      5140      5150
5098 TAGCGGGTTTCCTTGTTTTGGCATTTCCCATCGTTAGTCAGGTCATGTAC 2603_ail.seq
4867 TAGCGGGTTTCCTTGTTTTGGCATTTCCCATCGTTAGTCAGGTCATGTAC 18rs21_ail.seq
5099 TAGCGGGTTTCCTTGTTTTGGCATTTCCCATCGTTAGTCAGGTCATGTAC coh1_ail.seq
5099 TAGCGGGTTTCCTTGTTTTGGCATTTCCCATCGTTAGTCAGGTCATGTAC cjb111_ail.seq
5101 TAGCGGGTTTCCTTGTTTTGGCATTTCCCATCGTTAGTCAGGTCATGTAC nem316_ail.seq
5100 TAGCGGGTTTCCTTGTTTTGGCATTTCCCATCGTTAGTCAGGTCATGTAC a909_ail.seq

TTTCAAGCCTCTCACGCCAATATTAATGCTTTTAAAGAAGCTGTTACCAA Majority
          5160      5170      5180      5190      5200
5148 TTTCAAGCCTCTCACGCCAATATTAATGCTTTTAAAGAAGCTGTTACCAA 2603_ail.seq
4917 TTTCAAGCCTCTCACGCCAATATTAATGCTTTTAAAGAAGCTGTTACCAA 18rs21_ail.seq
5149 TTTCAAGCCTCTCACGCCAATATTAATGCTTTTAAAGAAGCTGTTACCAA coh1_ail.seq
5149 TTTCAAGCCTCTCACGCCAATATTAATGCTTTTAAAGAAGCTGTTACCAA cjb111_ail.seq
5151 TTTCAAGCCTCTCACGCCAATATTAATGCTTTTAAAGAAGCTGTTACCAA nem316_ail.seq
5150 TTTCAAGCCTCTCACGCCAATATTAATGCTTTTAAAGAAGCTGTTACCAA a909_ail.seq

GATTGACCGCGTGGAGATTAATCGGCGTTTAGAACTTGCTTATGCTTATA Majority
          5210      5220      5230      5240      5250
5198 GATTGACCGCGTGGAGATTAATCGGCGTTTAGAACTTGCTTATGCTTATA 2603_ail.seq
4967 GATTGACCGCGTGGAGATTAATCGGCGTTTAGAACTTGCTTATGCTTATA 18rs21_ail.seq
5199 GATTGACCGCGTGGAGATTAATCGGCGTTTAGAACTTGCTTATGCTTATA coh1_ail.seq
5199 GATTGACCGCGTGGAGATTAATCGGCGTTTAGAACTTGCTTATGCTTATA cjb111_ail.seq
5201 GATTGACCGCGTGGAGATTAATCGGCGTTTAGAACTTGCTTATGCTTATA nem316_ail.seq
5200 GATTGACCGCGTGGAGATTAATCGGCGTTTAGAACTTGCTTATGCTTATA a909_ail.seq
```

FIGURE 18 N

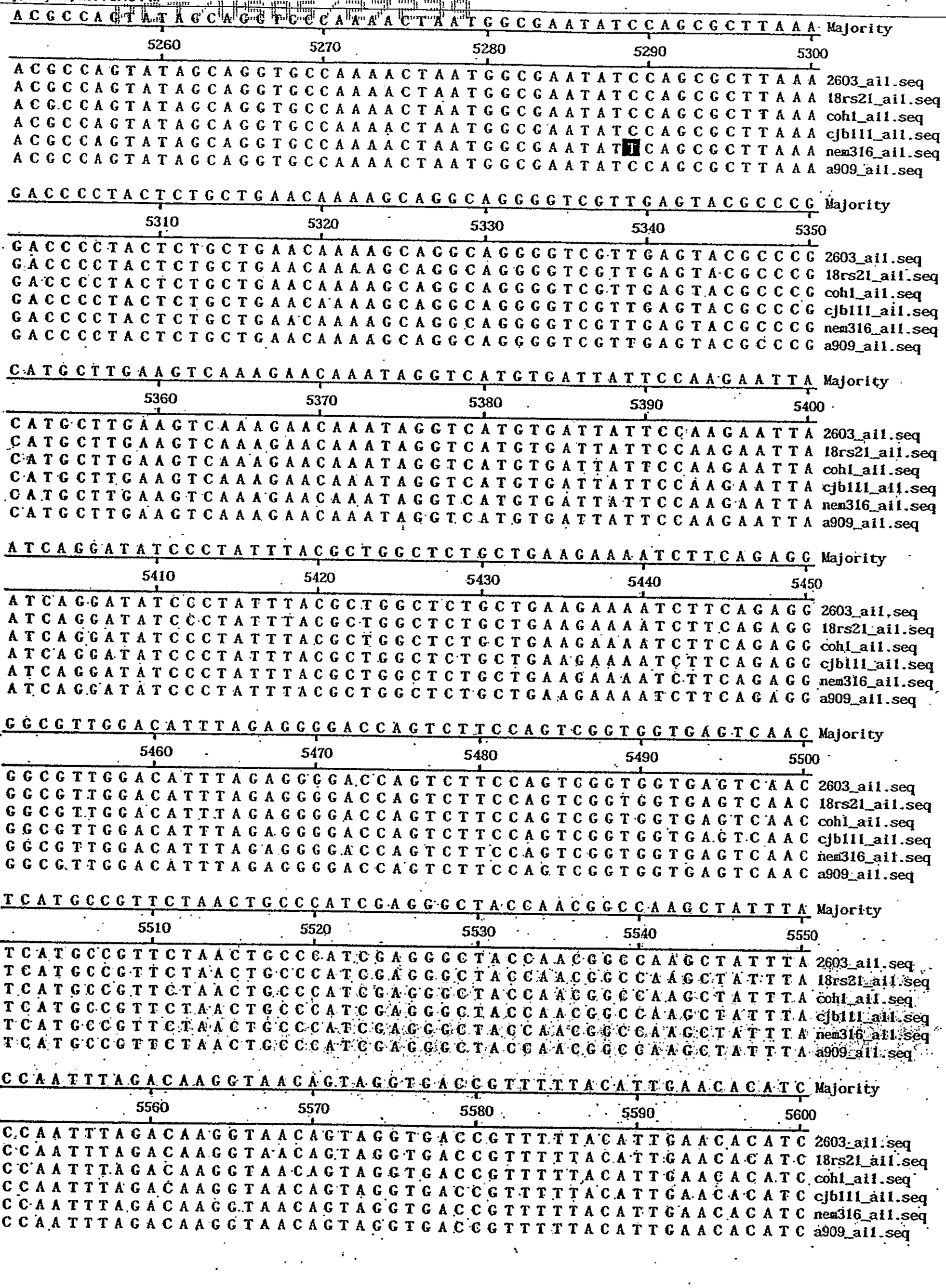

ACGCCAGTATAGCAGGTGCCAAAACTAATGGCGAATATCCAGCGCTTAAA Majority
5260 5270 5280 5290 5300
5248 ACGCCAGTATAGCAGGTGCCAAAACTAATGGCGAATATCCAGCGCTTAAA 2603_all.seq
5017 ACGCCAGTATAGCAGGTGCCAAAACTAATGGCGAATATCCAGCGCTTAAA 18rs21_all.seq
5249 ACGCCAGTATAGCAGGTGCCAAAACTAATGGCGAATATCCAGCGCTTAAA coh1_all.seq
5249 ACGCCAGTATAGCAGGTGCCAAAACTAATGGCGAATATCCAGCGCTTAAA cjb111_all.seq
5251 ACGCCAGTATAGCAGGTGCCAAAACTAATGGCGAATATTCAGCGCTTAAA nem316_all.seq
5250 ACGCCAGTATAGCAGGTGCCAAAACTAATGGCGAATATCCAGCGCTTAAA a909_all.seq
GACCCCTACTCTGCTGAACAAAAGCAGGCAGGGGTCGTTGAGTACGCCCG Majority
5310 5320 5330 5340 5350
5298 GACCCCTACTCTGCTGAACAAAAGCAGGCAGGGGTCGTTGAGTACGCCCG 2603_all.seq
5067 GACCCCTACTCTGCTGAACAAAAGCAGGCAGGGGTCGTTGAGTACGCCCG 18rs21_all.seq
5299 GACCCCTACTCTGCTGAACAAAAGCAGGCAGGGGTCGTTGAGTACGCCCG coh1_all.seq
5299 GACCCCTACTCTGCTGAACAAAAGCAGGCAGGGGTCGTTGAGTACGCCCG cjb111_all.seq
5301 GACCCCTACTCTGCTGAACAAAAGCAGGCAGGGGTCGTTGAGTACGCCCG nem316_all.seq
5300 GACCCCTACTCTGCTGAACAAAAGCAGGCAGGGGTCGTTGAGTACGCCCG a909_all.seq
CATGCTTGAAGTCAAAGAACAAATAGGTCATGTGATTATTCCAAGAATTA Majority
5360 5370 5380 5390 5400
5348 CATGCTTGAAGTCAAAGAACAAATAGGTCATGTGATTATTCCAAGAATTA 2603_all.seq
5117 CATGCTTGAAGTCAAAGAACAAATAGGTCATGTGATTATTCCAAGAATTA 18rs21_all.seq
5349 CATGCTTGAAGTCAAAGAACAAATAGGTCATGTGATTATTCCAAGAATTA coh1_all.seq
5349 CATGCTTGAAGTCAAAGAACAAATAGGTCATGTGATTATTCCAAGAATTA cjb111_all.seq
5351 CATGCTTGAAGTCAAAGAACAAATAGGTCATGTGATTATTCCAAGAATTA nem316_all.seq
5350 CATGCTTGAAGTCAAAGAACAAATAGGTCATGTGATTATTCCAAGAATTA a909_all.seq
ATCAGGATATCCCTATTTACGCTGGCTCTGCTGAAGAAAATCTTCAGAGG Majority
5410 5420 5430 5440 5450
5398 ATCAGGATATCCCTATTTACGCTGGCTCTGCTGAAGAAAATCTTCAGAGG 2603_all.seq
5167 ATCAGGATATCCCTATTTACGCTGGCTCTGCTGAAGAAAATCTTCAGAGG 18rs21_all.seq
5399 ATCAGGATATCCCTATTTACGCTGGCTCTGCTGAAGAAAATCTTCAGAGG coh1_all.seq
5399 ATCAGGATATCCCTATTTACGCTGGCTCTGCTGAAGAAAATCTTCAGAGG cjb111_all.seq
5401 ATCAGGATATCCCTATTTACGCTGGCTCTGCTGAAGAAAATCTTCAGAGG nem316_all.seq
5400 ATCAGGATATCCCTATTTACGCTGGCTCTGCTGAAGAAAATCTTCAGAGG a909_all.seq
GGCGTTGGACATTTAGAGGGGACCAGTCTTCCAGTCGGTGGTGAGTCAAC Majority
5460 5470 5480 5490 5500
448 GGCGTTGGACATTTAGAGGGGACCAGTCTTCCAGTCGGTGGTGAGTCAAC 2603_all.seq
217 GGCGTTGGACATTTAGAGGGGACCAGTCTTCCAGTCGGTGGTGAGTCAAC 18rs21_all.seq
449 GGCGTTGGACATTTAGAGGGGACCAGTCTTCCAGTCGGTGGTGAGTCAAC coh1_all.seq
449 GGCGTTGGACATTTAGAGGGGACCAGTCTTCCAGTCGGTGGTGAGTCAAC cjb111_all.seq
451 GGCGTTGGACATTTAGAGGGGACCAGTCTTCCAGTCGGTGGTGAGTCAAC nem316_all.seq
450 GGCGTTGGACATTTAGAGGGGACCAGTCTTCCAGTCGGTGGTGAGTCAAC a909_all.seq
TCATGCCGTTCTAACTGCCCATCGAGGGCTACCAACGGCCAAGCTATTTA Majority
5510 5520 5530 5540 5550
498 TCATGCCGTTCTAACTGCCCATCGAGGGCTACCAACGGCCAAGCTATTTA 2603_all.seq
267 TCATGCCGTTCTAACTGCCCATCGAGGGCTACCAACGGCCAAGCTATTTA 18rs21_all.seq
499 TCATGCCGTTCTAACTGCCCATCGAGGGCTACCAACGGCCAAGCTATTTA coh1_all.seq
499 TCATGCCGTTCTAACTGCCCATCGAGGGCTACCAACGGCCAAGCTATTTA cjb111_all.seq
501 TCATGCCGTTCTAACTGCCCATCGAGGGCTACCAACGGCCAAGCTATTTA nem316_all.seq
500 TCATGCCGTTCTAACTGCCCATCGAGGGCTACCAACGGCCAAGCTATTTA a909_all.seq
CCAATTTAGACAAGGTAACAGTAGGTGACCGTTTTTACATTGAACACATC Majority
5560 5570 5580 5590 5600
548 CCAATTTAGACAAGGTAACAGTAGGTGACCGTTTTTACATTGAACACATC 2603_all.seq
317 CCAATTTAGACAAGGTAACAGTAGGTGACCGTTTTTACATTGAACACATC 18rs21_all.seq
549 CCAATTTAGACAAGGTAACAGTAGGTGACCGTTTTTACATTGAACACATC coh1_all.seq
549 CCAATTTAGACAAGGTAACAGTAGGTGACCGTTTTTACATTGAACACATC cjb111_all.seq
551 CCAATTTAGACAAGGTAACAGTAGGTGACCGTTTTTACATTGAACACATC nem316_all.seq
550 CCAATTTAGACAAGGTAACAGTAGGTGACCGTTTTTACATTGAACACATC a909_all.seq

FIGURE 18 O

```
      G G C G G A A A G A T T G C T T A T C A G G T A G A C C A A A T C A A A G T T A T C G C C C C T G A Majority
                5610          5620          5630          5640          5650
598   G G C G G A A A G A T T G C T T A T C A G G T A G A C C A A A T C A A A G T T A T C G C C C C T G A 2603_ail.seq
367   G G C G G A A A G A T T G C T T A T C A G G T A G A C C A A A T C A A A G T T A T C G C C C C T G A 18rs21_ail.seq
599   G G C G G A A A G A T T G C T T A T C A G G T A G A C C A A A T C A A A G T T A T C G C C C C T G A coh1_ail.seq
599   G G C G G A A A G A T T G C T T A T C A G G T A G A C C A A A T C A A A G T T A T C G C C C C T G A cjb111_ail.seq
601   G G C G G A A A G A T T G C T T A T C A G G T A G A C C A A A T C A A A G T T A T C G C C C C T G A nem316_ail.seq
600   G G C G G A A A G A T T G C T T A T C A G G T A G A C C A A A T C A A A G T T A T C G C C C C T G A a909_ail.seq

      T C A G T T A G A G G A T T T G T A C G T G A T T C A A G G A G A A G A T C A C G T C A C C C T A T Majority
                5660          5670          5680          5690          5700
648   T C A G T T A G A G G A T T T G T A C G T G A T T C A A G G A G A A G A T C A C G T C A C C C T A T 2603_ail.seq
417   T C A G T T A G A G G A T T T G T A C G T G A T T C A A G G A G A A G A T C A C G T C A C C C T A T 18rs21_ail.seq
649   T C A G T T A G A G G A T T T G T A C G T G A T T C A A G G A G A A G A T C A C G T C A C C C T A T coh1_ail.seq
649   T C A G T T A G A G G A T T T G T A C G T G A T T C A A G G A G A A G A T C A C G T C A C C C T A T cjb111_ail.seq
651   T C A G T T A G A G G A T T T G T A C G T G A T T C A A G G A G A A G A T C A C G T C A C C C T A T nem316_ail.seq
650   T C A G T T A G A G G A T T T G T A C G T G A T T C A A G G A G A A G A T C A C G T C A C C C T A T a909_ail.seq

      T A A C T T G C A C A C C T T A T A T G A T A A A T A G T C A T C G C C T C C T C G T T C G A G G C Majority
                5710          5720          5730          5740          5750
698   T A A C T T G C A C A C C T T A T A T G A T A A A T A G T C A T C G C C T C C T C G T T C G A G G C 2603_ail.seq
467   T A A C T T G C A C A C C T T A T A T G A T A A A T A G T C A T C G C C T C C T C G T T C G A G G C 18rs21_ail.seq
699   T A A C T T G C A C A C C T T A T A T G A T A A A T A G T C A T C G C C T C C T C G T T C G A G G C coh1_ail.seq
699   T A A C T T G C A C A C C T T A T A T G A T A A A T A G T C A T C G C C T C C T C G T T C G A G G C cjb111_ail.seq
701   T A A C T T G C A C A C C T T A T A T G A T A A A T A G T C A T C G C C T C C T C G T T C G A G G C nem316_ail.seq
700   T A A C T T G C A C A C C T T A T A T G A T A A A T A G T C A T C G C C T C C T C G T T C G A G G C a909_ail.seq

      A A G C G A A T T C C T T A T G T G G A A A A A A C A G T G C A G A A A G A T T C A A A G A C C T T Majority
                5760          5770          5780          5790          5800
748   A A G C G A A T T C C T T A T G T G G A A A A A A C A G T G C A G A A A G A T T C A A A G A C C T T 2603_ail.seq
517   A A G C G A A T T C C T T A T G T G G A A A A A A C A G T G C A G A A A G A T T C A A A G A C C T T 18rs21_ail.seq
749   A A G C G A A T T C C T T A T G T G G A A A A A A C A G T G C A G A A A G A T T C A A A G A C C T T coh1_ail.seq
749   A A G C G A A T T C C T T A T G T G G A A A A A A C A G T G C A G A A A G A T T C A A A G A C C T T cjb111_ail.seq
751   A A G C G A A T T C C T T A T G T G G A A A A A A C A G T G C A G A A A G A T T C A A A G A C C T T nem316_ail.seq
750   A A G C G A A T T C C T T A T G T G G A A A A A A C A G T G C A G A A A G A T T C A A A G A C C T T a909_ail.seq

      C A G G C A A C A A C A A T A C C T A A C C T A T G C T A T G T G G G T A G T C G T T G G A C T T A Majority
                5810          5820          5830          5840          5850
798   C A G G C A A C A A C A A T A C C T A A C C T A T G C T A T G T G G G T A G T C G T T G G A C T T A 2603_ail.seq
567   C A G G C A A C A A C A A T A C C T A A C C T A T G C T A T G T G G G T A G T C G T T G G A C T T A 18rs21_ail.seq
799   C A G G C A A C A A C A A T A C C T A A C C T A T G C T A T G T G G G T A G T C G T T G G A C T T A coh1_ail.seq
799   C A G G C A A C A A C A A T A C C T A A C C T A T G C T A T G T G G G T A G T C G T T G G A C T T A cjb111_ail.seq
801   C A G G C A A C A A C A A T A C C T A A C C T A T G C T A T G T G G G T A G T C G T T G G A C T T A nem316_ail.seq
800   C A G G C A A C A A C A A T A C C T A A C C T A T G C T A T G T G G G T A G T C G T T G G A C T T A a909_ail.seq

      T C T T G C T G T C G C T T C T C A T T T G G T T T A A A A A G A C G A A A C A G A A A A A G C G G Majority
                5860          5870          5880          5890          5900
848   T C T T G C T G T C G C T T C T C A T T T G G T T T A A A A A G A C G A A A C A G A A A A A G C G G 2603_ail.seq
617   T C T T G C T G T C G C T T C T C A T T T G G T T T A A A A A G A C G A A A C A G A A A A A G C G G 18rs21_ail.seq
849   T C T T G C T G T C G C T T C T C A T T T G G T T T A A A A A G A C G A A A C A G A A A A A G C G G coh1_ail.seq
849   T C T T G C T G T C G C T T C T C A T T T G G T T T A A A A A G A C G A A A C A G A A A A A G C G G cjb111_ail.seq
851   T C T T G C T G T C G C T T C T C A T T T G G T T T A A A A A G A C G A A A C A G A A A A A G C G G nem316_ail.seq
850   T C T T G C T G T C G C T T C T C A T T T G G T T T A A A A A G A C G A A A C A G A A A A A G C G G a909_ail.seq

      A G A A A G A A T G A A A A A A G C G G C T A G T C A A A A T A G T C A C A A T A A T T C G A A A T A Majority
                5910          5920          5930          5940          5950
898   A G A A A G A A T G A A A A A G C G G C T A G T C A A A A T A G T C A C A A T A A T T C G A A A T A 2603_ail.seq
867   A G A A A G A A T G A A A A A G C G G C T A G T C A A A A T A G T C A C A A T A A T T C G A A A T A 18rs21_ail.seq
899   A G A A A G A A T G A A A A A G C G G C T A G T C A A A A T A G T C A C A A T A A T T C G A A A T A coh1_ail.seq
899   A G A A A G A A T G A A A A A G C G G C T A G T C A A A A T A G T C A C A A T A A T T C G A A A T A cjb111_ail.seq
901   A G A A A G A A T G A A A A A G C G G C T A G T C A A A A T A G T C A C A A T A A T T C G A A A T A nem316_ail.seq
900   A G A A A G A A T G A A A A A G C G G C T A G T C A A A A T A G T C A C A A T A A T T C G A A A T A a909_ail.seq
```

FIGURE 18 P

```
                ATAAAATCAGAACCCTCATTTTTGTGATGGGAAGTCTGATTCTCTTATTT Majority
                    5960      5970      5980      5990      6000
5948 ATAAAATCAGAACCCTCATTTTTGTGATGGGAAGTCTGATTCTCTTATTT 2603_ail.seq
5717 ATAAAATCAGAACCCTCATTTTTGTGATGGGAAGTCTGATTCTCTTATTT 18rs21_ail.seq
5949 ATAAAATCAGAACCCTCATTTTTGTGATGGGAAGTCTGATTCTCTTATTT cohl_ail.seq
5949 ATAAAATCAGAACCCTCATTTTTGTGATGCGAAGTCTGATTCTCTTATTT cjb111_ail.seq
5951 ATAAAATCAGAACCCTCATTTTTGTGATGGGAAGTCTGATTCTCTTATTT nem316_ail.seq
5950 ATAAAATCAGAACCCTCATTTTTGTGATGGGAAGTCTGATTCTCTTATTT a909_ail.seq

                CCGATTGTGAGCCAGGTAAGTTACTACCTTGCTTCGCATCAAAATATTAA Majority
                    6010      6020      6030      6040      6050
5998 CCGATTGTGAGCCAGGTAAGTTACTACCTTGCTTCGCATCAAAATATTAA 2603_ail.seq
5767 CCGATTGTGAGCCAGGTAAGTTACTACCTTGCTTCGCATCAAAATATTAA 18rs21_ail.seq
5999 CCGATTGTGAGCCAGGTAAGTTACTACCTTGCTTCGCATCAAAATATTAA cohl_ail.seq
5999 CCGATTGTGAGCCAGGTAAGTTACTACCTTGCTTCGCATCAAAATATTAA cjb111_ail.seq
6001 CCGATTGTGAGCCAGGTAAGTTACTACCTTGCTTCGCATCAAAATATTAA nem316_ail.seq
6000 CCGATTGTGAGCCAGGTAAGTTACTACCTTGCTTCGCATCAAAATATTAA a909_ail.seq

                TCAATTTAAGCGGGAAGTCGCTAAGATTGATACTAATACGGTTGAACGAC Majority
                    6060      6070      6080      6090      6100
6048 TCAATTTAAGCGGGAAGTCGCTAAGATTGATACTAATACGGTTGAACGAC 2603_ail.seq
5817 TCAATTTAAGCGGGAAGTCGCTAAGATTGATACTAATACGGTTGAACGAC 18rs21_ail.seq
6049 TCAATTTAAGCGGGAAGTCGCTAAGATTGATACTAATACGGTTGAACGAC cohl_ail.seq
6049 TCAATTTAAGCGGGAACTCGCTAAGATTGATACTAATACGGTTGAACGAC cjb111_ail.seq
6051 TCAATTTAAGCGGGAAGTCGCTAAGATTGATACTAATACGGTTGAACGAC nem316_ail.seq
6050 TCAATTTAAGCGGGAAGTCGCTAAGATTGATACTAATACGGTTGAACGAC a909_ail.seq

                GCATCGCTTTAGCTAATGCTTACAATGAGACGTTATCAAGGAATCCCTTG Majority
                    6110      6120      6130      6140      6150
6098 GCATCGCTTTAGCTAATGCTTACAATGAGACGTTATCAAGGAATCCCTTG 2603_ail.seq
5867 GCATCGCTTTAGCTAATGCTTACAATGAGACGTTATCAAGGAATCCCTTG 18rs21_ail.seq
6099 GCATCGCTTTAGCTAATGCTTACAATGAGACGTTATCAAGGAATCCCTTG cohl_ail.seq
6099 GCATCGCTTTAGCTAATGCTTACAATGAGACGTTATCAAGGAATCCCTTG cjb111_ail.seq
6101 GCATCGCTTTAGCTAATGCTTACAATGAGACGTTATCAAGGAATCCCTTG nem316_ail.seq
6100 GCATCGCTTTAGCTAATGCTTACAATGAGACGTTATCAAGGAATCCCTTG a909_ail.seq

                CTTATAGACCCTTTTACCAGTAAGCAAAAAGAAGGTTTGAGAGAGTATGC Majority
                    6160      6170      6180      6190      6200
6148 CTTATAGACCCTTTTACCAGTAAGCAAAAAGAAGGTTTGAGAGAGTATGC 2603_ail.seq
5917 CTTATAGACCCTTTTACCAGTAAGCAAAAAGAAGGTTTGAGAGAGTATGC 18rs21_ail.seq
6149 CTTATAGACCCTTTTACCAGTAAGCAAAAAGAAGGTTTGAGAGAGTATGC cohl_ail.seq
6149 CTTATAGACCCTTTTACCAGTAAGCAAAAAGAAGGTTTGAGAGAGTATGC cjb111_ail.seq
6151 CTTATAGACCCTTTTACCAGTAAGCAAAAAGAAGGTTTGAGAGAGTATGC nem316_ail.seq
6150 CTTATAGACCCTTTTACCAGTAAGCAAAAAGAAGGTTTGAGAGAGTATGC a909_ail.seq

                TCGTATGCTTGAAGTTCATGAGCAAATAGGTCATGTGGCAATCCCAAGTA Majority
                    6210      6220      6230      6240      6250
6198 TCGTATGCTTGAAGTTCATGAGCAAATAGGTCATGTGGCAATCCCAAGTA 2603_ail.seq
5967 TCGTATGCTTGAAGTTCATGAGCAAATAGGTCATGTGGCAATCCCAAGTA 18rs21_ail.seq
6199 TCGTATGCTTGAAGTTCATGAGCAAATAGGTCATGTGGCAATCCCAAGTA cohl_ail.seq
6199 TCGTATGCTTGAAGTTCATGAGCAAATAGGTCATGTGGCAATCCCAAGTA cjb111_ail.seq
6201 TCGTATGCTTGAAGTTCATGAGCAAATAGGTCATGTGGCAATCCCAAGTA nem316_ail.seq
6200 TCGTATGCTTGAAGTTCATGAGCAAATAGGTCATGTGGCAATCCCAAGTA a909_ail.seq

                TTGGGGTTGATATTCCAATTTATGCTGGAACATCCGAAACTGTGCTTCAG Majority
                    6260      6270      6280      6290      6300
6248 TTGGGGTTGATATTCCAATTTATGCTGGAACATCCGAAACTGTGCTTCAG 2603_ail.seq
6017 TTGGGGTTGATATTCCAATTTATGCTGGAACATCCGAAACTGTGCTTCAG 18rs21_ail.seq
6249 TTGGGGTTGATATTCCAATTTATGCTGGAACATCCGAAACTGTGCTTCAG cohl_ail.seq
6249 TTGGGGTTGATATTCCAATTTATGCTGGAACATCCGAAACTGTGCTTCAG cjb111_ail.seq
6251 TTGGGGTTGATATTCCAATTTATGCTGGAACATCCGAAACTGTGCTTCAG nem316_ail.seq
6250 TTGGGGTTGATATTCCAATTTATGCTGGAACATCCGAAACTGTGCTTCAG a909_ail.seq
```

FIGURE 18 Q

```
                AAAGGTAGTGGGCATTTGGAGGGAACCAGTCTTCCAGTGGGAGGTTTGTC Majority
                      6310      6320      6330      6340      6350
6298 AAAGGTAGTGGGCATTTGGAGGGAACCAGTCTTCCAGTGGGAGGTTTGTC 2603_ail.seq
6067 AAAGGTAGTGGGCATTTGGAGGGAACCAGTCTTCCAGTGGGAGGTTTGTC 2603_ail.seq
6299 AAAGGTAGTGGGCATTTGGAGGGAACCAGTCTTCCAGTGGGAGGTTTGTC 18rs21_ail.seq
6299 AAAGGTAGTGGGCATTTGGAGGGAACCAGTCTTCCAGTGGGAGGTTTGTC cohl_ail.seq
6301 AAAGGTAGTGGGCATTTGGAGGGAACCAGTCTTCCAGTGGGAGGTTTGTC cjbl11_ail.seq
6300 AAAGGTAGTGGGCATTTGGAGGGAACCAGTCTTCCAGTGGGAGGTTTGTC nem316_ail.seq
6298 AAAGGTAGTGGGCATTTGGAGGGAACCAGTCTTCCAGTGGCAGGTTTGTC a909_ail.seq

                AACCCATTCAGTACTAACTGCCCACCGTGGCTTGCCAACAGCTAGGCTAT Majority
                      6360      6370      6380      6390      6400
6348 AACCCATTCAGTACTAACTGCCCACCGTGGCTTGCCAACAGCTAGGCTAT 2603_ail.seq
6117 AACCCATTCAGTACTAACTGCCCACCGTGGCTTGCCAACAGCTAGGCTAT 2603_ail.seq
6349 AACCCATTCAGTACTAACTGCCCACCGTGGCTTGCCAACAGCTAGGCTAT 18rs21_ail.seq
6349 AACCCATTCAGTACTAACTGCCCACCGTGGCTTGCCAACAGCTAGGCTAT cohl_ail.seq
6351 AACCCATTCAGTACTAACTGCCCACCGTGGCTTGCCAACAGCTAGGCTAT cjbl11_ail.seq
6350 AACCCATTCAGTACTAACTGCCCACCGTGGCTTGCCAACAGCTAGGCTAT nem316_ail.seq
6348 AACCCATTCAGTACTAACTGCCCACCGTGGCTTGCCAACAGCTAGGCTAT a909_ail.seq

                TTACCGACTTAAATAAAGTTAAAAAAAGGCCAGATTTTCTATGTGACGAAC Majority
                      6410      6420      6430      6440      6450
6398 TTACCGACTTAAATAAAGTTAAAAAAAGGCCAGATTTTCTATGTGACGAAC 2603_ail.seq
6167 TTACCGACTTAAATAAAGTTAAAAAAAGGCCAGATTTTCTATGTGACGAAC 2603_ail.seq
6399 TTACCGACTTAAATAAAGTTAAAAAAAGGCCAGATTTTCTATGTGACGAAC 18rs21_ail.seq
6399 TTACCGACTTAAATAAAGTTAAAAAAAGGCCAGATTTTCTATGTGACGAAC cohl_ail.seq
6401 TTACCGACTTAAATAAAGTTAAAAAAAGGCCAGATTTTCTATGTGACGAAC cjbl11_ail.seq
6400 TTACCGACTTAAATAAAGTTAAAAAAAGGCCAGATTTTCTATGTGACGAAC nem316_ail.seq
6398 TTACCGACTTAAATAAAGTTAAAAAAAGGCCAGATTTTCTATGTGACGAAC a909_ail.seq

                ATCAAGGAAACACTTGCCTACAAAGTCGTGTCTATCAAAGTTGTGGATCC Majority
                      6460      6470      6480      6490      6500
6448 ATCAAGGAAACACTTGCCTACAAAGTCGTGTCTATCAAAGTTGTGGATCC 2603_ail.seq
6217 ATCAAGGAAACACTTGCCTACAAAGTCGTGTCTATCAAAGTTGTGGATCC 2603_ail.seq
6449 ATCAAGGAAACACTTGCCTACAAAGTCGTGTCTATCAAAGTTGTGGATCC 18rs21_ail.seq
6449 ATCAAGGAAACACTTGCCTACAAAGTCGTGTCTATCAAAGTTGTGGATCC cohl_ail.seq
6451 ATCAAAGAAACACTTGCCTACAAAGTCGTGTCTATCAAAGTTGTGGATCC cjbl11_ail.seq
6450 ATCAAGGAAACACTTGCCTACAAAGTCGTGTCTATCAAAGTTGTGGATCC nem316_ail.seq
6448 ATCAAGGAAACACTTGCCTACAAAGTCGTGTCTATCAAAGTTGTGGATCC a909_ail.seq

                AACAGCTTTAAGTGAGGTTAAGATTGTCAATGGTAAGGATTATATAACCT Majority
                      6510      6520      6530      6540      6550
6498 AACAGCTTTAAGTGAGGTTAAGATTGTCAATGGTAAGGATTATATAACCT 2603_ail.seq
6267 AACAGCTTTAAGTGAGGTTAAGATTGTCAATGGTAAGGATTATATAACCT 2603_ail.seq
6499 AACAGCTTTAAGTGAGGTTAAGATTGTCAATGGTAAGGATTATATAACCT 18rs21_ail.seq
6499 AACAGCTTTAAGTGAGGTTAAGATTGTCAATGGTAAGGATTATATAACCT cohl_ail.seq
6501 AACAGCTTTAAGTGAGGTTAAGATTGTCAATGGTAAGGATTATATAACCT cjbl11_ail.seq
6500 AACAGCTTTAAGTGAGGTTAAGATTGTCAATGGTAAGGATTATATAACCT nem316_ail.seq
6498 AACAGCTTTAAGTGAGGTTAAGATTGTCAATGGTAAGGATTATATAACCT a909_ail.seq

                TGCTGACTTGCACACCTTACATGATCAATAGTCATCGTCTCTTGGTAAAA Majority
                      6560      6570      6580      6590      6600
6548 TGCTGACTTGCACACCTTACATGATCAATAGTCATCGTCTCTTGGTAAAA 2603_ail.seq
6317 TGCTGACTTGCACACCTTACATGATCAATAGTCATCGTCTCTTGGTAAAA 2603_ail.seq
6549 TGCTGACTTGCACACCTTACATGATCAATAGTCATCGTCTCTTGGTAAAA 18rs21_ail.seq
6549 TGCTGACTTGCACACCTTACATGATCAATAGTCATCGTCTCTTGGTAAAA cohl_ail.seq
6551 TGCTGACTTGCACACCTTACATGATCAATAGTCATCGTCTCTTGGTAAAA cjbl11_ail.seq
6550 TGCTGACTTGCACACCTTACATGATCAATAGTCATCGTCTCTTGGTAAAA nem316_ail.seq
6548 TGCTGACTTGCACACCTTACATGATCAATAGTCATCGTCTCTTGGTAAAA a909_ail.seq

                GGAGAGCGTATTCCTTATGATTCTACCGAGGCGGAAAAGCACAAAGAACA Majority
                      6610      6620      6630      6640      6650
6598 GGAGAGCGTATTCCTTATGATTCTACCGAGGCGGAAAAGCACAAAGAACA 2603_ail.seq
6367 GGAGAGCGTATTCCTTATGATTCTACCGAGGCGGAAAAGCACAAAGAACA 2603_ail.seq
6599 GGAGAGCGTATTCCTTATGATTCTACCGAGGCGGAAAAGCACAAAGAACA 18rs21_ail.seq
6599 GGAGAGCGTATTCCTTATGATTCTACCGAGGCGGAAAAGCACAAAGAACA cohl_ail.seq
6601 GGAGAGCGTATTCCTTATGATTCTACCGAGGCGGAAAAGCACAAAGAACA cjbl11_ail.seq
6600 GGAGAGCGTATTCCTTATGATTCTACCGAGGCGGAAAAGCACAAAGAACA nem316_ail.seq
6598 GGAGAGCGTATTCCTTATGATTCTACCGAGGCGGAAAAGCACAAAGAACA a909_ail.seq
```

FIGURE 18 R

```
     AACCGTACAAGATTATCGTTTGTCACTAGTGTTGAAGATACTACTAGTAT Majority
          6660      6670      6680      6690      6700
6648 AACCGTACAAGATTATCGTTTGTCACTAGTGTTGAAGATACTACTAGTAT 2603_all.seq
6417 AACCGTACAAGATTATCGTTTGTCACTAGTGTTGAAGATACTACTAGTAT 18rs21_all.seq
6649 AACCGTACAAGATTATCGTTTGTCACTAGTGTTGAAGATACTACTAGTAT coh1_all.seq
6649 AACCGTACAAGATTATCGTTTGTCACTAGTGTTGAAGATACTACTAGTAT cjb111_all.seq
6651 AACCGTACAAGATTATCGTTTGTCACTAGTGTTGAAGATACTACTAGTAT nem316_all.seq
6650 AACCGTACAAGATTATCGTTTGTCACTAGTGTTGAAGATACTACTAGTAT a909_all.seq

     TATTAATTGGACTCTTCATCGTGATAATGATGAGAAGATGGATGCAACAT Majority
          6710      6720      6730      6740      6750
6698 TATTAATTGGACTCTTCATCGTGATAATGATGAGAAGATGGATGCAACAT 2603_all.seq
6467 TATTAATTGGACTCTTCATCGTGATAATGATGAGAAGATGGATGCAACAT 18rs21_all.seq
6699 TATTAATTGGACTCTTCATCGTGATAATGATGAGAAGATGGATGCAACAT coh1_all.seq
6699 TATTAATTGGACTCTTCATCGTGATAATGATGAGAAGATGGATGCAACAT cjb111_all.seq
6701 TATTAATTGGACTCTTCATCGTGATAATGATGAGAAGATGGATGCAACAT nem316_all.seq
6700 TATTAATTGGACTCTTCATCGTGATAATGATGAGAAGATGGATGCAACAT a909_all.seq

     CGTCAATAACGATGTTGTGAATGGCTTACTTACTTATCAAATAGGTGACT Majority
          6760      6770      6780      6790      6800
6748 CGTCAATAACGATGTTGTGAATGGCTTACTT....ATCAAATAGGTGACT 2603_all.seq
6517 CGTCAATAACGATGTTGTGAATGGCTTACTTACTTATCAAATAGGTGACT 18rs21_all.seq
6749 CGTCAATAACGATGTTGTGAATGGCTTACTTACTTATCAAATAGGTGACT coh1_all.seq
6749 CGTCAATAACGATGTTGTGAATGGCTTACTTACTTATCAAATAGGTGACT cjb111_all.seq
6751 CGTCAATAACGATGTTGTGAATGGCTTACTTACTTATCAAATAGGTGACT nem316_all.seq
6750 CGTCAATAACGATGTTGTGAATGGCTTACTTACTTATCAAATAGGTGACT a909_all.seq

     AATGATGATTGTGAATAATGGTTATCTAGAAGGGAGAAAAATGAAAAAGA Majority
          6810      6820      6830      6840      6850
6794 AATGATGATTGTGAATAATGGTTATCTAGAAGGGAGAAAAATGAAAAAGA 2603_all.seq
6567 AATGATGATTGTGAATAATGGTTATCTAGAAGGGAGAAAAATGAAAAAGA 18rs21_all.seq
6799 AATGATGATTGTGAATAATGGTTATCTAGAAGGGAGAAAAATGAAAAAGA coh1_all.seq
6799 AATGATGATTGTGAATAATGGTTATCTAGAAGGGAGAAAAATGAAAAAGA cjb111_all.seq
6801 AATGATGATTGTGAATAATGGTTATCTAGAAGGGAGAAAAATGAAAAAGA nem316_all.seq
6800 AATGATGATTGTGAATAATGGTTATCTAGAAGGGAGAAAAATGAAAAAGA a909_all.seq

     GACAAAAAATATGGAGAGGGTTATCAGTTACTTTACTAATCCTGTCCCAA Majority
          6860      6870      6880      6890      6900
6844 GACAAAAAATATGGAGAGGGTTATCAGTTACTTTACTAATCCTGTCCCAA 2603_all.seq
6617 GACAAAAAATATGGAGAGGGTTATCAGTTACTTTACTAATCCTGTCCCAA 18rs21_all.seq
6849 GACAAAAAATATGGAGAGGGTTATCAGTTACTTTACTAATCCTGTCCCAA coh1_all.seq
6849 GACAAAAAATATGGAGAGGGTTATCAGTTACTTTACTAATCCTGTCCCAA cjb111_all.seq
6851 GACAAAAAATATGGAGAGGGTTATCAGTTACTTTACTAATCCTGTCCCAA nem316_all.seq
6850 GACAAAAAATATGGAGAGGGTTATCAGTTACTTTACTAATCCTGTCCCAA a909_all.seq

     ATTCCATTTGGTATATTGGTACAAGGTGAAACCCAAGATACCAATCAAGC Majority
          6910      6920      6930      6940      6950
6894 ATTCCATTTGGTATATTGGTACAAGGTGAAACCCAAGATACCAATCAAGC 2603_all.seq
6667 ATTCCATTTGGTATATTGGTACAAGGTGAAACCCAAGATACCAATCAAGC 18rs21_all.seq
6899 ATTCCATTTGGTATATTGGTACAAGGTGAAACCCAAGATACCAATCAAGC coh1_all.seq
6899 ATTCCATTTGGTATATTGGTACAAGGTGAAACCCAAGATACCAATCAAGC cjb111_all.seq
6901 ATTCCATTTGGTATATTGGTACAAGGTGAAACCCAAGATACCAATCAAGC nem316_all.seq
6900 ATTCCATTTGGTATATTGGTACAAGGTGAAACCCAAGATACCAATCAAGC a909_all.seq

     ACTTGGAAAAGTAATTGTTAAAAAAAACGGGAGACAATGCTACACCATTAG Majority
          6960      6970      6980      6990      7000
6944 ACTTGGAAAAGTAATTGTTAAAAAAAACGGGAGACAATGCTACACCATTAG 2603_all.seq
6717 ACTTGGAAAAGTAATTGTTAAAAAAAACGGGAGACAATGCTACACCATTAG 18rs21_all.seq
6949 ACTTGGAAAAGTAATTGTTAAAAAAAACGGGAGACAATGCTACACCATTAG coh1_all.seq
6949 ACTTGGAAAAGTAATTGTTAAAAAAAACGGGAGACAATGCTACACCATTAG cjb111_all.seq
6951 ACTTGGAAAAGTAATTGTTAAAAAAAACGGGAGACAATGCTACACCATTAG nem316_all.seq
6950 ACTTGGAAAAGTAATTGTTAAAAAAAACGGGGGACAATGCTACACCATTAG a909_all.seq
```

FIGURE 18 S

```
      GCAAAGCGACTTTTGTGTTAAAAAATGACAATGATAAGTCAGAAACAAGT Majority
                7010      7020      7030      7040      7050
6994  GCAAAGCGACTTTTGTGTTAAAAAATGACAATGATAAGTCAGAAACAAGT 2603_ail.seq
6767  GCAAAGCGACTTTTGTGTTAAAAAATGACAATGATAAGTCAGAAACAAGT 18rs21_ail.seq
6999  GCAAAGCGACTTTTGTGTTAAAAAATGACAATGATAAGTCAGAAACAAGT cohl_ail.seq
6999  GCAAAGCGACTTTTGTGTTAAAAAATGACAATGATAAGTCAGAAACAAGT cjb111_ail.seq
7001  GCAAAGCGACTTTTGTGTTAAAAAATGACAATGATAAGTCAGAAACAAGT nem316_ail.seq
7000  GCAAAGCGACTTTTGTGTTAAAAAATGACAATGATAAGTCAGAAACAAGT a909_ail.seq

      CACGAAACGGTAGAGGGTTCTGGAGAAGCAACCTTTGAAAACATAAAACC Majority
                7060      7070      7080      7090      7100
7044  CACGAAACGGTAGAGGGTTCTGGAGAAGCAACCTTTGAAAACATAAAACC 2603_ail.seq
6817  CACGAAACGGTAGAGGGTTCTGGAGAAGCAACCTTTGAAAACATAAAACC 18rs21_ail.seq
7049  CACGAAACGGTAGAGGGTTCTGGAGAAGCAACCTTTGAAAACATAAAACC cohl_ail.seq
7049  CACGAAACGGTAGAGGGTTCTGGAGAAGCAACCTTTGAAAACATAAAACC cjb111_ail.seq
7051  CACGAAACGGTAGAGGGTTCTGGAGAAGCAACCTTTGAAAACATAAAACC nem316_ail.seq
7050  CACGAAACGGTAGAGGGTTCTGGAGAAGCAACCTTTGAAAACATAAAACC a909_ail.seq

      TGGAGACTACACATTAAGAGAAGAAACAGCACCAATTGGTTATAAAAAAA Majority
                7110      7120      7130      7140      7150
7094  TGGAGACTACACATTAAGAGAAGAAACAGCACCAATTGGTTATAAAAAAA 2603_ail.seq
6867  TGGAGACTACACATTAAGAGAAGAAACAGCACCAATTGGTTATAAAAAAA 18rs21_ail.seq
7099  TGGAGACTACACATTAAGAGAAGAAACAGCACCAATTGGTTATAAAAAAA cohl_ail.seq
7099  TGGAGACTACACATTAAGAGAAGAAACAGCACCAATTGGTTATAAAAAAA cjb111_ail.seq
7101  TGGAGACTACACATTAAGAGAAGAAACAGCACCAATTGGTTATAAAAAAA nem316_ail.seq
7100  TGGAGACTACACATTAAGAGAAGAAACAGCACCAATTGGTTATAAAAAAA a909_ail.seq

      CTGATAAAACCTGGAAAGTTAAAGTTGCAGATAACGGAGCAACAATAATC Majority
                7160      7170      7180      7190      7200
7144  CTGATAAAACCTGGAAAGTTAAAGTTGCAGATAACGGAGCAACAATAATC 2603_ail.seq
6917  CTGATAAAACCTGGAAAGTTAAAGTTGCAGATAACGGAGCAACAATAATC 18rs21_ail.seq
7149  CTGATAAAACCTGGAAAGTTAAAGTTGCAGATAACGGAGCAACAATAATC cohl_ail.seq
7149  CTGATAAAACCTGGAAAGTTAAAGTTGCAGATAACGGAGCAACAATAATC cjb111_ail.seq
7151  CTGATAAAACCTGGAAAGTTAAAGTTGCAGATAACGGAGCAACAATAATC nem316_ail.seq
7150  CTGATAAAACCTGGAAAGTTAAAGTTGCAGATAACGGAGCAACAATAATC a909_ail.seq

      GAGGGTATGGATGCAGATAAAGCAGAGAAACGAAAAGAAGTTTTGAATGC Majority
                7210      7220      7230      7240      7250
7194  GAGGGTATGGATGCAGATAAAGCAGAGAAACGAAAAGAAGTTTTGAATGC 2603_ail.seq
6967  GAGGGTATGGATGCAGATAAAGCAGAGAAACGAAAAGAAGTTTTGAATGC 18rs21_ail.seq
7199  GAGGGTATGGATGCAGATAAAGCAGAGAAACGAAAAGAAGTTTTGAATGC cohl_ail.seq
7199  GAGGGTATGGATGCAGATAAAGCAGAGAAACGAAAAGAAGTTTTGAATGC cjb111_ail.seq
7201  GAGGGTATGGATGCAGATAAAGCAGAGAAACGAAAAGAAGTTTTGAATGC nem316_ail.seq
7200  GAGGGTATGGATGCAGATAAAGCAGAGAAACGAAAAGAAGTTTTGAATGC a909_ail.seq

      CCAATATCCAAAATCAGCTATTTATGAGGATACAAAAGAAAATTACCCAT Majority
                7260      7270      7280      7290      7300
7244  CCAATATCCAAAATCAGCTATTTATGAGGATACAAAAGAAAATTACCCAT 2603_ail.seq
7017  CCAATATCCAAAATCAGCTATTTATGAGGATACAAAAGAAAATTACCCAT 18rs21_ail.seq
7249  CCAATATCCAAAATCAGCTATTTATGAGGATACAAAAGAAAATTACCCAT cohl_ail.seq
7249  CCAATATCCAAAATCAGCTATTTATGAGGATACAAAAGAAAATTACCCAT cjb111_ail.seq
7251  CCAATATCCAAAATCAGCTATTTATGAGGATACAAAAGAAAATTACCCAT nem316_ail.seq
7250  CCAATATCCAAAATCAGCTATTTATGAGGATACAAAAGAAAATTACCCAT a909_ail.seq

      TAGTTAATGTAGAGGGTTCCAAAGTTGGTGAACAATACAAAGCATTGAAT Majority
                7310      7320      7330      7340      7350
7294  TAGTTAATGTAGAGGGTTCCAAAGTTGGTGAACAATACAAAGCATTGAAT 2603_ail.seq
7067  TAGTTAATGTAGAGGGTTCCAAAGTTGGTGAACAATACAAAGCATTGAAT 18rs21_ail.seq
7299  TAGTTAATGTAGAGGGTTCCAAAGTTGGTGAACAATACAAAGCATTGAAT cohl_ail.seq
7299  TAGTTAATGTAGAGGGTTCCAAAGTTGGTGAACAATACAAAGCATTGAAT cjb111_ail.seq
7301  TAGTTAATGTAGAGGGTTCCAAAGTTGGTGAACAATACAAAGCATTGAAT nem316_ail.seq
7300  TAGTTAATGTAGAGGGTTCCAAAGTTGGTGAACAATACAAAGCATTGAAT a909_ail.seq
```

FIGURE 18 T

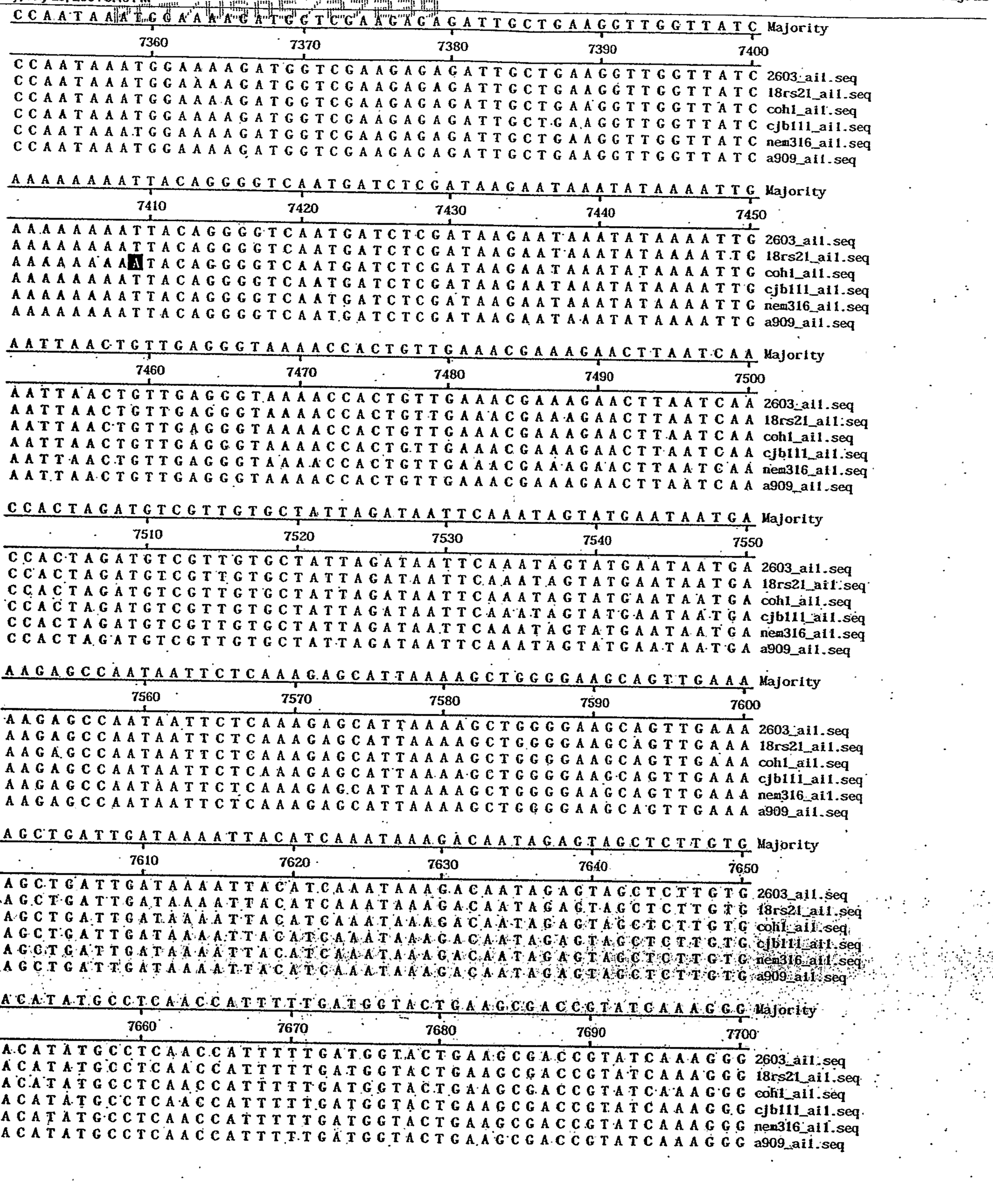

CCAATAAATGGAAAAGATGGTCGAAGAGAGATTGCTGAAGGTTGGTTATC Majority
7360 7370 7380 7390 7400
7344 CCAATAAATGGAAAAGATGGTCGAAGAGAGATTGCTGAAGGTTGGTTATC 2603_ail.seq
7117 CCAATAAATGGAAAAGATGGTCGAAGAGAGATTGCTGAAGGTTGGTTATC 2603_ail.seq
7349 CCAATAAATGGAAAAGATGGTCGAAGAGAGATTGCTGAAGGTTGGTTATC 18rs21_ail.seq
7349 CCAATAAATGGAAAAGATGGTCGAAGAGAGATTGCTGAAGGTTGGTTATC coh1_ail.seq
7351 CCAATAAATGGAAAAGATGGTCGAAGAGAGATTGCTGAAGGTTGGTTATC cjb111_ail.seq
7350 CCAATAAATGGAAAAGATGGTCGAAGAGAGATTGCTGAAGGTTGGTTATC nem316_ail.seq
7350 CCAATAAATGGAAAAGATGGTCGAAGAGAGATTGCTGAAGGTTGGTTATC a909_ail.seq
AAAAAAAAATTACAGGGGTCAATGATCTCGATAAGAATAAATATAAAATTG Majority
7410 7420 7430 7440 7450
7394 AAAAAAAAATTACAGGGGTCAATGATCTCGATAAGAATAAATATAAAATTG 2603_ail.seq
7167 AAAAAAAAATTACAGGGGTCAATGATCTCGATAAGAATAAATATAAAATTG 2603_ail.seq
7399 AAAAAAAAATTACAGGGGTCAATGATCTCGATAAGAATAAATATAAAATTG 18rs21_ail.seq
7399 AAAAAAAAATTACAGGGGTCAATGATCTCGATAAGAATAAATATAAAATTG coh1_ail.seq
7401 AAAAAAAAATTACAGGGGTCAATGATCTCGATAAGAATAAATATAAAATTG cjb111_ail.seq
7400 AAAAAAAAATTACAGGGGTCAATGATCTCGATAAGAATAAATATAAAATTG nem316_ail.seq
7400 AAAAAAAAATTACAGGGGTCAATGATCTCGATAAGAATAAATATAAAATTG a909_ail.seq
AATTAACTGTTGAGGGTAAAACCACTGTTGAAACGAAAGAACTTAATCAA Majority
7460 7470 7480 7490 7500
7444 AATTAACTGTTGAGGGTAAAACCACTGTTGAAACGAAAGAACTTAATCAA 2603_ail.seq
7217 AATTAACTGTTGAGGGTAAAACCACTGTTGAAACGAAAGAACTTAATCAA 2603_ail.seq
7449 AATTAACTGTTGAGGGTAAAACCACTGTTGAAACGAAAGAACTTAATCAA 18rs21_ail.seq
7449 AATTAACTGTTGAGGGTAAAACCACTGTTGAAACGAAAGAACTTAATCAA coh1_ail.seq
7451 AATTAACTGTTGAGGGTAAAACCACTGTTGAAACGAAAGAACTTAATCAA cjb111_ail.seq
7450 AATTAACTGTTGAGGGTAAAACCACTGTTGAAACGAAAGAACTTAATCAA nem316_ail.seq
7450 AATTAACTGTTGAGGGTAAAACCACTGTTGAAACGAAAGAACTTAATCAA a909_ail.seq
CCACTAGATGTCGTTGTGCTATTAGATAATTCAAATAGTATGAATAATGA Majority
7510 7520 7530 7540 7550
7494 CCACTAGATGTCGTTGTGCTATTAGATAATTCAAATAGTATGAATAATGA 2603_ail.seq
7267 CCACTAGATGTCGTTGTGCTATTAGATAATTCAAATAGTATGAATAATGA 2603_ail.seq
7499 CCACTAGATGTCGTTGTGCTATTAGATAATTCAAATAGTATGAATAATGA 18rs21_ail.seq
7499 CCACTAGATGTCGTTGTGCTATTAGATAATTCAAATAGTATGAATAATGA coh1_ail.seq
7501 CCACTAGATGTCGTTGTGCTATTAGATAATTCAAATAGTATGAATAATGA cjb111_ail.seq
7500 CCACTAGATGTCGTTGTGCTATTAGATAATTCAAATAGTATGAATAATGA nem316_ail.seq
7500 CCACTAGATGTCGTTGTGCTATTAGATAATTCAAATAGTATGAATAATGA a909_ail.seq
AAGAGCCAATAATTCTCAAAGAGCATTAAAAGCTGGGGAAGCAGTTGAAA Majority
7560 7570 7580 7590 7600
7544 AAGAGCCAATAATTCTCAAAGAGCATTAAAAGCTGGGGAAGCAGTTGAAA 2603_ail.seq
7317 AAGAGCCAATAATTCTCAAAGAGCATTAAAAGCTGGGGAAGCAGTTGAAA 2603_ail.seq
7549 AAGAGCCAATAATTCTCAAAGAGCATTAAAAGCTGGGGAAGCAGTTGAAA 18rs21_ail.seq
7549 AAGAGCCAATAATTCTCAAAGAGCATTAAAAGCTGGGGAAGCAGTTGAAA coh1_ail.seq
7551 AAGAGCCAATAATTCTCAAAGAGCATTAAAAGCTGGGGAAGCAGTTGAAA cjb111_ail.seq
7550 AAGAGCCAATAATTCTCAAAGAGCATTAAAAGCTGGGGAAGCAGTTGAAA nem316_ail.seq
7550 AAGAGCCAATAATTCTCAAAGAGCATTAAAAGCTGGGGAAGCAGTTGAAA a909_ail.seq
AGCTGATTGATAAAATTACATCAAATAAAGACAATAGAGTAGCTCTTGTG Majority
7610 7620 7630 7640 7650
7594 AGCTGATTGATAAAATTACATCAAATAAAGACAATAGAGTAGCTCTTGTG 2603_ail.seq
7367 AGCTGATTGATAAAATTACATCAAATAAAGACAATAGAGTAGCTCTTGTG 2603_ail.seq
7599 AGCTGATTGATAAAATTACATCAAATAAAGACAATAGAGTAGCTCTTGTG 18rs21_ail.seq
7599 AGCTGATTGATAAAATTACATCAAATAAAGACAATAGAGTAGCTCTTGTG coh1_ail.seq
7601 AGCTGATTGATAAAATTACATCAAATAAAGACAATAGAGTAGCTCTTGTG cjb111_ail.seq
7600 AGCTGATTGATAAAATTACATCAAATAAAGACAATAGAGTAGCTCTTGTG nem316_ail.seq
7600 AGCTGATTGATAAAATTACATCAAATAAAGACAATAGAGTAGCTCTTGTG a909_ail.seq
ACATATGCCTCAACCATTTTTGATGGTACTGAAGCGACCGTATCAAAGGG Majority
7660 7670 7680 7690 7700
7644 ACATATGCCTCAACCATTTTTGATGGTACTGAAGCGACCGTATCAAAGGG 2603_ail.seq
7417 ACATATGCCTCAACCATTTTTGATGGTACTGAAGCGACCGTATCAAAGGG 2603_ail.seq
7649 ACATATGCCTCAACCATTTTTGATGGTACTGAAGCGACCGTATCAAAGGG 18rs21_ail.seq
7649 ACATATGCCTCAACCATTTTTGATGGTACTGAAGCGACCGTATCAAAGGG coh1_ail.seq
7651 ACATATGCCTCAACCATTTTTGATGGTACTGAAGCGACCGTATCAAAGGG cjb111_ail.seq
7650 ACATATGCCTCAACCATTTTTGATGGTACTGAAGCGACCGTATCAAAGGG nem316_ail.seq
7650 ACATATGCCTCAACCATTTTTGATGGTACTGAAGCGACCGTATCAAAGGG a909_ail.seq

FIGURE 18 U

```
AGTTGCCGATCAAAATGGTAAAGCCGCTGAATGATAGTGTATCATGGGATT Majority
        7710      7720      7730      7740      7750
7694 AGTTGCCGATCAAAATGGTAAAGCCGCTGAATGATAGTGTATCATGGGATT 2603_all.seq
7467 AGTTGCCGATCAAAATGGTAAAGCCGCTGAATGATAGTGTATCATGGGATT 18rs21_all.seq
7699 AGTTGCCGATCAAAATGGTAAAGCCGCTGAATGATAGTGTATCATGGGATT coh1_all.seq
7699 AGTTGCCGATCAAAATGGTAAAGCCGCTGAATGATAGTGTATCATGGGATT cjb111_all.seq
7701 AGTTGCCGATCAAAATGGTAAAGCCGCTGAATGATAGTGTATCATGGGATT nem316_all.seq
7700 AGTTGCCGATCAAAATGGTAAAGCCGCTGAATGATAGTGTATCATGGGATT a909_all.seq

ATCATAAAACTACTTTTACAGCAACTACACATAATTACAGTTATTTAAAT Majority
        7760      7770      7780      7790      7800
7744 ATCATAAAACTACTTTTACAGCAACTACACATAATTACAGTTATTTAAAT 2603_all.seq
7517 ATCATAAAACTACTTTTACAGCAACTACACATAATTACAGTTATTTAAAT 18rs21_all.seq
7749 ATCATAAAACTACTTTTACAGCAACTACACATAATTACAGTTATTTAAAT coh1_all.seq
7749 ATCATAAAACTACTTTTACAGCAACTACACATAATTACAGTTATTTAAAT cjb111_all.seq
7751 ATCATAAAACTACTTTTACAGCAACTACACATAATTACAGTTATTTAAAT nem316_all.seq
7750 ATCATAAAACTACTTTTACAGCAACTACACATAATTACAGTTATTTAAAT a909_all.seq

TTAACAAATGATGCTAACGAAGTTAATATTCTAAAGTCAAGAATTCCAAA Majority
        7810      7820      7830      7840      7850
7794 TTAACAAATGATGCTAACGAAGTTAATATTCTAAAGTCAAGAATTCCAAA 2603_all.seq
7567 TTAACAAATGATGCTAACGAAGTTAATATTCTAAAGTCAAGAATTCCAAA 18rs21_all.seq
7799 TTAACAAATGATGCTAACGAAGTTAATATTCTAAAGTCAAGAATTCCAAA coh1_all.seq
7799 TTAACAAATGATGCTAACGAAGTTAATATTCTAAAGTCAAGAATTCCAAA cjb111_all.seq
7801 TTAACAAATGATGCTAACGAAGTTAATATTCTAAAGTCAAGAATTCCAAA nem316_all.seq
7800 TTAACAAATGATGCTAACGAAGTTAATATTCTAAAGTCAAGAATTCCAAA a909_all.seq

GGAAGCGGAGCATATAAATGGGGATCGCACGCTCTATCAATTTGGTGCGA Majority
        7860      7870      7880      7890      7900
7844 GGAAGCGGAGCATATAAATGGGGATCGCACGCTCTATCAATTTGGTGCGA 2603_all.seq
7617 GGAAGCGGAGCATATAAATGGGGATCGCACGCTCTATCAATTTGGTGCGA 18rs21_all.seq
7849 GGAAGCGGAGCATATAAATGGGGATCGCACGCTCTATCAATTTGGTGCGA coh1_all.seq
7849 GGAAGCGGAGCATATAAATGGGGATCGCACGCTCTATCAATTTGGTGCGA cjb111_all.seq
7851 GGAAGCGGAGCATATAAATGGGGATCGCACGCTCTATCAATTTGGTGCGA nem316_all.seq
7850 GGAAGCGGAGCATATAAATGGGGATCGCACGCTCTATCAATTTGGTGCGA a909_all.seq

CATTTACTCAAAAAGCTCTAATGAAAGCAAATGAAATTTTAGAGACACAA Majority
        7910      7920      7930      7940      7950
7894 CATTTACTCAAAAAGCTCTAATGAAAGCAAATGAAATTTTAGAGACACAA 2603_all.seq
7667 CATTTACTCAAAAAGCTCTAATGAAAGCAAATGAAATTTTAGAGACACAA 18rs21_all.seq
7899 CATTTACTCAAAAAGCTCTAATGAAAGCAAATGAAATTTTAGAGACACAA coh1_all.seq
7899 CATTTACTCAAAAAGCTCTAATGAAAGCAAATGAAATTTTAGAGACACAA cjb111_all.seq
7901 CATTTACTCAAAAAGCTCTAATGAAAGCAAATGAAATTTTAGAGACACAA nem316_all.seq
7900 CATTTACTCAAAAAGCTCTAATGAAAGCAAATGAAATTTTAGAGACACAA a909_all.seq

AGTTCTAATGCTAGAAAAAAACTTATTTTTCACGTAACTGATGGTGTCCC Majority
        7960      7970      7980      7990      8000
7944 AGTTCTAATGCTAGAAAAAAACTTATTTTTCACGTAACTGATGGTGTCCC 2603_all.seq
7717 AGTTCTAATGCTAGAAAAAAACTTATTTTTCACGTAACTGATGGTGTCCC 18rs21_all.seq
7949 AGTTCTAATGCTAGAAAAAAACTTATTTTTCACGTAACTGATGGTGTCCC coh1_all.seq
7949 AGTTCTAATGCTAGAAAAAAACTTATTTTTCACGTAACTGATGGTGTCCC cjb111_all.seq
7951 AGTTCTAATGCTAGAAAAAAACTTATTTTTCACGTAACTGATGGTGTCCC nem316_all.seq
7950 AGTTCTAATGCTAGAAAAAAACTTATTTTTCACGTAACTGATGGTGTCCC a909_all.seq

TACGATGTCTTATGCCATAAATTTTAATCCTTATATATCAACATCTTACC Majority
        8010      8020      8030      8040      8050
7994 TACGATGTCTTATGCCATAAATTTTAATCCTTATATATCAACATCTTACC 2603_all.seq
7767 TACGATGTCTTATGCCATAAATTTTAATCCTTATATATCAACATCTTACC 18rs21_all.seq
7999 TACGATGTCTTATGCCATAAATTTTAATCCTTATATATCAACATCTTACC coh1_all.seq
7999 TACGATGTCTTATGCCATAAATTTTAATCCTTATATATCAACATCTTACC cjb111_all.seq
8001 TACGATGTCTTATGCCATAAATTTTAATCCTTATATATCAACATCTTACC nem316_all.seq
8000 TACGATGTCTTATGCCATAAATTTTAATCCTTATATATCAACATCTTACC a909_all.seq
```

FIGURE 18 V

```
      AAAACCAGTTAATTCTTTTTTAAATAAAATACCAGATAGAAGTGGTATT Majority
                8060      8070      8080      8090      8100
8044  AAAACCAGTTTAATTCTTTTTTAAATAAAATACCAGATAGAAGTGGTATT 2603_all.seq
7817  AAAACCAGTTTAATTCTTTTTTAAATAAAATACCAGATAGAAGTGGTATT 18rs21_all.seq
8049  AAAACCAGTTTAATTCTTTTTTAAATAAAATACCAGATAGAAGTGGTATT coh1_all.seq
8049  AAAACCAGTTTAATTCTTTTTTAAATAAAATACCAGATAGAAGTGGTATT cjb111_all.seq
8051  AAAACCAGTTTAATTCTTTTTTAAATAAAATACCAGATAGAAGTGGTATT nem316_all.seq
8050  AAAACCAGTTTAATTCTTTTTTAAATAAAATACCAGATAGAAGTGGTATT a909_all.seq

      CTCCAAGAGGATTTTATAATCAATGGTGATGATTATCAAATAGTAAAAGG Majority
                8110      8120      8130      8140      8150
8094  CTCCAAGAGGATTTTATAATCAATGGTGATGATTATCAAATAGTAAAAGG 2603_all.seq
7867  CTCCAAGAGGATTTTATAATCAATGGTGATGATTATCAAATAGTAAAAGG 18rs21_all.seq
8099  CTCCAAGAGGATTTTATAATCAATGGTGATGATTATCAAATAGTAAAAGG coh1_all.seq
8099  CTCCAAGAGGATTTTATAATCAATGGTGATGATTATCAAATAGTAAAAGG cjb111_all.seq
8101  CTCCAAGAGGATTTTATAATCAATGGTGATGATTATCAAATAGTAAAAGG nem316_all.seq
8100  CTCCAAGAGGATTTTATAATCAATGGTGATGATTATCAAATAGTAAAAGG a909_all.seq

      AGATGGAGAGAGTTTTAAACTGTTTTCGGATAGAAAAGTTCCTGTTACTG Majority
                8160      8170      8180      8190      8200
8144  AGATGGAGAGAGTTTTAAACTGTTTTCGGATAGAAAAGTTCCTGTTACTG 2603_all.seq
7917  AGATGGAGAGAGTTTTAAACTGTTTTCGGATAGAAAAGTTCCTGTTACTG 18rs21_all.seq
8149  AGATGGAGAGAGTTTTAAACTGTTTTCGGATAGAAAAGTTCCTGTTACTG coh1_all.seq
8149  AGATGGAGAGAGTTTTAAACTGTTTTCGGATAGAAAAGTTCCTGTTACTG cjb111_all.seq
8151  AGATGGAGAGAGTTTTAAACTGTTTTCGGATAGAAAAGTTCCTGTTACTG nem316_all.seq
8150  AGATGGAGAGAGTTTTAAACTGTTTTCGGATAGAAAAGTTCCTGTTACTG a909_all.seq

      GAGGAACGACACAAGCAGCTTATCGAGTACCGCAAAATCAACTCTCTGTA Majority
                8210      8220      8230      8240      8250
8194  GAGGAACGACACAAGCAGCTTATCGAGTACCGCAAAATCAACTCTCTGTA 2603_all.seq
7967  GAGGAACGACACAAGCAGCTTATCGAGTACCGCAAAATCAACTCTCTGTA 18rs21_all.seq
8199  GAGGAACGACACAAGCAGCTTATCGAGTACCGCAAAATCAACTCTCTGTA coh1_all.seq
8199  GAGGAACGACACAAGCAGCTTATCGAGTACCGCAAAATCAACTCTCTGTA cjb111_all.seq
8201  GAGGAACGACACAAGCAGCTTATCGAGTACCGCAAAATCAACTCTCTGTA nem316_all.seq
8200  GAGGAACGACACAAGCAGCTTATCGAGTACCGCAAAATCAACTCTCTGTA a909_all.seq

      ATGAGTAATGAGGGATATGCAATTAATAGTGGATATATTTATCTCTATTG Majority
                8260      8270      8280      8290      8300
8244  ATGAGTAATGAGGGATATGCAATTAATAGTGGATATATTTATCTCTATTG 2603_all.seq
8017  ATGAGTAATGAGGGATATGCAATTAATAGTGGATATATTTATCTCTATTG 18rs21_all.seq
8249  ATGAGTAATGAGGGATATGCAATTAATAGTGGATATATTTATCTCTATTG coh1_all.seq
8249  ATGAGTAATGAGGGATATGCAATTAATAGTGGATATATTTATCTCTATTG cjb111_all.seq
8251  ATGAGTAATGAGGGATATGCAATTAATAGTGGATATATTTATCTCTATTG nem316_all.seq
8250  ATGAGTAATGAGGGATATGCAATTAATAGTGGATATATTTATCTCTATTG a909_all.seq

      GAGAGATTACAACTGGGTCTATCCATTTGATCCTAAGACAAAGAAAGTTT Majority
                8310      8320      8330      8340      8350
8294  GAGAGATTACAACTGGGTCTATCCATTTGATCCTAAGACAAAGAAAGTTT 2603_all.seq
8067  GAGAGATTACAACTGGGTCTATCCATTTGATCCTAAGACAAAGAAAGTTT 18rs21_all.seq
8299  GAGAGATTACAACTGGGTCTATCCATTTGATCCTAAGACAAAGAAAGTTT coh1_all.seq
8299  GAGAGATTACAACTGGGTCTATCCATTTGATCCTAAGACAAAGAAAGTTT cjb111_all.seq
8301  GAGAGATTACAACTGGGTCTATCCATTTGATCCTAAGACAAAGAAAGTTT nem316_all.seq
8300  GAGAGATTACAACTGGGTCTATCCATTTGATCCTAAGACAAAGAAAGTTT a909_all.seq

      CTGCAACGAAACAAATCAAAACTCATGGTGAGCCAACAACATTATACTTT Majority
                8360      8370      8380      8390      8400
8344  CTGCAACGAAACAAATCAAAACTCATGGTGAGCCAACAACATTATACTTT 2603_all.seq
8117  CTGCAACGAAACAAATCAAAACTCATGGTGAGCCAACAACATTATACTTT 18rs21_all.seq
8349  CTGCAACGAAACAAATCAAAACTCATGGTGAGCCAACAACATTATACTTT coh1_all.seq
8349  CTGCAACGAAACAAATCAAAACTCATGGTGAGCCAACAACATTATACTTT cjb111_all.seq
8351  CTGCAACGAAACAAATCAAAACTCATGGTGAGCCAACAACATTATACTTT nem316_all.seq
8350  CTGCAACGAAACAAATCAAAACTCATGGTGAGCCAACAACATTATACTTT a909_all.seq
```

FIGURE 18 W

```
      AATGGAAATATAAGACCTAAAGGTTATGACATTTTTACTGTTGGGATTGG Majority
               8410      8420      8430      8440      8450
8394  AATGGAAATATAAGACCTAAAGGTTATGACATTTTTACTGTTGGGATTGG 2603_all.seq
8167  AATGGAAATATAAGACCTAAAGGTTATGACATTTTTACTGTTGGGATTGG 18rs21_all.seq
8399  AATGGAAATATAAGACCTAAAGGTTATGACATTTTTACTGTTGGGATTGG coh1_all.seq
8399  AATGGAAATATAAGACCTAAAGGTTATGACATTTTTACTGTTGGGATTGG cjb111_all.seq
8401  AATGGAAATATAAGACCTAAAGGTTATGACATTTTTACTGTTGGGATTGG nem316_all.seq
8400  AATGGAAATATAAGACCTAAAGGTTATGACATTTTTACTGTTGGGATTGG a909_all.seq

      TGTAAACGGAGATCCTGGTGCAACTCCTCTTGAAGCTGAGAAATTTATGC Majority
               8460      8470      8480      8490      8500
8444  TGTAAACGGAGATCCTGGTGCAACTCCTCTTGAAGCTGAGAAATTTATGC 2603_all.seq
8217  TGTAAACGGAGATCCTGGTGCAACTCCTCTTGAAGCTGAGAAATTTATGC 18rs21_all.seq
8449  TGTAAACGGAGATCCTGGTGCAACTCCTCTTGAAGCTGAGAAATTTATGC coh1_all.seq
8449  TGTAAACGGAGATCCTGGTGCAACTCCTCTTGAAGCTGAGAAATTTATGC cjb111_all.seq
8451  TGTAAACGGAGATCCTGGTGCAACTCCTCTTGAAGCTGAGAAATTTATGC nem316_all.seq
8450  TGTAAACGGAGATCCTGGTGCAACTCCTCTTGAAGCTGAGAAATTTATGC a909_all.seq

      AATCAATATCAAGTAAAACAGAAAATTATACTAATGTTGATGATACAAAT Majority
               8510      8520      8530      8540      8550
8494  AATCAATATCAAGTAAAACAGAAAATTATACTAATGTTGATGATACAAAT 2603_all.seq
8267  AATCAATATCAAGTAAAACAGAAAATTATACTAATGTTGATGATACAAAT 18rs21_all.seq
8499  AATCAATATCAAGTAAAACAGAAAATTATACTAATGTTGATGATACAAAT coh1_all.seq
8499  AATCAATATCAAGTAAAACAGAAAATTATACTAATGTTGATGATACAAAT cjb111_all.seq
8501  AATCAATATCAAGTAAAACAGAAAATTATACTAATGTTGATGATACAAAT nem316_all.seq
8500  AATCAATATCAAGTAAAACAGAAAATTATACTAATGTTGATGATACAAAT a909_all.seq

      AAAATTTATGATGAGCTAAATAAATACTTTAAAACAATTGTTGAGGAAAA Majority
               8560      8570      8580      8590      8600
8544  AAAATTTATGATGAGCTAAATAAATACTTTAAAACAATTGTTGAGGAAAA 2603_all.seq
8317  AAAATTTATGATGAGCTAAATAAATACTTTAAAACAATTGTTGAGGAAAA 18rs21_all.seq
8549  AAAATTTATGATGAGCTAAATAAATACTTTAAAACAATTGTTGAGGAAAA coh1_all.seq
8549  AAAATTTATGATGAGCTAAATAAATACTTTAAAACAATTGTTGAGGAAAA cjb111_all.seq
8551  AAAATTTATGATGAGCTAAATAAATACTTTAAAACAATTGTTGAGGAAAA nem316_all.seq
8550  AAAATTTATGATGAGCTAAATAAATACTTTAAAACAATTGTTGAGGAAAA a909_all.seq

      ACATTCTATTGTTGATGGAAATGTGACTGATCCTATGGGAGAGATGATTG Majority
               8610      8620      8630      8640      8650
8594  ACATTCTATTGTTGATGGAAATGTGACTGATCCTATGGGAGAGATGATTG 2603_all.seq
8367  ACATTCTATTGTTGATGGAAATGTGACTGATCCTATGGGAGAGATGATTG 18rs21_all.seq
8599  ACATTCTATTGTTGATGGAAATGTGACTGATCCTATGGGAGAGATGATTG coh1_all.seq
8599  ACATTCTATTGTTGATGGAAATGTGACTGATCCTATGGGAGAGATGATTG cjb111_all.seq
8601  ACATTCTATTGTTGATGGAAATGTGACTGATCCTATGGGAGAGATGATTG nem316_all.seq
8600  ACATTCTATTGTTGATGGAAATGTGACTGATCCTATGGGAGAGATGATTG a909_all.seq

      AATTCCAATTAAAAAATGGTCAAAGTTTTACACATGATGATTACGTTTTG Majority
               8660      8670      8680      8690      8700
8644  AATTCCAATTAAAAAATGGTCAAAGTTTTACACATGATGATTACGTTTTG 2603_all.seq
8417  AATTCCAATTAAAAAATGGTCAAAGTTTTACACATGATGATTACGTTTTG 18rs21_all.seq
8649  AATTCCAATTAAAAAATGGTCAAAGTTTTACACATGATGATTACGTTTTG coh1_all.seq
8649  AATTCCAATTAAAAAATGGTCAAAGTTTTACACATGATGATTACGTTTTG cjb111_all.seq
8651  AATTCCAATTAAAAAATGGTCAAAGTTTTACACATGATGATTACGTTTTG nem316_all.seq
8650  AATTCCAATTAAAAAATGGTCAAAGTTTTACACATGATGATTACGTTTTG a909_all.seq

      GTTGGAAATGATGGCAGTCAATTAAAAAATGGTGTGGCTCTTGGTGGACC Majority
               8710      8720      8730      8740      8750
8694  GTTGGAAATGATGGCAGTCAATTAAAAAATGGTGTGGCTCTTGGTGGACC 2603_all.seq
8467  GTTGGAAATGATGGCAGTCAATTAAAAAATGGTGTGGCTCTTGGTGGACC 18rs21_all.seq
8699  GTTGGAAATGATGGCAGTCAATTAAAAAATGGTGTGGCTCTTGGTGGACC coh1_all.seq
8699  GTTGGAAATGATGGCAGTCAATTAAAAAATGGTGTGGCTCTTGGTGGACC cjb111_all.seq
8701  GTTGGAAATGATGGCAGTCAATTAAAAAATGGTGTGGCTCTTGGTGGACC nem316_all.seq
8700  GTTGGAAATGATGGCAGTCAATTAAAAAATGGTGTGGCTCTTGGTGGACC a909_all.seq
```

FIGURE 18 X

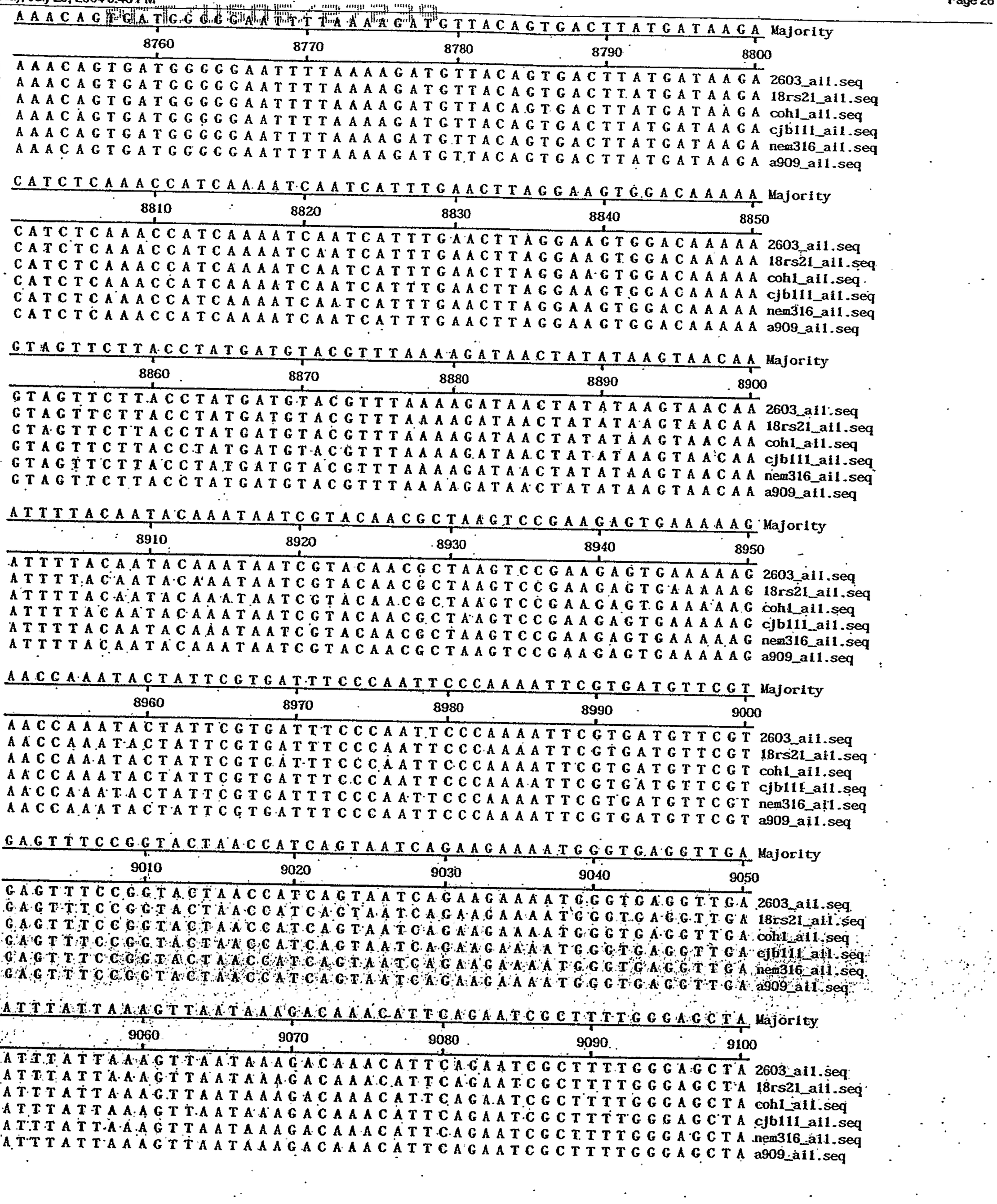

AAACAGTGATGGCGCGAATTTTAAAAGATGTTACAGTGACTTATGATAAGA Majority
8760 8770 8780 8790 8800
8744 AAACAGTGATGGCGCGAATTTTAAAAGATGTTACAGTGACTTATGATAAGA
8517 AAACAGTGATGCGCGCAATTTTAAAAGATGTTACAGTGACTTATGATAAGA 2603_all.seq
8749 AAACAGTGATGGGCGAATTTTAAAAGATGTTACAGTGACTTATGATAAGA 18rs21_all.seq
8749 AAACAGTGATGGGCGAATTTTAAAAGATGTTACAGTGACTTATGATAAGA coh1_all.seq
8751 AAACAGTGATGGGCGAATTTTAAAAGATGTTACAGTGACTTATGATAAGA cjb111_all.seq
8750 AAACAGTGATGGCGCGAATTTTAAAAGATGTTACAGTGACTTATGATAAGA nem316_all.seq
a909_all.seq
CATCTCAAACCATCAAAATCAATCATTTGAACTTAGGAAGTGGACAAAAAA Majority
8810 8820 8830 8840 8850
8794 CATCTCAAACCATCAAAATCAATCATTTGAACTTAGGAAGTGGACAAAAAA
8567 CATCTCAAACCATCAAAATCAATCATTTGAACTTAGGAAGTGGACAAAAAA 2603_all.seq
8799 CATCTCAAACCATCAAAATCAATCATTTGAACTTAGGAAGTGGACAAAAAA 18rs21_all.seq
8799 CATCTCAAACCATCAAAATCAATCATTTGAACTTAGGAAGTGGACAAAAAA coh1_all.seq
8801 CATCTCAAACCATCAAAATCAATCATTTGAACTTAGGAAGTGGACAAAAAA cjb111_all.seq
8800 CATCTCAAACCATCAAAATCAATCATTTGAACTTAGGAAGTGGACAAAAAA nem316_all.seq
a909_all.seq
GTAGTTCTTACCTATGATGTACGTTTAAAAGATAACTATATAAGTAACAA Majority
8860 8870 8880 8890 8900
8844 GTAGTTCTTACCTATGATGTACGTTTAAAAGATAACTATATAAGTAACAA
8617 GTAGTTCTTACCTATGATGTACGTTTAAAAGATAACTATATAAGTAACAA 2603_all.seq
8849 GTAGTTCTTACCTATGATGTACGTTTAAAAGATAACTATATAAGTAACAA 18rs21_all.seq
8849 GTAGTTCTTACCTATGATGTACGTTTAAAAGATAACTATATAAGTAACAA coh1_all.seq
8851 GTAGTTCTTACCTATGATGTACGTTTAAAAGATAACTATATAAGTAACAA cjb111_all.seq
8850 GTAGTTCTTACCTATGATGTACGTTTAAAAGATAACTATATAAGTAACAA nem316_all.seq
a909_all.seq
ATTTTACAATACAAATAATCGTACAACGCTAAGTCCGAAGAGTGAAAAAG Majority
8910 8920 8930 8940 8950
8894 ATTTTACAATACAAATAATCGTACAACGCTAAGTCCGAAGAGTGAAAAAG
8667 ATTTTACAATACAAATAATCGTACAACGCTAAGTCCGAAGAGTGAAAAAG 2603_all.seq
8899 ATTTTACAATACAAATAATCGTACAACGCTAAGTCCGAAGAGTGAAAAAG 18rs21_all.seq
8899 ATTTTACAATACAAATAATCGTACAACGCTAAGTCCGAAGAGTGAAAAAG coh1_all.seq
8901 ATTTTACAATACAAATAATCGTACAACGCTAAGTCCGAAGAGTGAAAAAG cjb111_all.seq
8900 ATTTTACAATACAAATAATCGTACAACGCTAAGTCCGAAGAGTGAAAAAG nem316_all.seq
a909_all.seq
AACCAAATACTATTCGTGATTTCCCAATTCCCAAAATTCGTGATGTTCGT Majority
8960 8970 8980 8990 9000
8944 AACCAAATACTATTCGTGATTTCCCAATTCCCAAAATTCGTGATGTTCGT
8717 AACCAAATACTATTCGTGATTTCCCAATTCCCAAAATTCGTGATGTTCGT 2603_all.seq
8949 AACCAAATACTATTCGTGATTTCCCAATTCCCAAAATTCGTGATGTTCGT 18rs21_all.seq
8949 AACCAAATACTATTCGTGATTTCCCAATTCCCAAAATTCGTGATGTTCGT coh1_all.seq
8951 AACCAAATACTATTCGTGATTTCCCAATTCCCAAAATTCGTGATGTTCGT cjb111_all.seq
8950 AACCAAATACTATTCGTGATTTCCCAATTCCCAAAATTCGTGATGTTCGT nem316_all.seq
a909_all.seq
GAGTTTCCGGTACTAACCATCAGTAATCAGAAGAAAATGGGTGAGGTTGA Majority
9010 9020 9030 9040 9050
8994 GAGTTTCCGGTACTAACCATCAGTAATCAGAAGAAAATGGGTGAGGTTGA
8767 GAGTTTCCGGTACTAACCATCAGTAATCAGAAGAAAATGGGTGAGGTTGA 2603_all.seq
8999 GAGTTTCCGGTACTAACCATCAGTAATCAGAAGAAAATGGGTGAGGTTGA 18rs21_all.seq
8999 GAGTTTCCGGTACTAACCATCAGTAATCAGAAGAAAATGGGTGAGGTTGA coh1_all.seq
9001 GAGTTTCCGGTACTAACCATCAGTAATCAGAAGAAAATGGGTGAGGTTGA cjb111_all.seq
9000 GAGTTTCCGGTACTAACCATCAGTAATCAGAAGAAAATGGGTGAGGTTGA nem316_all.seq
a909_all.seq
ATTTATTAAAGTTAATAAAGACAAACATTCAGAATCGCTTTTGGGAGCTA Majority
9060 9070 9080 9090 9100
9044 ATTTATTAAAGTTAATAAAGACAAACATTCAGAATCGCTTTTGGGAGCTA
8817 ATTTATTAAAGTTAATAAAGACAAACATTCAGAATCGCTTTTGGGAGCTA 2603_all.seq
9049 ATTTATTAAAGTTAATAAAGACAAACATTCAGAATCGCTTTTGGGAGCTA 18rs21_all.seq
9049 ATTTATTAAAGTTAATAAAGACAAACATTCAGAATCGCTTTTGGGAGCTA coh1_all.seq
9051 ATTTATTAAAGTTAATAAAGACAAACATTCAGAATCGCTTTTGGGAGCTA cjb111_all.seq
9050 ATTTATTAAAGTTAATAAAGACAAACATTCAGAATCGCTTTTGGGAGCTA nem316_all.seq
a909_all.seq

FIGURE 18 Y

```
          A G T T T C A A C T T C A G A T A G A A A A A G A T T T T C T G G G T A T A A G C A A T T T G T T  Majority
                    9110        9120        9130        9140        9150
9094  A G T T T C A A C T T C A G A T A G A A A A A G A T T T T C T G G G T A T A A G C A A T T T G T T  2603_all.seq
8867  A G T T T C A A C T T C A G A T A G A A A A A G A T T T T C T G G G T A T A A G C A A T T T G T T  18rs21_all.seq
9099  A G T T T C A A C T T C A G A T A G A A A A A G A T T T T C T G G G T A T A A G C A A T T T G T T  coh1_all.seq
9099  A G T T T C A A C T T C A G A T A G A A A A A G A T T T T C T G G G T A T A A G C A A T T T G T T  cjb111_all.seq
9101  A G T T T C A A C T T C A G A T A G A A A A A G A T T T T C T G G G T A T A A G C A A T T T G T T  nem316_all.seq
9100  A G T T T C A A C T T C A G A T A G A A A A A G A T T T T C T G G G T A T A A G C A A T T T G T T  a909_all.seq

          C C A G A G G G A A G T G A T G T T A C A A C A A A G A A T G A T G G T A A A A T T T A T T T T A A  Majority
                    9160        9170        9180        9190        9200
9144  C C A G A G G G A A G T G A T G T T A C A A C A A A G A A T G A T G G T A A A A T T T A T T T T A A  2603_all.seq
8917  C C A G A G G G A A G T G A T G T T A C A A C A A A G A A T G A T G G T A A A A T T T A T T T T A A  18rs21_all.seq
9149  C C A G A G G G A A G T G A T G T T A C A A C A A A G A A T G A T G G T A A A A T T T A T T T T A A  coh1_all.seq
9149  C C A G A G G G A A G T G A T G T T A C A A C A A A G A A T G A T G G T A A A A T T T A T T T T A A  cjb111_all.seq
9151  C C A G A G G G A A G T G A T G T T A C A A C A A A G A A T G A T G G T A A A A T T T A T T T T A A  nem316_all.seq
9150  C C A G A G G G A A G T G A T G T T A C A A C A A A G A A T G A T G G T A A A A T T T A T T T T A A  a909_all.seq
```

```
          A G C A C T T C A A G A T G G T A A C T A T A A A T T A T A T G A A A T T T C A A G T C C A G A T G  Majority
                    9210        9220        9230        9240        9250
9194  A G C A C T T C A A G A T G G T A A C T A T A A A T T A T A T G A A A T T T C A A G T C C A G A T G  2603_all.seq
8967  A G C A C T T C A A G A T G G T A A C T A T A A A T T A T A T G A A A T T T C A A G T C C A G A T G  18rs21_all.seq
9199  A G C A C T T C A A G A T G G T A A C T A T A A A T T A T A T G A A A T T T C A A G T C C A G A T G  coh1_all.seq
9199  A G C A C T T C A A G A T G G T A A C T A T A A A T T A T A T G A A A T T T C A A G T C C A G A T G  cjb111_all.seq
9201  A G C A C T T C A A G A T G G T A A C T A T A A A T T A T A T G A A A T T T C A A G T C C A G A T G  nem316_all.seq
9200  A G C A C T T C A A G A T G G T A A C T A T A A A T T A T A T G A A A T T T C A A G T C C A G A T G  a909_all.seq

          G C T A T A T A G A G G T T A A A A C G A A A C C T G T T G T G A C A T T T A C A A T T C A A A A T  Majority
                    9260        9270        9280        9290        9300
9244  G C T A T A T A G A G G T T A A A A C G A A A C C T G T T G T G A C A T T T A C A A T T C A A A A T  2603_all.seq
9017  G C T A T A T A G A G G T T A A A A C G A A A C C T G T T G T G A C A T T T A C A A T T C A A A A T  18rs21_all.seq
9249  G C T A T A T A G A G G T T A A A A C G A A A C C T G T T G T G A C A T T T A C A A T T C A A A A T  coh1_all.seq
9249  G C T A T A T A G A G G T T A A A A C G A A A C C T G T T G T G A C A T T T A C A A T T C A A A A T  cjb111_all.seq
9251  G C T A T A T A G A G G T T A A A A C G A A A C C T G T T G T G A C A T T T A C A A T T C A A A A T  nem316_all.seq
9250  G C T A T A T A G A G G T T A A A A C G A A A C C T G T T G T G A C A T T T A C A A T T C A A A A T  a909_all.seq
```

```
          G G A G A A G T T A C G A A C C T G A A A G C A G A T C C A A A T G C T A A T A A A A A T C A A A T  Majority
                    9310        9320        9330        9340        9350
294   G G A G A A G T T A C G A A C C T G A A A G C A G A T C C A A A T G C T A A T A A A A A T C A A A T  2603_all.seq
067   G G A G A A G T T A C G A A C C T G A A A G C A G A T C C A A A T G C T A A T A A A A A T C A A A T  18rs21_all.seq
299   G G A G A A G T T A C G A A C C T G A A A G C A G A T C C A A A T G C T A A T A A A A A T C A A A T  coh1_all.seq
299   G G A G A A G T T A C G A A C C T G A A A G C A G A T C C A A A T G C T A A T A A A A A T C A A A T  cjb111_all.seq
301   G G A G A A G T T A C G A A C C T G A A A G C A G A T C C A A A T G C T A A T A A A A A T C A A A T  nem316_all.seq
300   G G A G A A G T T A C G A A C C T G A A A G C A G A T C C A A A T G C T A A T A A A A A T C A A A T  a909_all.seq

          C G G G T A T C T T G A A G G A A A T G G T A A A C A T C T T A T T A C C A A C A C T C C C A A A C  Majority
                    9360        9370        9380        9390        9400
344   C G G G T A T C T T G A A G G A A A T G G T A A A C A T C T T A T T A C C A A C A C T C C C A A A C  2603_all.seq
117   C G G G T A T C T T G A A G G A A A T G G T A A A C A T C T T A T T A C C A A C A C T C C C A A A C  18rs21_all.seq
349   C G G G T A T C T T G A A G G A A A T G G T A A A C A T C T T A T T A C C A A C A C T C C C A A A C  coh1_all.seq
349   C G G G T A T C T T G A A G G A A A T G G T A A A C A T C T T A T T A C C A A C A C T C C C A A A C  cjb111_all.seq
351   C G G G T A T C T T G A A G G A A A T G G T A A A C A T C T T A T T A C C A A C A C T C C C A A A C  nem316_all.seq
350   C G G G T A T C T T G A A G G A A A T G G T A A A C A T C T T A T T A C C A A C A C T C C C A A A C  a909_all.seq

          G C C C A C C A G G T G T T T T T C C T A A A A C A G G G G G A A T T G G T A C A A T T G T C T A T  Majority
                    9410        9420        9430        9440        9450
394   G C C C A C C A G G T G T T T T T C C T A A A A C A G G G G G A A T T G G T A C A A T T G T C T A T  2603_all.seq
167   G C C C A C C A G G T G T T T T T C C T A A A A C A G G G G G A A T T G G T A C A A T T G T C T A T  18rs21_all.seq
399   G C C C A C C A G G T G T T T T T C C T A A A A C A G G G G G A A T T G G T A C A A T T G T C T A T  coh1_all.seq
399   G C C C A C C A G G T G T T T T T C C T A A A A C A G G G G G A A T T G G T A C A A T T G T C T A T  cjb111_all.seq
401   G C C C A C C A G G T G T T T T T C C T A A A A C A G G G G G A A T T G G T A C A A T T G T C T A T  nem316_all.seq
400   G C C C A C C A G G T G T T T T T C C T A A A A C A G G G G G A A T T G G T A C A A T T G T C T A T  a909_all.seq
```

FIGURE 18 Z

```
ATATTAGTTGCTTCTACTTTTATGATACTTACCATTTGTTCTTTCCGTCG Majority
          9460      9470      9480      9490      9500
9444 ATATTAGTTGCTTCTACTTTTATGATACTTACCATTTGTTCTTTCCGTCG 2603_all.seq
9217 ATATTAGTTGGTTCTACTTTTATGATACTTACCATTTGTTCTTTCCGTCG 2603_all.seq
9449 ATATTAGTTGGTTCTACTTTTATGATACTTACCATTTGTTCTTTCCGTCG 18rs21_all.seq
9449 ATATTAGTTGGTTCTACTTTTATGATACTTACCATTTGTTCTTTCCGTCG coh1_all.seq
9451 ATATTAGTTGGTTCTACTTTTATGATACTTACCATTTGTTCTTTCCGTCG cjb111_all.seq
9450 ATATTAGTTGGTTCTACTTTTATGATACTTACCATTTGTTCTTTCCGTCG nem316_all.seq
9450 ATATTAGTTGGTTCTACTTTTATGATACTTACCATTTGTTCTTTCCGTCG a909_all.seq

TAAACAATTGTAAGGTGTCGTTGAAATTATTAAATAATAGAAAATGACTA Majority
          9510      9520      9530      9540      9550
9494 TAAACAATTGTAAGGTGTCGTTGAAATTATTAAATAATAGAAAATGACTA 2603_all.seq
9267 TAAACAATTGTAAGGTGTCGTTGAAATTATTAAATAATAGAAAATGACTA 18rs21_all.seq
9499 TAAACAATTGTAAGGTGTCGTTGAAATTATTAAATAATAGAAAATGACTA coh1_all.seq
9499 TAAACAATTGTAAGGTGTCGTTGAAATTATTAAATAATAGAAAATGACTA cjb111_all.seq
9501 TAAACAATTGTAAGGTGTCGTTGAAATTATTAAATAATAGAAAATGACTA nem316_all.seq
9500 TAAACAATTGTAAGGTGTCGTTGAAATTATTAAATAATAGAAAATGACTA a909_all.seq

GTTTTGGCCTTTCCCTATTGTCAGTCAGATTAGTTATTATCAAGCTTCTC Majority
          9560      9570      9580      9590      9600
9544 GTTTTGGCCTTTCCCTATTGTCAGTCAGATTAGTTATTATCAAGCTTCTC 2603_all.seq
9317 GTTTTGGCCTTTCCCTATTGTCAGTCAGATTAGTTATTATCAAGCTTCTC 18rs21_all.seq
9549 GTTTTGGCCTTTCCCTATTGTCAGTCAGATTAGTTATTATCAAGCTTCTC coh1_all.seq
9549 GTTTTGGCCTTTCCCTATTGTCAGTCAGATTAGTTATTATCAAGCTTCTC cjb111_all.seq
9551 GTTTTGGCCTTTCCCTATTGTCAGTCAGATTAGTTATTATCAAGCTTCTC nem316_all.seq
9550 GTTTTGGCCTTTCCCTATTGTCAGTCAGATTAGTTATTATCAAGCTTCTC a909_all.seq

ATGCCAATATTAATGCCTTTAAAAGAACAGTTACAACCATTGACCGTACG Majority
          9610      9620      9630      9640      9650
9594 ATGCCAATATTAATGCCTTTAAAAGAACAGTTACAACCATTGACCGTACG 2603_all.seq
9367 ATGCCAATATTAATGCCTTTAAAAGAACAGTTACAACCATTGACCGTACG 18rs21_all.seq
9599 ATGCCAATATTAATGCCTTTAAAAGAACAGTTACAACCATTGACCGTACG coh1_all.seq
9599 ATGCCAATATTAATGCCTTTAAAAGAACAGTTACAACCATTGACCGTACG cjb111_all.seq
9601 ATGCCAATATTAATGCCTTTAAAAGAACAGTTACAACCATTGACCGTACG nem316_all.seq
9600 ATGCCAATATTAATGCCTTTAAAAGAACAGTTACAACCATTGACCGTACG a909_all.seq

GAGATTCAACGACGGTTAGGTCTGGCTAAAGCCTACAATGCTAGTATTTC Majority
          9660      9670      9680      9690      9700
9644 GAGATTCAACGACGGTTAGGTCTGGCTAAAGCCTACAATGCTAGTATTTC 2603_all.seq
9417 GAGATTCAACGACGGTTAGGTCTGGCTAAAGCCTACAATGCTAGTATTTC 18rs21_all.seq
9649 GAGATTCAACGACGGTTAGGTCTGGCTAAAGCCTACAATGCTAGTATTTC coh1_all.seq
9649 GAGATTCAACGACGGTTAGGTCTGGCTAAAGCCTACAATGCTAGTATTTC cjb111_all.seq
9651 GAGATTCAACGACGGTTAGGTCTGTCTAAAGCCTACAATGCTAGTATTTC nem316_all.seq
9650 GAGATTCAACGACGGTTAGGTCTGGCTAAAGCCTACAATGCTAGTATTTC a909_all.seq

TGGAACAAGTAGTCAGTCGACTCAATCTGTGCTGAGAGATTCTTATTCTG Majority
          9710      9720      9730      9740      9750
9694 TGGAACAAGTAGTCAGTCGACTCAATCTGTGCTGAGAGATTCTTATTCTG 2603_all.seq
9467 TGGAACAAGTAGTCAGTCGACTCAATCTGTGCTGAGAGATTCTTATTCTG 18rs21_all.seq
9699 TGGAACAAGTAGTCAGTCGACTCAATCTGTGCTGAGAGATTCTTATTCTG coh1_all.seq
9699 TGGAACAAGTAGTCAGTCGACTCAATCTGTGCTGAGAGATTCTTATTCTG cjb111_all.seq
9701 TGGAACAAGTAGTCAGTCGACTCAATCTGTGCTGAGAGATTCTTATTCTG nem316_all.seq
9700 TGGAACAAGTAGTCAGTCGACTCAATCTGTGCTGAGAGATTCTTATTCTG a909_all.seq

AGGAGCAAAAAAAGGCAGGGTGGACTGAATACGCTAGGATGTTAGAAGTCA Majority
          9760      9770      9780      9790      9800
9744 AGGAGCAAAAAAAGGCAGGGTGGACTGAATACGCTAGGATGTTAGAAGTCA 2603_all.seq
9517 AGGAGCAAAAAAAGGCAGGGTGGACTGAATACGCTAGGATGTTAGAAGTCA 18rs21_all.seq
9749 AGGAGCAAAAAAAGGCAGGGTGGACTGAATACGCTAGGATGTTAGAAGTCA coh1_all.seq
9749 AGGAGCAAAAAAAGGCAGGGTGGACTGAATACGCTAGGATGTTAGAAGTCA cjb111_all.seq
9751 AGGAGCAAAAAAAGGCAGGGTGGACTGAATACGCTAGGATGTTAGAAGTCA nem316_all.seq
9750 AGGAGCAAAAAAAGGCAGGGTGGACTGAATACGCTAGGATGTTAGAAGTCA a909_all.seq
```

FIGURE 18 AA

```
      GAGAGCAGGTTGACCATGTGATGATTCCAAAAATCAATCAGGATTTACCA Majority
                9810      9820      9830      9840      9850
9794  GAGAGCAGGTTGACCATGTGATGATTCCAAAAATCAATCAGGATTTACCA 2603_a11.seq
9567  GAGAGCAGGTTGACCATGTGATGATTCCAAAAATCAATCAGGATTTACCA 18rs21_a11.seq
9799  GAGAGCAGGTTGACCATGTGATGATTCCAAAAATCAATCAGGATTTACCA coh1_a11.seq
9799  GAGAGCAGGTTGACCATGTGATGATTCCAAAAATCAATCAGGATTTACCA cjb111_a11.seq
9801  GAGAGCAGGTTGACCATGTGATGATTCCAAAAATCAATCAGGATTTACCA nem316_a11.seq
9800  GAGAGCAGGTTGACCATGTGATGATTCCAAAAATCAATCAGGATTTACCA a909_a11.seq

      ATCTACGCTGGTTCAGAAGAGGACAATCTGCAACGGGGAGTTGGTCATCT Majority
                9860      9870      9880      9890      9900
9844  ATCTACGCTGGTTCAGAAGAGGACAATCTGCAACGGGGAGTTGGTCATCT 2603_a11.seq
9617  ATCTACGCTGGTTCAGAAGAGGACAATCTGCAACGGGGAGTTGGTCATCT 18rs21_a11.seq
9849  ATCTACGCTGGTTCAGAAGAGGACAATCTGCAACGGGGAGTTGGTCATCT coh1_a11.seq
9849  ATCTACGCTGGTTCAGAAGAGGACAATCTGCAACGGGGAGTTGGTCATCT cjb111_a11.seq
9851  ATCTACGCTGGTCCAGAAGAGGACAATCTGCAACGGGGAGTTGGTCATCT nem316_a11.seq
9850  ATCTACGCTGGTTCAGAAGAGGACAATCTGCAACGGGGAGTTGGTCATCT a909_a11.seq

      AGAAGGGATAAGTTTGCCGATTGGAGGGGCTTCTACACATGCGGTCTTGA Majority
                9910      9920      9930      9940      9950
9894  AGAAGGGATAAGTTTGCCGATTGGAGGGGCTTCTACACATGCGGTCTTGA 2603_a11.seq
9667  AGAAGGGATAAGTTTGCCGATTGGAGGGGCTTCTACACATGCGGTCTTGA 18rs21_a11.seq
9899  AGAAGGGATAAGTTTGCCGATTGGAGGGGCTTCTACACATGCGGTCTTGA coh1_a11.seq
9899  AGAAGGGATAAGTTTGCCGATTGGAGGGGCTTCTACACATGCGGTCTTGA cjb111_a11.seq
9901  AGAAGGGATAAGTTTGCCGATTGGAGGGGCTTCTACACATGCGGTCTTGA nem316_a11.seq
9900  AGAAGGGATAAGTTTGCCGATTGGAGGGGCTTCTACACATGCGGTCTTGA a909_a11.seq

      GCGGTCAAAGAGGTATGCCAGCTGCTCGGTTGTTTGCGGATTTGGATAAG Majority
                9960      9970      9980      9990      10000
9944  GCGGTCAAAGAGGTATGCCAGCTGCTCGGTTGTTTGCGGATTTGGATAAG 2603_a11.seq
9717  GCGGTCAAAGAGGTATGCCAGCTGCTCGGTTGTTTGCGGATTTGGATAAG 18rs21_a11.seq
9949  GCGGTCAAAGAGGTATGCCAGCTGCTCGGTTGTTTGCGGATTTGGATAAG coh1_a11.seq
9949  GCGGTCAAAGAGGTATGTCAGCTGCTCGGTTGTTTGCGGATTTGGATAAG cjb111_a11.seq
9951  GCGGTCAAAGAGGTATGCCAGCTGCTCGGTTGTTTGCGGATTTGGATAAG nem316_a11.seq
9950  GCGGTCAAAGAGGTATGTCAGCTGCTCGGTTGTTTGCGGATTTGGATAAG a909_a11.seq

      ATGAAAAAAAGGTGATTATTTTTATGTTACCAATCTGAAAGAAACCTTGGC Majority
                10010     10020     10030     10040     10050
994   ATGAAAAAAAGGTGATTATTTTTATGTTACCAATCTGAAAGAAACCTTGGC 2603_a11.seq
767   ATGAAAAAAAGGTGATTATTTTTATGTTACCAATCTGAAAGAAACCTTGGC 18rs21_a11.seq
999   ATGAAAAAAAGGTGATTATTTTTATGTTACCAATCTGAAAGAAACCTTGGC coh1_a11.seq
999   ATGAAAAAAAGGTGATTATTTTTATGTTACCAATCTGAAAGAAACCTTGGC cjb111_a11.seq
0001  ATGAAAAAAAGGTGATTATTTTTATGTTACCAATCTGAAAGAAACCTTGGC nem316_a11.seq
0000  ATGAAAAAAAGGTGATTATTTTTATGTTACCAATCTGAAAGAAACCTTGGC a909_a11.seq

      TTATCAAGTGGATCGTATCATGGTGATTGAACCTAGCCAATTGGATGCCG Majority
                10060     10070     10080     10090     10100
0044  TTATCAAGTGGATCGTATCATGGTGATTGAACCTAGCCAATTGGATGCCG 2603_a11.seq
817   TTATCAAGTGGATCGTATCATGGTGATTGAACCTAGCCAATTGGATGCCG 18rs21_a11.seq
0049  TTATCAAGTGGATCGTATCATGGTGATTGAACCTAGCCAATTGGATGCCG coh1_a11.seq
0049  TTATCAAGTGGATCGTATCATGGTGATTGAACCTAGCCAATTGGATGCCG cjb111_a11.seq
0051  TTATCAAGTGGATCGTATCATGGTGATTGAACCTAGCCAATTGGATGCCG nem316_a11.seq
0050  TTATCAAGTGGATCGTATCATGGTGATTGAACCTAGCCAATTGGATGCCG a909_a11.seq

      TGAGCATTGAAGAGGATAAAGATTATGTTACCCTTCTGACCTGTACACCT Majority
                10110     10120     10130     10140     10150
0094  TGAGCATTGAAGAGGATAAAGATTATGTTACCCTTCTGACCTGTACACCT 2603_a11.seq
867   TGAGCATTGAAGAGGATAAAGATTATGTTACCCTTCTGACCTGTACACCT 18rs21_a11.seq
0099  TGAGCATTGAAGAGGATAAAGATTATGTTACCCTTCTGACCTGTACACCT coh1_a11.seq
0099  TGAGCATTGAAGAGGATAAAGATTATGTTACCCTTCTGACCTGTACACCT cjb111_a11.seq
0101  TGAGCATTGAAGAGGATAAAGATTATGTTACCCTTCTGACCTGTACACCT nem316_a11.seq
0100  TGAGCATTGAAGAGGATAAAGATTATGTTACCCTTCTGACCTGTACACCT a909_a11.seq
```

FIGURE 18 AB

```
                                                
```

```
      TAAAGGTCGTTATAATAGCGGAGCTCATCTGAGAAAACTTTTATACCTCA Majority
               10510      10520      10530      10540      10550
10494 TAAAGGTCGTTATAATAGCGGAGCTCATCTGAGAAAACTTTTATACCTCA 2603_ail.seq
10267 TAAAGGTCGTTATAATAGCGGAGCTCATCTGAGAAAACTTTTATACCTCA 18rs21_ail.seq
10499 TAAAGGTCGTTATAATAGCGGAGCTCATCTGAGAAAACTTTTATACCTCA coh1_ail.seq
10499 TAAAGGTCGTTATAATAGCGGAGCTCATCTGAGAAAACTTTTATACCTCA cjb111_ail.seq
10501 TAAAGGTCGTTATAATAGCGGAGCTCATCTGAGAAAACTTTTATACCTCA nem316_ail.seq
10500 TAAAGGTCGTTATAATAGCGGAGCTCATCTGAGAAAACTTTTATACCTCA a909_ail.seq

      AAGTCAGTCTAGCTTTGATATCATGAAGCCATTAGGAGTTATTCCTTATC Majority
               10560      10570      10580      10590      10600
10544 AAGTCAGTCTAGCTTTGATATCATGAAGCCATTAGGAGTTATTCCTTATC 2603_ail.seq
10317 AAGTCAGTCTAGCTTTGATATCATGAAGCCATTAGGAGTTATTCCTTATC 18rs21_ail.seq
10549 AAGTCAGTCTAGCTTTGATATCATGAAGCCATTAGGAGTTATTCCTTATC coh1_ail.seq
10549 AAGTCAGTCTAGCTTTGATATCATGAAGCCATTAGGAGTTATTCCTTATC cjb111_ail.seq
10551 AAGTCAGTCTAGCTTTGATATCATGAAGCCATTAGGAGTTATTCCTTATC nem316_ail.seq
10550 AAGTCAGTCTAGCTTTGATATCATGAAGCCATTAGGAGTTATTCCTTATC a909_ail.seq

      TTTTAGTGGCGCGCGATCCATATAGTGATAGATCGAGATATTTAGATCCA Majority
               10610      10620      10630      10640      10650
10594 TTTTAGTGGCGCGCGATCCATATAGTGATAGATCGAGATATTTAGATCCA 2603_ail.seq
10367 TTTTAGTGGCGCGCGATCCATATAGTGATAGATCGAGATATTTAGATCCA 18rs21_ail.seq
10599 TTTTAGTGGCGCGCGATCCATATAGTGATAGATCGAGATATTTAGATCCA coh1_ail.seq
10599 TTTTAGTGGCGCGCGATCCATATAGTGATAGATCGAGATATTTAGATCCA cjb111_ail.seq
10601 TTTTAGTGGCGCGCGATCCATATAGTGATAGATCGAGATATTTAGATCCA nem316_ail.seq
10600 TTTTAGTGGCGCGCGATCCATATAGTGATAGATCGAGATATTTAGATCCA a909_ail.seq

      AAAGTTCTATCATCCTCTTTGGCGCCTTTTTTCCAGCAGATAATATTAA Majority
               10660      10670      10680      10690      10700
10644 AAAGTTCTATCATCCTCTTTGGCGCCTTTTTTCCAGCAGATAATATTAA 2603_ail.seq
10417 AAAGTTCTATCATCCTCTTTGGCGCCTTTTTTCCAGCAGATAATATTAA 18rs21_ail.seq
10649 AAAGTTCTATCATCCTCTTTGGCGCCTTTTTTCCAGCAGATAATATTAA coh1_ail.seq
10649 AAAGTTCTATCATCCTCTTTGGCGCCTTTTTTCCAGCAGATAATATTAA cjb111_ail.seq
10651 AAAGTTCTATCATCCTCTTTGGCGCCTTTTTTCCAGCAGATAATATTAA nem316_ail.seq
10650 AAAGTTCTATCATCCTCTTTGGCGCCTTTTTTCCAGCAGATAATATTAA a909_ail.seq

      GGTAGCTTGGTCTAACAACTCCAGCACTTTATTTACACCACCTATTAATG Majority
               10710      10720      10730      10740      10750
0694  GGTAGCTTGGTCTAACAACTCCAGCACTTTATTTACACCACCTATTAATG 2603_ail.seq
0467  GGTAGCTTGGTCTAACAACTCCAGCACTTTATTTACACCACCTATTAATG 18rs21_ail.seq
0699  GGTAGCTTGGTCTAACAACTCCAGCACTTTATTTACACCACCTATTAATG coh1_ail.seq
0699  GGTAGCTTGGTCTAACAACTCCAGCACTTTATTTACACCACCTATTAATG cjb111_ail.seq
0701  GGTAGCTTGGTCTAACAACTCCAGCACTTTATTTACACCACCTATTAATG nem316_ail.seq
0700  GGTAGCTTGGTCTAACAACTCCAGCACTTTATTTACACCACCTATTAATG a909_ail.seq

      CAAACTACACCACTCAGATTCAAGCTATTGGGACAACGATTAAGTCACAA Majority
               10760      10770      10780      10790      10800
0744  CAAACTACACCACTCAGATTCAAGCTATTGGGACAACGATTAAGTCACAA 2603_ail.seq
0517  CAAACTACACCACTCAGATTCAAGCTATTGGGACAACGATTAAGTCACAA 18rs21_ail.seq
0749  CAAACTACACCACTCAGATTCAAGCTATTGGGACAACGATTAAGTCACAA coh1_ail.seq
0749  CAAACTACACCACTCAGATTCAAGCTATTGGGACAACGATTAAGTCACAA cjb111_ail.seq
0751  CAAACTACACCACTCAGATTCAAGCTATTGGGACAACGATTAAGTCACAA nem316_ail.seq
0750  CAAACTACACCACTCAGATTCAAGCTATTGGGACAACGATTAAGTCACAA a909_ail.seq

      ATTCCGGAATCGATTTTGACGGTTACGGATAAAAAGAGCAGGAAGTTCAG Majority
               10810      10820      10830      10840      10850
0794  ATTCCGGAATCGATTTTGACGGTTACGGATAAAAAGAGCAGGAAGTTCAG 2603_ail.seq
0567  ATTCCGGAATCGATTTTGACGGTTACGGATAAAAAGAGCAGGAAGTTCAG 18rs21_ail.seq
0799  ATTCCGGAATCGATTTTGACGGTTACGGATAAAAAGAGCAGGAAGTTCAG coh1_ail.seq
0799  ATTCCGGAATCGATTTTGACGGTTACGGATAAAAAGAGCAGGAAGTTCAG cjb111_ail.seq
0801  ATTCCGGAATCGATTTTGACGATTACGGATAAAAAGAGCAGGAAGTTCAG nem316_ail.seq
0800  ATTCCGGAATCGATTTTGACGGTTACGGATAAAAAGAGCAGGAAGTTCAG a909_ail.seq
```

FIGURE 18 AD

```
                                                
```

```
      GCAAAGGGTAAAAATCCTTTATTTTAAGCACTTTTTCAAGCATTTTGTCT Majority
                11210       11220       11230       11240       11250
11194 GCAAAGGGTAAAAATCCTTTATTTTAAGCACTTTTTCAAGCATTTTGTCT 2603_a11.seq
10967 GCAAAGGGTAAAAATCCTTTATTTTAAGCACTTTTTCAAGCATTTTGTCT 18rs21_a11.seq
11199 GCAAAGGGTAAAAATCCTTTATTTTAAGCACTTTTTCAAGCATTTTGTCT coh1_a11.seq
11199 GCAAAGGGTAAAAATCCTTTATTTTAAGCACTTTTTCAAGCATTTTGTCT cjb111_a11.seq
11201 GCAAAGGGTAAAAATCCTTTATTTTAAGCACTTTTTCAAGCATTTTGTCT nem316_a11.seq
11200 GCAAAGGGTAAAAATCCTTTATTTTAAGCACTTTTTCAAGCATTTTGTCT a909_a11.seq

      TTATTGAAAAGAGTGATTTTAACATAAAAAAAGGTATTAAAAAACATATTG Majority
                11260       11270       11280       11290       11300
11244 TTATTGAAAAGAGTGATTTTAACATAAAAAAAGGTATTAAAAAACATATTG 2603_a11.seq
11017 TTATTGAAAAGAGTGATTTTAACATAAAAAAAGGTATTAAAAAACATATTG 18rs21_a11.seq
11249 TTATTGAAAAGAGTGATTTTAACATAAAAAAAGGTATTAAAAAACATATTG coh1_a11.seq
11249 TTATTGAAAAGAGTGATTTTAACATAAAAAAAGGTATTAAAAAACATATTG cjb111_a11.seq
11251 TTATTGAAAAGAGTGATTTTAACATAAAAAAAGGTATTAAAAAACATATTG nem316_a11.seq
11250 TTATTGAAAAGAGTGATTTTAACATAAAAAAAGGTATTAAAAAACATATTG a909_a11.seq

      ACGTGACCGTTTGTTTTGAAGTGGCTTGCGTAGACAAAAAAATAGATACG Majority
                11310       11320       11330       11340       11350
11294 ACGTGACCGTTTGTTTTGAAGTGGCTTGCGTAGACAAAAAAATAGATACG 2603_a11.seq
11067 ACGTGACCGTTTGTTTTGAAGTGGCTTGCGTAGACAAAAAAATAGATACG 18rs21_a11.seq
11299 ACGTGACCGTTTGTTTTGAAGTGGCTTGCGTAGACAAAAAAATAGATACG coh1_a11.seq
11299 ACGTGACCGTTTGTTTTGAAGTGGCTTGCGTAGACAAAAAAATAGATACG cjb111_a11.seq
11301 ACGTGACCGTTTGTTTTGAAGTGGCTTGCGTAGACAAAAAAATAGATACG nem316_a11.seq
11300 ACGTGACCGTTTGTTTTGAAGTGGCTTGCGTAGACAAAAAAATAGATACG a909_a11.seq

      TCAGATAAATTTCTGGCATTACGAGAACATTTTTAGAGTGTTCTCTTTTT Majority
                11360       11370       11380       11390       11400
11344 TCAGATAAATTTCTGGCATTACGAGAACATTTTTAGAGTGTTCTCTTTTT 2603_a11.seq
11117 TCAGATAAATTTCTGGCATTACGAGAACATTTTTAGAGTGTTCTCTTTTT 18rs21_a11.seq
11349 TCAGATAAATTTCTGGCATTACGAGAACATTTTTAGAGTGTTCTCTTTTT coh1_a11.seq
11349 TCAGATAAATTTCTGGCATTACGAGAACATTTCTAGAGTGTTCTCTTTTT cjb111_a11.seq
11351 TCAGATAAATTTCTGGCATTACGAGAACATTTTTAGAGTGTTCTCTTTTT nem316_a11.seq
11350 TCAGATAAATTTCTGGCATTACGAGAACATTTCTAGAGTGTTCTCTTTTT a909_a11.seq

      TTAGTTTACGGAGGAAAAATATATATGGAAAAACAGGATTCACGAGTTCT Majority
                11410       11420       11430       11440       11450
11394 TTAGTTTACGGAGGAAAAATATATATGGAAAAACAGGATTCACGAGTTCT 2603_a11.seq
11167 TTAGTTTACGGAGGAAAAATATATATGGAAAAACAGGATTCACGAGTTCT 18rs21_a11.seq
11399 TTAGTTTACGGAGGAAAAATATATATGGAAAAACAGGATTCACGAGTTCT coh1_a11.seq
11399 TTAGTTTACGGAGGAAAAATATATATGGAAAAACAGGATTCACGAGTTCT cjb111_a11.seq
11401 TTAGTTTACGGAGGAAAAATATATATGGAAAAACAGGATTCACGAGTTCT nem316_a11.seq
11400 TTAGTTTACGGAGGAAAAATATATATGGAAAAACAGGATTCACGAGTTCT a909_a11.seq

      CATCCATTGGGAGGGGAATTCTGGGGACAAGCTCATTGAACACCAAACCA Majority
                11460       11470       11480       11490       11500
11444 CATCCATTGGGAGGGGAATTCTGGGGACAAGCTCATTGAACACCAAACCA 2603_a11.seq
11217 CATCCATTGGGAGGGGAATTCTGGGGACAAGCTCATTGAACACCAAACCA 18rs21_a11.seq
11449 CATCCATTGGGAGGGGAATTCTGGGGACAAGCTCATTGAACACCAAACCA coh1_a11.seq
11449 CATCCATTGGGAGGGGAATTCTGGGGACAAGCTCATTGAACACCAAACCA cjb111_a11.seq
11451 CATCCATTGGGAGGGGAATTCTGGGGACAAGCTCATTGAACACCAAACCA nem316_a11.seq
11450 CATCCATTGGGAGGGGAATTCTGGGGACAAGCTCATTGAACACCAAACCA a909_a11.seq

      GCGCAACGGGGTGGTACTACCAAGTCGATCGTAGCTTTAGTCAACCAAAA Majority
                11510       11520       11530       11540       11550
11494 GCGCAACGGGGTGGTACTACCAAGTCGATCGTAGCTTTAGTCAACCAAAA 2603_a11.seq
11267 GCGCAACGGGGTGGTACTACCAAGTCGATCGTAGCTTTAGTCAACCAAAA 18rs21_a11.seq
11499 GCGCAATGGGGTGGTACTACCAAGTCGATCGTAGCTTTAGTCAACCAAAA coh1_a11.seq
11499 GCGCAACGGGGTGGTACTACCAAGTCGATCGTAGCTTTAGTCAACCAAAA cjb111_a11.seq
11501 GCGCAACGGGGTGGTACTACCAAGTCGATCGTAGCTTTAGTCAACCAAAA nem316_a11.seq
11500 GCGCAACGGGGTGGTACTACCAAGTCGATCGTAGCTTTAGTCAACCAAAA a909_a11.seq
```

FIGURE 18 AF

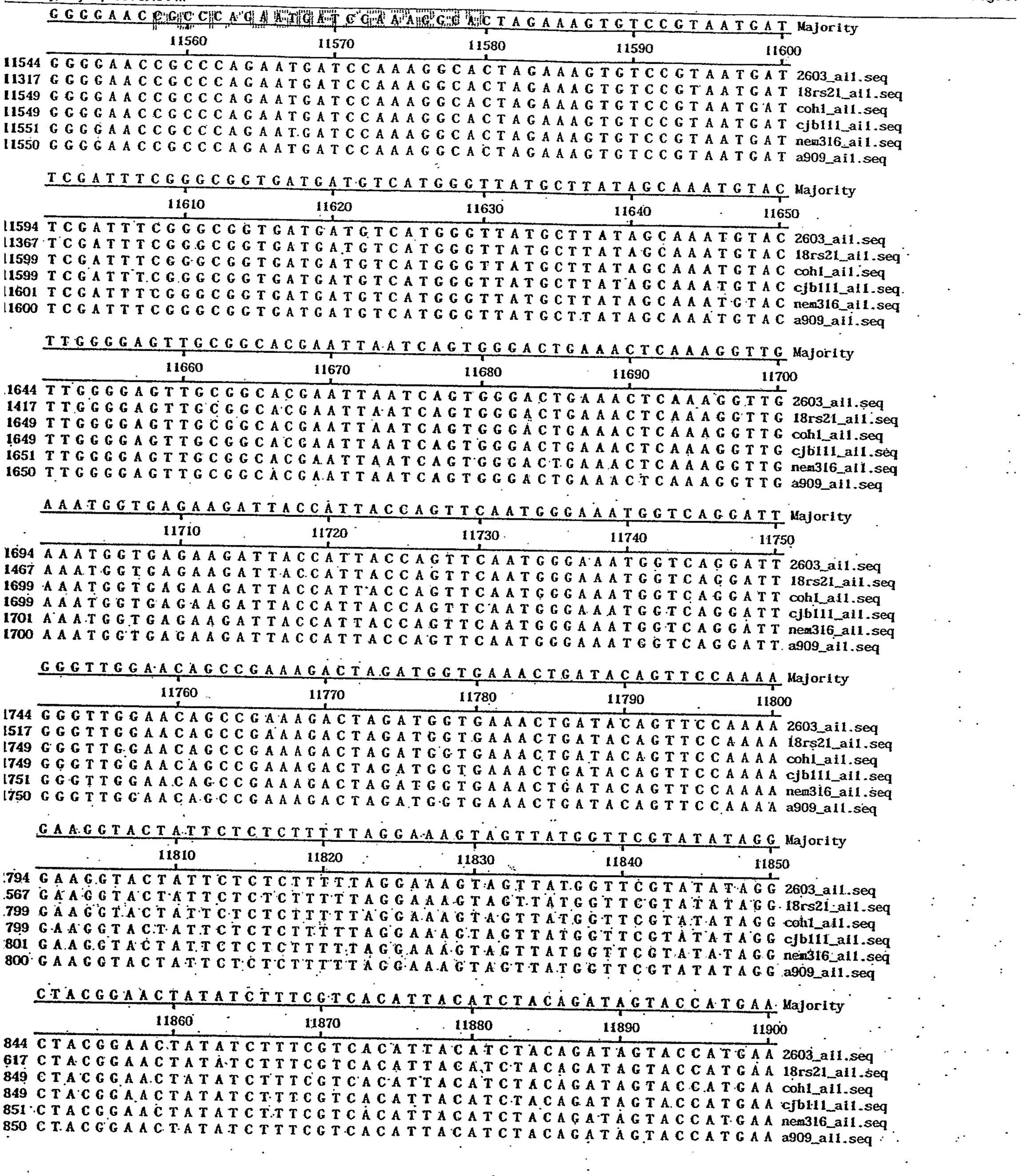

GGGCAACCGCCCAGAATGATCCAAAGGCACTAGAAAGTGTCCGTAATGAT Majority
11560 11570 11580 11590 11600
11544 GGGCAACCGCCCAGAATGATCCAAAGGCACTAGAAAGTGTCCGTAATGAT
11317 GGGCAACCGCCCAGAATGATCCAAAGGCACTAGAAAGTGTCCGTAATGAT 2603_all.seq
11549 GGGCAACCGCCCAGAATGATCCAAAGGCACTAGAAAGTGTCCGTAATGAT 18rs21_all.seq
11549 GGGCAACCGCCCAGAATGATCCAAAGGCACTAGAAAGTGTCCGTAATGAT coh1_all.seq
11551 GGGCAACCGCCCAGAATGATCCAAAGGCACTAGAAAGTGTCCGTAATGAT cjb111_all.seq
11550 GGGCAACCGCCCAGAATGATCCAAAGGCACTAGAAAGTGTCCGTAATGAT nem316_all.seq
11550 GGGCAACCGCCCAGAATGATCCAAAGGCACTAGAAAGTGTCCGTAATGAT a909_all.seq
TCGATTTCGGGCGGTGATGATGTCATGGGTTATGCTTATAGCAAATGTAC Majority
11610 11620 11630 11640 11650
11594 TCGATTTCGGGCGGTGATGATGTCATGGGTTATGCTTATAGCAAATGTAC
11367 TCGATTTCGGGCGGTGATGATGTCATGGGTTATGCTTATAGCAAATGTAC 2603_all.seq
11599 TCGATTTCGGGCGGTGATGATGTCATGGGTTATGCTTATAGCAAATGTAC 18rs21_all.seq
11599 TCGATTTCGGGCGGTGATGATGTCATGGGTTATGCTTATAGCAAATGTAC coh1_all.seq
11601 TCGATTTCGGGCGGTGATGATGTCATGGGTTATGCTTATAGCAAATGTAC cjb111_all.seq
11600 TCGATTTCGGGCGGTGATGATGTCATGGGTTATGCTTATAGCAAATGTAC nem316_all.seq
11600 TCGATTTCGGGCGGTGATGATGTCATGGGTTATGCTTATAGCAAATGTAC a909_all.seq
TTGGGGAGTTGCGGCACGAATTAATCAGTGGGACTGAAACTCAAAGGTTG Majority
11660 11670 11680 11690 11700
1644 TTGGGGAGTTGCGGCACGAATTAATCAGTGGGACTGAAACTCAAAGGTTG
1417 TTGGGGAGTTGCGGCACGAATTAATCAGTGGGACTGAAACTCAAAGGTTG 2603_all.seq
1649 TTGGGGAGTTGCGGCACGAATTAATCAGTGGGACTGAAACTCAAAGGTTG 18rs21_all.seq
1649 TTGGGGAGTTGCGGCACGAATTAATCAGTGGGACTGAAACTCAAAGGTTG coh1_all.seq
1651 TTGGGGAGTTGCGGCACGAATTAATCAGTGGGACTGAAACTCAAAGGTTG cjb111_all.seq
1650 TTGGGGAGTTGCGGCACGAATTAATCAGTGGGACTGAAACTCAAAGGTTG nem316_all.seq
1650 TTGGGGAGTTGCGGCACGAATTAATCAGTGGGACTGAAACTCAAAGGTTG a909_all.seq
AAATGGTGAGAAGATTACCATTACCAGTTCAATGGGAAATGGTCAGGATT Majority
11710 11720 11730 11740 11750
1694 AAATGGTGAGAAGATTACCATTACCAGTTCAATGGGAAATGGTCAGGATT
1467 AAATGGTGAGAAGATTACCATTACCAGTTCAATGGGAAATGGTCAGGATT 2603_all.seq
1699 AAATGGTGAGAAGATTACCATTACCAGTTCAATGGGAAATGGTCAGGATT 18rs21_all.seq
1699 AAATGGTGAGAAGATTACCATTACCAGTTCAATGGGAAATGGTCAGGATT coh1_all.seq
1701 AAATGGTGAGAAGATTACCATTACCAGTTCAATGGGAAATGGTCAGGATT cjb111_all.seq
1700 AAATGGTGAGAAGATTACCATTACCAGTTCAATGGGAAATGGTCAGGATT nem316_all.seq
1700 AAATGGTGAGAAGATTACCATTACCAGTTCAATGGGAAATGGTCAGGATT a909_all.seq
GGGTTGGAACAGCCGAAAGACTAGATGGTGAAACTGATACAGTTCCAAAA Majority
11760 11770 11780 11790 11800
1744 GGGTTGGAACAGCCGAAAGACTAGATGGTGAAACTGATACAGTTCCAAAA
1517 GGGTTGGAACAGCCGAAAGACTAGATGGTGAAACTGATACAGTTCCAAAA 2603_all.seq
1749 GGGTTGGAACAGCCGAAAGACTAGATGGTGAAACTGATACAGTTCCAAAA 18rs21_all.seq
1749 GGGTTGGAACAGCCGAAAGACTAGATGGTGAAACTGATACAGTTCCAAAA coh1_all.seq
1751 GGGTTGGAACAGCCGAAAGACTAGATGGTGAAACTGATACAGTTCCAAAA cjb111_all.seq
1750 GGGTTGGAACAGCCGAAAGACTAGATGGTGAAACTGATACAGTTCCAAAA nem316_all.seq
1750 GGGTTGGAACAGCCGAAAGACTAGATGGTGAAACTGATACAGTTCCAAAA a909_all.seq
GAAGGTACTATTCTCTCTTTTTAGGAAAGTAGTTATGGTTCGTATATAGG Majority
11810 11820 11830 11840 11850
1794 GAAGGTACTATTCTCTCTTTTTAGGAAAGTAGTTATGGTTCGTATATAGG
1567 GAAGGTACTATTCTCTCTTTTTAGGAAAGTAGTTATGGTTCGTATATAGG 2603_all.seq
1799 GAAGGTACTATTCTCTCTTTTTAGGAAAGTAGTTATGGTTCGTATATAGG 18rs21_all.seq
1799 GAAGGTACTATTCTCTCTTTTTAGGAAACTAGTTATGGTTCGTATATAGG coh1_all.seq
1801 GAAGGTACTATTCTCTCTTTTTAGGAAAGTAGTTATGGTTCGTATATAGG cjb111_all.seq
1800 GAAGGTACTATTCTCTCTTTTTAGGAAAGTAGTTATGGTTCGTATATAGG nem316_all.seq
1800 GAAGGTACTATTCTCTCTTTTTAGGAAAGTAGTTATGGTTCGTATATAGG a909_all.seq
CTACGGAACTATATCTTTCGTCACATTACATCTACAGATAGTACCATGAA Majority
11860 11870 11880 11890 11900
844 CTACGGAACTATATCTTTCGTCACATTACATCTACAGATAGTACCATGAA
617 CTACGGAACTATATCTTTCGTCACATTACATCTACAGATAGTACCATGAA 2603_all.seq
849 CTACGGAACTATATCTTTCGTCACATTACATCTACAGATAGTACCATGAA 18rs21_all.seq
849 CTACGGAACTATATCTTTCGTCACATTACATCTACAGATAGTACCATGAA coh1_all.seq
851 CTACGGAACTATATCTTTCGTCACATTACATCTACAGATAGTACCATGAA cjb111_all.seq
850 CTACGGAACTATATCTTTCGTCACATTACATCTACAGATAGTACCATGAA nem316_all.seq
850 CTACGGAACTATATCTTTCGTCACATTACATCTACAGATAGTACCATGAA a909_all.seq

FIGURE 18 AG

Alignment Report of AI-1_alignment, using J. Hein method with Weighted residue weight table.
Thursday, July 29, 2004 5:46 PM


```
      TTTTGCTTATATGACCAAGTAAAGTGAGGATATACTAACAAATGAAATAT Majority
               11910      11920      11930      11940      11950
11894 TTTTGCTTATATGACCAAGTAAAGTGAGGATATACTAACAAATGAAATAT 2603_ail.seq
11667 TTTTGCTTATATGACCAAGTAAAGTGAGGATATACTAACAAATGAAATAT 18rs21_ail.seq
11899 TTTTGCTTATATGACCAAGTAAAGTGAGGATATACTAACAAATGAAATAT coh1_ail.seq
11899 TTTTGCTTATATGACCAAGTAAAGTGAGGATATACTAACAAATGAAATAT cjb111_ail.seq
11901 TTTTGCTTATATGACCAAGTAAAGTGAGGATATACTAACAAATGAAATAT nem316_ail.seq
11900 TTTTGCTTATATGACCAAGTAAAGTGAGGATATACTAACAAATGAAATAT a909_ail.seq

      TTATTATCGTATTTGTCCATTTTATCGAAAAGTTTGCATATTATCATTAT Majority
               11960      11970      11980      11990      12000
11944 TTATTATCGTATTTGTCCATTTTATCGAAAAGTTTGCATATTATCATTAT 2603_ail.seq
11717 TTATTATCGTATTTGTCCATTTTATCGAAAAGTTTGCATATTATCATTAT 18rs21_ail.seq
11949 TTATTATCGTATTTGTCCATTTTATCGAAAAGTTTGCATATTATCATTAT coh1_ail.seq
11949 TTATTATCGTATTTGTCCATTTTATCGAAAAGTTTGCATATTATCATTAT cjb111_ail.seq
11951 TTATTATCGTATTTGTCCATTTTATCGAAAAGTTTGCATATTATCATTAT nem316_ail.seq
11950 TTATTATCGTATTTGTCCATTTTATCGAAAAGTTTGCATATTATCATTAT a909_ail.seq

      GTTTGATAAGATGCAAATATAATGATAGTAGGAGCTAAATATGGATATTT Majority
               12010      12020      12030      12040      12050
11994 GTTTGATAAGATGCAAATATAATGATAGTAGGAGCTAAATATGGATATTT 2603_ail.seq
11767 GTTTGATAAGATGCAAATATAATGATAGTAGGAGCTAAATATGGATATTT 18rs21_ail.seq
11999 GTTTGATAAGATGCAAATATAATGATAGTAGGAGCTAAATATGGATATTT coh1_ail.seq
11999 GTTTGATAAGATGCAAATATAATGATAGTAGGAGCTAAATATGGATATTT cjb111_ail.seq
12001 GTTTGATAAGATGCAAATATAATGATAGTAGGAGCTAAATATGGATATTT nem316_ail.seq
12000 GTTTGATAAGATGCAAATATAATGATAGTAGGAGCTAAATATGGATATTT a909_ail.seq

      AAAAAATCAAGAGTATCCTAAGTGCTTTCCATTTTGAAATTCAAATATAG Majority
               12060      12070      12080      12090      12100
12044 AAAAAATCAAGAGTATCCTAAGTGCTTTCCATTTTGAAATTCAAATATAG 2603_ail.seq
11817 AAAAAATCAAGAGTATCCTAAGTGCTTTCCATTTTGAAATTCAAATATAG 18rs21_ail.seq
12049 AAAAAATCAAGAGTATCCTAAGTGCTTTCCATTTTGAAATTCAAATATAG coh1_ail.seq
12049 AAAAAATCAAGAGTATCCTAAGTGCTTTCCATTTTGAAATTCAAATATAG cjb111_ail.seq
12051 AAAAAATCAAGAGTATCCTAAGTGCTTTCCATTTTGAAATTCAAATATAG nem316_ail.seq
12050 AAAAAATCAAGAGTATCCTAAGTGCTTTCCATTTTGAAATTCAAATATAG a909_ail.seq

      CTAATAGTTCTAGAACTTCTAATTGTTTTTTCGTCGACGATATGAATTTTC Majority
               12110      12120      12130      12140      12150
12094 CTAATAGTTCTAGAACTTCTAATTGTTTTTTCGTCGACGATATGAATTTTC 2603_ail.seq
11867 CTAATAGTTCTAGAACTTCTAATTGTTTTTTCGTCGACGATATGAATTTTC 18rs21_ail.seq
12099 CTAATAGTTCTAGAACTTCTAATTGTTTTTTCGTCGACGATATGAATTTTC coh1_ail.seq
12099 CTAATAGTTCTAGAACTTCTAATTGTTTTTTCGTCGACGATATGAATTTTC cjb111_ail.seq
12101 CTAATAGTTCTAGAACTTCTAATTGTTTTTTCGTCGACGATATGAATTTTC nem316_ail.seq
12100 CTAATAGTTCTAGAACTTCTAATTGTTTTTTCGTCGACGATATGAATTTTC a909_ail.seq

      AATCTTAACTGTTAGGATTCCACCTCCCTTTGGTTAAAGAAAAAAGGTCA Majority
               12160      12170      12180      12190      12200
12144 AATCTTAACTGTTAGGATTCCACCTCCCTTTGGTTAAAGAAAAAAGGTCA 2603_ail.seq
11917 AATCTTAACTGTTAGGATTCCACCTCCCTTTGGTTAAAGAAAAAAGGTCA 18rs21_ail.seq
12149 AATCTTAACTGTTAGGATTCCACCTCCCTTTGGTTAAAGAAAAAAGGTCA coh1_ail.seq
12149 AATCTTAACTGTTAGGATTCCACCTCCCTTTGGTTAAAGAAAAAAGGTCA cjb111_ail.seq
12151 AATCTTAACTGTTAGGATTCCACCTCCCTTTGGTTAAAGAAAAAAGGTCA nem316_ail.seq
12150 AATCTTAACTGTTAGGATTCCACCTCCCTTTGGTTAAAGAAAAAAGGTCA a909_ail.seq

      GGTCGTTTAGATAACTTTGTCAAACAAGCTCAAGCTATCTAAAAATAGTT Majority
               12210      12220      12230      12240      12250
12194 GGTCGTTTAGATAACTTTGTCAAACAAGCTCAAGCTATCTAAAAATAGTT 2603_ail.seq
11967 GGTCGTTTAGATAACTTTGTCAAACAAGCTCAAGCTATCTAAAAATAGTT 18rs21_ail.seq
12199 GGTCGTTTAGATAACTTTGTCAAACAAGCTCAAGCTATCTAAAAATAGTT coh1_ail.seq
12199 GGTCGTTTAGATAACTTTGTCAAACAAGCTCAAGCTATCTAAAAATAGTT cjb111_ail.seq
12201 GGTCGTTTAGATAACTTTGTCAAACAAGCTCAAGCTATCTAAAAATAGTT nem316_ail.seq
12200 GGTCGTTTAGATAACTTTGTCAAACAAGCTCAAGCTATCTAAAAATAGTT a909_ail.seq
```

FIGURE 18 AH

```
     TGAAATGGCATTACTCTAGTTTTTAATAAGCTATCTGATGAGCAGAAGG Majority
                12260      12270      12280      12290      12300
12244 TGAAATGGCATTACTCTAGTTTTTAATAAGCTATCTGATGAGCAGAAGG 2603_ail.seq
12017 TGAAATGGCATTACTCTAGTTTTTAATAAGCTATCTGATGAGCAGAAGG 18rs21_ail.seq
12249 TGAAATGGCATTACTCTAGTTTTTAATAAGCTATCTGATGAGCAGAAGG coh1_ail.seq
12249 TGAAATGGCATTACTCTAGTTTTTAATAAGCTATCTGATGAGCAGAAGG cjb111_ail.seq
12251 TGAAATGGCATTACTCTAGTTTTTAATAAGCTATCTGATGAGCAGAAGG nem316_ail.seq
12250 TGAAATGGCATTACTCTAGTTTTTAATAAGCTATCTGATGAGCAGAAGG a909_ail.seq

     AGAAGTTAATGCATGTTGGGAAGTCTTATTTTGACTATCAAGAAAATGCT Majority
                12310      12320      12330      12340      12350
12294 AGAAGTTAATGCATGTTGGGAAGTCTTATTTTGACTATCAAGAAAATGCT 2603_ail.seq
12067 AGAAGTTAATGCATGTTGGGAAGTCTTATTTTGACTATCAAGAAAATGCT 18rs21_ail.seq
12299 AGAAGTTAATGCATGTTGGGAAGTCTTATTTTGACTATCAAGAAAATGCT coh1_ail.seq
12299 AGAAGTTAATGCATGTTGGGAAGTCTTATTTTGACTATCAAGAAAATGCT cjb111_ail.seq
12301 AGAAGTTAATGCATGTTGGGAAGTCTTATTTTGACTATCAAGAAAATGCT nem316_ail.seq
12300 AGAAGTTAATGCATGTTGGGAAGTCTTATTTTGACTATCAAGAAAATGCT a909_ail.seq

     CTTATCCCACAATTAGGTTTTCTATATTCTAAATTAACTAAAAAAATTGA Majority
                12360      12370      12380      12390      12400
2344 CTTATCCCACAATTAGGTTTTCTATATTCTAAATTAACTAAAAAAATTGA 2603_ail.seq
2117 CTTATCCCACAATTAGGTTTTCTATATTCTAAATTAACTAAAAAAATTGA 18rs21_ail.seq
2349 CTTATCCCACAATTAGGTTTTCTATATTCTAAATTAACTAAAAAAATTGA coh1_ail.seq
2349 CTTATCCCACAATTAGGTTTTCTATATTCTAAATTAACTAAAAAAATTGA cjb111_ail.seq
2351 CTTATCCCACAATTAGGTTTTCTATATTCTAAATTAACTAAAAAAATTGA nem316_ail.seq
2350 CTTATCCCACAATTAGGTTTTCTATATTCTAAATTAACTAAAAAAATTGA a909_ail.seq

     ACTTGATAATCGGTTGTCTCCGACTGAACAAAAGTTATTGATTACCTTAT Majority
                12410      12420      12430      12440      12450
2394 ACTTGATAATCGGTTGTCTCCGACTGAAAAAAAGTTATTGATTACCTTAT 2603_ail.seq
2167 ACTTGATAATCGGTTGTCTCCGACTGAACAAAAGTTATTGATTACCTTAT 18rs21_ail.seq
2399 ACTTGATAATCGGTTGTCTCCGACTGAACAAAAGTTATTGATTACCTTAT coh1_ail.seq
2399 ACTTGATAATCGGTTGTCTCCGACTGAACAAAAGTTATTGATTACCTTAT cjb111_ail.seq
2401 ACTTGATAATCGGTTGTCTCCGACTGAACAAAAGTTATTGATTACCTTAT nem316_ail.seq
2400 ACTTGATAATCGGTTGTCTCCGACTGAACAAAAGTTATTGATTACCTTAT a909_ail.seq

     TATTACATACTAAAGGTTTAATCATTGATATGTAAGAAGTAAGTCAGCTA Majority
                12460      12470      12480      12490      12500
2444 TATTACATACTAAAGGTTTAATCATTGATATGTAAGAAGTAAGTCAGCTA 2603_ail.seq
2217 TATTACATACTAAAGGTTTAATCATTGATATGTAAGAAGTAAGTCAGCTA 18rs21_ail.seq
2449 TATTACATACTAAAGGTTTAATCATTGATATGTAAGAAGTAAGTCAGCTA coh1_ail.seq
2449 TATTACATACTAAAGGTTTAATCATTGATATGTAAGAAGTAAGTCAGCTA cjb111_ail.seq
2451 TATTACATACTAAAGGTTTAATCATTGATATGTAAGAAGTAAGTCAGCTA nem316_ail.seq
2450 TATTACATACTAAAGGTTTAATCATTGATATGTAAGAAGTAAGTCAGCTA a909_ail.seq

     ACCGATCTTTCTATTCTAAAACTTATATTGTTGCTTTAGAAATTTTAAAG Majority
                12510      12520      12530      12540      12550
2494 ACCGATCTTTCTATTCTAAAACTTATATTGTTGCTTTAGAAATTTTAAAG 2603_ail.seq
2267 ACCGATCTTTCTATTCTAAAACTTATATTGTTGCTTTAGAAATTTTAAAG 18rs21_ail.seq
2499 ACCGATCTTTCTATTCTAAAACTTATATTGTTGCTTTAGAAATTTTAAAG coh1_ail.seq
2499 ACCGATCTTTCTATTCTAAAACTTATATTGTTGCTTTAGAAATTTTAAAG cjb111_ail.seq
2501 ACCGATCTTTCTATTCTAAAACTTATATTGTTGCTTTAGAAATTTTAAAG nem316_ail.seq
2500 ACCGATCTTTCTATTCTAAAACTTATATTGTTGCTTTAGAAATTTTAAAG a909_ail.seq

     AGCGTGGATGGCTTCATAATAAACAGAAATCTTACCAATTTGCGAAGCCA Majority
                12560      12570      12580      12590      12600
2544 AGCGTGGATGGCTTCATAATAAACAGAAATCTTACCAATTTGCGAAGCCA 2603_ail.seq
2317 AGCGTGGATGGCTTCATAATAAACAGAAATCTTACCAATTTGCGAAGCCA 18rs21_ail.seq
2549 AGCGTGGATGGCTTCATAATAAACAGAAATCTTACCAATTTGCGAAGCCA coh1_ail.seq
2549 AGCGTGGATGGCTTCATAATAAACAGAAATCTTACCAATTTGCGAAGCCA cjb111_ail.seq
2551 AGCGTGGATGGCTTCATAATAAACAGAAATCTTACCAATTTGCGAAGCCA nem316_ail.seq
2550 AGCGTGGATGGCTTCATAATAAACAGAAATCTTACCAATTTGCGAAGCCA a909_ail.seq
```

FIGURE 18 AI

```
      A A A A A T A T G A T A T T T G A A G A G T C C A A A G A T C T A A T A G A T A G T C C A G T T A G  Majority
                12610       12620       12630       12640       12650
2594  A A A A A T A T G A T A T T T G A A G A G T C C A A A G A T C T A A T A G A T A G T C C A G T T A G  2603_all.seq
2367  A A A A A T A T G A T A T T T G A A G A G T C C A A A G A T C T A A T A G A T A G T C C A G T T A G  18rs21_all.seq
2599  A A A A A T A T G A T A T T T G A A G A G T C C A A A G A T C T A A T A G A T A G T C C A G T T A G  coh1_all.seq
2599  A A A A A T A T G A T A T T T G A A G A G T C C A A A G A T C T A A T A G A T A G T C C A G T T A G  cjb111_all.seq
2601  A A A A A T A T G A T A T T T G A A G A G T C C A A A G A T C T A A T A G A T A G T C C A G T T A G  nem316_all.seq
2600  A A A A A T A T G A T A T T T G A A G A G T C C A A A G A T C T A A T A G A T A G T C C A G T T A G  a909_all.seq

      A G A A G C G T T G A T T A T A A G T G A T A A G G A T T T T C A A A A A T T A A A A C A A G A G C  Majority
                12660       12670       12680       12690       12700
2644  A G A A G C G T T G A T T A T A A G T G A T A A G G A T T T T C A A A A A T T A A A A C A A G A G C  2603_all.seq
2417  A G A A G C G T T G A T T A T A A G T G A T A A G G A T T T T C A A A A A T T A A A A C A A G A G C  18rs21_all.seq
2649  A G A A G C G T T G A T T A T A A G T G A T A A G G A T T T T C A A A A A T T A A A A C A A G A G C  coh1_all.seq
2649  A G A A G C G T T G A T T A T A A G T G A T A A G G A T T T T C A A A A A T T A A A A C A A G A G C  cjb111_all.seq
2651  A G A A G C G T T G A T T A T A A G T G A T A A G G A T T T T C A A A A A T T A A A A C A A G A G C  nem316_all.seq
2650  A G A A G C G T T G A T T A T A A G T G A T A A G G A T T T T C A A A A A T T A A A A C A A G A G C  a909_all.seq

      T A T T A T T T T A A C C G A C T T A T T T T A A A G A C T T A T C A T A T C T A G G C T T G C T T  Majority
                12710       12720       12730       12740       12750
2694  T A T T A T T T T A A C C G A C T T A T T T T A A A G A C T T A T C A T A T C T A G G C T T G C T T  2603_all.seq
2467  T A T T A T T T T A A C C G A C T T A T T T T A A A G A C T T A T C A T A T C T A G G C T T G C T T  18rs21_all.seq
2699  T A T T A T T T T A A C C G A C T T A T T T T A A A G A C T T A T C A T A T C T A G G C T T G C T T  coh1_all.seq
2699  T A T T A T T T T A A C C G A C T T A T T T T A A A G A C T T A T C A T A T C T A G G C T T G C T T  cjb111_all.seq
2701  T A T T A T T T T A A C C G A C T T A T T T T A A A G A C T T A T C A T A T C T A G G C T T G C T T  nem316_all.seq
2700  T A T T A T T T T A A C C G A C T T A T T T T A A A G A C T T A T C A T A T C T A G G C T T G C T T  a909_all.seq

      G A T G A T T C G G A A A A A T A C G G A G A C T A T A C T A T T T C A A G G A A A A G A T A C A A  Majority
                12760       12770       12780       12790       12800
2744  G A T G A T T C G G A A A A A T A C G G A G A C T A T A C T A T T T C A A G G A A A A G A T A C A A  2603_all.seq
2517  G A T G A T T C G G A A A A A T A C G G A G A C T A T A C T A T T T C A A G G A A A A G A T A C A A  18rs21_all.seq
2749  G A T G A T T C G G A A A A A T A C G G A G A C T A T A C T A T T T C A A G G A A A A G A T A C A A  coh1_all.seq
2749  G A T G A T T C G G A A A A A T A C G G A G A C T A T A C T A T T T C A A G G A A A A G A T A C A A  cjb111_all.seq
2751  G A T G A T T C G G A A A A A T A C G G A G A C T A T A C T A T T T C A A G G A A A A G A T A C A A  nem316_all.seq
2750  G A T G A T T C G G A A A A A T A C G G A G A C T A T A C T A T T T C A A G G A A A A G A T A C A A  a909_all.seq

      A A G T T T C G A A T C A A G T C T T C A A C T A T A C A T C C T T C A A A G T C A T C G G C T A G  Majority
                12810       12820       12830       12840       12850
2794  A A G T T T C G A A T C A A G T C T T C A A C T A T A C A T C C T T C A A A G T C A T C G G C T A G  2603_all.seq
2567  A A G T T T C G A A T C A A G T C T T C A A C T A T A C A T C C T T C A A A G T C A T C G G C T A G  18rs21_all.seq
2799  A A G T T T C G A A T C A A G T C T T C A A C T A T A C A T C C T T C A A A G T C A T C G G C T A G  coh1_all.seq
2799  A A G T T T C G A A T C A A G T C T T C A A C T A T A C A T C C T T C A A A G T C A T C G G C T A G  cjb111_all.seq
2801  A A G T T T C G A A T C A A G T C T T C A A C T A T A C A T C C T T C A A A G T C A T C G G C T A G  nem316_all.seq
2800  A A G T T T C G A A T C A A G T C T T C A A C T A T A C A T C C T T C A A A G T C A T C G G C T A G  a909_all.seq

      A G A T T T G G A A T T A T G A A C C A A T C C C T T T G A T T A C T A G A A A A A T A A A T A G C  Majority
                12860       12870       12880       12890       12900
2844  A G A T T T G G A A T T A T G A A C C A A T C C C T T T G A T T A C T A G A A A A A T A A A T A G C  2603_all.seq
2617  A G A T T T G G A A T T A T G A A C C A A T C C C T T T G A T T A C T A G A A A A A T A A A T A G C  18rs21_all.seq
2849  A G A T T T G G A A T T A T G A A C C A A T C C C T T T G A T T A C T A G A A A A A T A A A T A G C  coh1_all.seq
2849  A G A T T T G G A A T T A T G A A C C A A T C C C T T T G A T T A C T A G A A A A A T A A A T A G C  cjb111_all.seq
2851  A G A T T T G G A A T T A T G A A C C A A T C C C T T T G A T T A C T A G A A A A A T A A A T A G C  nem316_all.seq
2850  A G A T T T G G A A T T A T G A A C C A A T C C C T T T G A T T A C T A G A A A A A T A A A T A G C  a909_all.seq

      T T G G A G A G G C T A A C G T G A C A C T G G T T G A T C C A A T C T C G C T T T A T T T A A C A  Majority
                12910       12920       12930       12940       12950
2894  T T G G A G A G G C T A A C G T G A C A C T G G T T G A T C C A A T C T C G C T T T A T T T A A C A  2603_all.seq
2667  T T G G A G A G G C T A A C G T G A C A C T G G T T G A T C C A A T C T C G C T T T A T T T A A C A  18rs21_all.seq
2899  T T G G A G A G G C T A A C G T G A C A C T G G T T G A T C C A A T C T C G C T T T A T T T A A C A  coh1_all.seq
2899  T T G G A G A G G C T A A C G T G A C A C T G G T T G A T C C A A T C T C G C T T T A T T T A A C A  cjb111_all.seq
2901  T T G G A G A G G C T A A C G T G A C A C T G G T T G A T C C A A T C T C G C T T T A T T T A A C A  nem316_all.seq
2900  T T G G A G A G G C T A A C G T G A C A C T G G T T G A T C C A A T C T C G C T T T A T T T A A C A  a909_all.seq
```

FIGURE 18 AJ

Alignment Report of AI-1_alignment, using J. Hein method with Weighted residue weight table.
Thursday, July 29, 2004 5:46 PM


```
     C T A A G A A T G A T G A A G A C C C T C G T A T T G A A G A A G A A G T T G A G C A G C T A G A A  Majority
                    12960         12970         12980         12990         13000
12944 C T A A G A A T G A T G A A G A C C C T C G T A T T G A A G A A G A A G T T G A G C A G C T A G A A  2603_all.seq
12717 C T A A G A A T G A T G A A G A C C C T C G T A T T G A A G A A G A A G T T G A G C A G C T A G A A  18rs21_all.seq
12949 C T A A G A A T G A T G A A G A C C C T C G T A T T G A A G A A G A A G T T G A G C A G C T A G A A  coh1_all.seq
12949 C T A A G A A T G A T G A A G A C C C T C G T A T T G A A G A A G A A G T T G A G C A G C T A G A A  cjb111_all.seq
12951 C T A A G A A T G A T G A A G A C C C T C G T A T T G A A G A A G A A G T T G A G C A G C T A G A .  nem316_all.seq
12950 C T A A G A A T G A T G A A G A C C C T C G T A T T G A A G A A G A A G T T G A G C A G C T A G A A  a909_all.seq

     G A T A A G A T                                                                                Majority

12994 G A T A A G A T                                                                               2603_all.seq
12767 G A T A A G A T                                                                               18rs21_all.seq
12999 G A T A A G A T                                                                               coh1_all.seq
12999 G A                                                                                           cjb111_all.seq
13000 . A                                                                                           nem316_all.seq
13000 G A T A A G A T                                                                               a909_all.seq
```

Decoration 'Decoration #1': Shade (with solid black) residues that differ from the Consensus.

FIGURE 18 AK

```
GGCCTTGTTCCGATGTTGATCCGATAACTCCTGGCTCATTAATAGCCTG Majority
         10        20        30        40        50
1   GGCCTTGTTCCGATGTTGATCCGATAACTCCTGGCTCATTAATAGCCTG 2603_a12.seq
1   GGCCTTGTTCCGATGTTGATCCGATAACTCCTGGCTCATTAATAGCCTG 18rs21_a12.seq
1   GGCCTTGTTCCGATGTTGATCCCGATAACTCCTGGCTCATTAATAGCCTG 515_a12.seq
1   GGCCTTGTTCCGATGTTGATCCCGATAACTCCTGGCTCATTAATAGCCTG cjb111_a12.seq
1   GGCCTTGTTCCGATGTTGATCCCGATAACTCCTGGCTCATTAATAGCCTG h36b_a12.seq

TTCGTAACGCTCTTTAATTATCTCTAACTTAGCATGGGTATTGGTAAAAT Majority
         60        70        80        90       100
51  TTCGTAACGCTCTTTAATTATCTCTAACTTAGCATGGGTATTGGTAAAAT 2603_a12.seq
51  TTCGTAACGCTCTTTAATTATCTCTAACTTAGCATGGGTATTGGTAAAAT 18rs21_a12.seq
51  TTCGTAACGCTCTTTAATTATCTCTAACTTAGCATGGGTATTGGTAAAAT 515_a12.seq
51  TTCGTAACGCTCTTTAATTATCTCTAACTTAGCATGGGTATTGGTAAAAT cjb111_a12.seq
51  TTCGTAACGCTCTTTAATTATCTCTAACTTAGCATGGGTATTGGTAAAAT h36b_a12.seq

TTTGAAAATAGACTAAGTATTTATTAACCTCAGGCCACTTTCTATGCATG Majority
        110       120       130       140       150
101 TTTGAAAATAGACTAAGTATTTATTAACCTCAGGCCACTTTCTATGCATG 2603_a12.seq
101 TTTGAAAATAGACTAAGTATTTATTAACCTCAGGCCACTTTCTATGCATG 18rs21_a12.seq
101 TTTGAAAATAGACTAAGTATTTATTAACTTCAGGCCACTTTCTATGCATG 515_a12.seq
101 TTTGAAAATAGACTAAGTATTTATTAACCTCAGGCCACTTTCTATGCATG cjb111_a12.seq
101 TTTGAAAATAGACTAAGTATTTATTAACCTCAGGCCACTTTCTATGCATG h36b_a12.seq

AAATCAATTTCTTTATAGAATTGTTCACGAATAGGAGCTTCTGGAGCAAC Majority
        160       170       180       190       200
151 AAATCAATTTCTTTATAGAATTGTTCACGAATAGGAGCTTCTGGAGCAAC 2603_a12.seq
151 AAATCAATTTCTTTATAGAATTGTTCACGAATAGGAGCTTCTGGAGCAAC 18rs21_a12.seq
151 AAATCAATTTCTTTATAGAATTGTTCACGAATAGGAGCTTCTGGAGCAAC 515_a12.seq
151 AAATCAATTTCTTTATAGAATTGTTCACGAATAGGAGCTTCTGGAGCAAC cjb111_a12.seq
151 AAATCAATTTCTTTATAGAATTGTTCACGAATAGGAGCTTCTGGAGCAAC h36b_a12.seq

TATAGCATCCCCTGAACCAGAAACTGTGCAAAAAGTGCACCCTCCTCTAG Majority
        210       220       230       240       250
201 TATAGCATCCCCTGAACCAGAAACTGTGCAAAAAGTGCACCCTCCTCTAG 2603_a12.seq
201 TATAGCATCCCCTGAACCAGAAACTGTGCAAAAAAGTGCACCCTCCTCTAG 18rs21_a12.seq
201 TATAGCATCCCCTGAACCAGAAACTGTGCAAAAAAGTGCACCCTCCTCTAG 515_a12.seq
201 TATAGCATCCCCTGAACCAGAAACTGTGCAAAAAAGTGCACCCTCCTCTAG cjb111_a12.seq
201 TATAGCATCCCCTGAACCAGAAACTGTGCAAAAAAGTGCACCCTCCTCTAG h36b_a12.seq

CAACTGTTCCGTCTCTGTTAGGACAGTCAAAACCAGCATCTATAGGTAAT Majority
        260       270       280       290       300
251 CAACTGTTCCGTCTCTGTTAGGACAGTCAAAACCAGCATCTATAGGTAAT 2603_a12.seq
251 CAACTGTTCCGTCTCTGTTAGGACAGTCAAAACCAGCATCTATAGGTAAT 18rs21_a12.seq
251 CAACTGTTCCGTCTCTGTTAGGACAGTCAAAACCAGCATCTATAGGTAAT 515_a12.seq
251 CAACTGTTCCGTCTCTGTTAGGACAGTCAAAACCAGCATCTATAGGTAAT cjb111_a12.seq
251 CAACTGTTCCGTCTCTGTTAGGACAGTCAAAACCAGCATCTATAGGTAAT h36b_a12.seq

TTAAATATTTTTTCTCCAAAGAGTTCTCGATAATAATCATTAATCGCACG Majority
        310       320       330       340       350
301 TTAAATATTTTTTCTCCAAAGAGTTCTCGATAATAATCATTAATCGCACG 2603_a12.seq
301 TTAAATATTTTTTCTCCAAAGAGTTCTCGATAATAATCATTAATCGCACG 18rs21_a12.seq
301 TTAAATATTTTTTCTCCAAAGAGTTCTCGATAATAATCATTAATCGCACG 515_a12.seq
301 TTAAATATTTTTTCTCCAAAGAGTTCTCGATAATAATCATTAATCGCACG cjb111_a12.seq
301 TTAAATATTTTTTCTCCAAAGAGTTCTCGATAATAATCATTAATCGCACG h36b_a12.seq

ATAACGTTTTTTTCATAGGATAATTGTATCACAATTTTAACTAAAATAACC Majority
        360       370       380       390       400
351 ATAACGTTTTTTTCATAGGATAATTGTATCACAATTTTAACTAAAATAACC 2603_a12.seq
351 ATAACGTTTTTTTCATAGGATAATTGTATCACAATTTTAACTAAAATAACC 18rs21_a12.seq
351 ATAACGTTTTTTTCATAGGATAATTGTATCACAATTTTAACTAAAATAACC 515_a12.seq
351 ATAACGTTTTTTTCATAGGATAATTGTATCACAATTTTAACTAAAATAACC cjb111_a12.seq
351 ATAACGTTTTTTTCATAGGATAATTGTATCACAATTTTAACTAAAATAACC h36b_a12.seq

TCACTACTACAATAAAACTAAAAAAGATTGGAACGTCAGTTAGTCCCAAT Majority
        410       420       430       440       450
401 TCACTACTACAATAAAACTAAAAAAGATTGGAACGTCAGTTAGTCCCAAT 2603_a12.seq
401 TCACTACTACAATAAAACTAAAAAAGATTGGAACGTCAGTTAGTCCCAAT 18rs21_a12.seq
401 TCACTACTACAATAAAACTAAAAAAGATTGGAACATCAGTTAGTCCCAAT 515_a12.seq
401 TCACTACTACAATAAAACTAAAAAAGATTGGAACTAACTGACGTTCCAAT cjb111_a12.seq
401 TCACTACTACAATAAAACTAAAAAAGATTGGAACGTCAGTTAGTCCCAAT h36b_a12.seq
```

Figure 19

```
                                                
```

```
TTACAGTATTACCAATCACAGTGATTAACTTGAAAAATCTTGTAGAAAGA Majority
          910       920       930       940       950
901  TTACAGTATTACCAATCACAGTGATTAACTTGAAAAATCTTGTAGAAAGA
901  TTACAGTATTACCAATCACAGTGATTAACTTGAAAAATCTTGTAGAAAGA 2603_a12.seq
901  TTACAGTATTACCAATCACAGTGATTAACTTGAAAAATCTTGTAGAAAGA 18rs21_a12.seq
901  TTACAGTATTACCAATCACAGTGATTAACTTGAAAAATCTTGTAGAAAGA 515_a12.seq
901  TTACAGTATTACCAATCACAGTGATTAACTTGAAAAATCTTGTAGAAAGA cjb111_a12.seq
901  TTACAGTATTACCAATCACAGTGATTAACTTGAAAAATCTTGTAGAAAGA h36b_a12.seq

TTTGGCAACTGTCCTCTAACACTTTCTTGAATAGTTTGGTCAAATGAAAT Majority
          960       970       980       990      1000
951  TTTGGCAACTGTCCTCTAACACTTTCTTGAATAGTTTGGTCAAATGAAAT
951  TTTGGCAACTGTCCTCTAACACTTTCTTGAATAGTTTGGTCAAATGAAAT 2603_a12.seq
951  TTTGGCAACTGTCCTCTAACACTTTCTTGAATAGTTTGGTCAAATGAAAT 18rs21_a12.seq
951  TTTGGCAACTGTCCTCTAACACTTTCTTGAATAGTTTGGTCAAATGCAAT 515_a12.seq
951  TTTGGCAACTGTCCTCTAACACTTTCTTGAATAGTTTGGTCAAATGAAAT cjb111_a12.seq
951  TTTGGCAACTGTCCTCTAACACTTTCTTGAATAGTTTGGTCAAATGAAAT h36b_a12.seq

TACAGTGTCGGGCCAATATTTGATGACCAATCCTAAACTGAAAAATAAGA Majority
         1010      1020      1030      1040      1050
1001 TACAGTGTCGGGCCAATATTTGATGACCAATCCTAAACTGAAAAATAAGA
1001 TACAGTGTCGGGCCAATATTTGATGACCAATCCTAAACTGAAAAATAAGA 2603_a12.seq
1001 TACAGTGTCGGGCCAATATTTGATGACCAATCCTAAACTGAAAAATAAGA 18rs21_a12.seq
1001 TACAGTGTCGGGCCAATATTTGATGACCAATCCTAAACTGAAAAATAAGA 515_a12.seq
1001 TACAGTGTCGGGCCAATATTTGATGACCAATCCTAAACTGAAAAATAAGA cjb111_a12.seq
1001 TACAGTGTCGGGCCAATATTTGATGACCAATCCTAAACTGAAAAATAAGA h36b_a12.seq

TAATAGCAATAAATGCTTGAATAAGTTTACTATTTTGACGAGATAACATT Majority
         1060      1070      1080      1090      1100
1051 TAATAGCAATAAATGCTTGAATAAGTTTACTATTTTGACGAGATAACATT
1051 TAATAGCAATAAATGCTTGAATAAGTTTACTATTTTGACGAGATAACATT 2603_a12.seq
1051 TAATAGCAATAAATGCTTGAATAAGTTTACTATTTTGACGAGATAACATT 18rs21_a12.seq
1051 TAATAGCAATAAATGCTTGAATAAGTTTACTATTTTGACGAGATAACATT 515_a12.seq
1051 TAATAGCAATAAATGCTTGAATAAGTTTACTATTTTGACGAGATAACATT cjb111_a12.seq
1051 TAATAGCAATAAATGCTTGAATAAGTTTACTATTTTGACGAGATAACATT h36b_a12.seq

AGTCTTTTTATATCTTTCTAATATTGGCAAACAAGCCACGTAAGTTAGAT Majority
         1110      1120      1130      1140      1150
1101 AGTCTTTTTATATCTTTCTAATATTGGCAAACAAGCCACGTAAGTTAGAT
1101 AGTCTTTTTATATCTTTCTAATATTGGCAAACAAGCCACGTAAGTTAGAT 2603_a12.seq
1101 AGTCTTTTTATATCTTTCTAATATTGGCAAACAAGCCACGTAAGTTAGAT 18rs21_a12.seq
1101 AGTCTTTTTATATCTTTCTAATATTGGCAAACAAGCCACGTAAGTTAGAT 515_a12.seq
1101 AGTCTTTTTATATCTTTCTAATATTGGCAAACAAGCCACGTAAGTTAGAT cjb111_a12.seq
1101 AGTCTTTTTATATCTTTCTAATATTGGCAAACAAGCCACGTAAGTTAGAT h36b_a12.seq

AGAAAACAATCGAAATTAAAATTCCCTCAACGATATTAAATGGAATAACC Majority
         1160      1170      1180      1190      1200
1151 AGAAAACAATCGAAATTAAAATTCCCTCAACGATATTAAATGGAATAACC
1151 AGAAAACAATCGAAATTAAAATTCCCTCAACGATATTAAATGGAATAACC 2603_a12.seq
1151 AGAAAACAATCGAAATTAAAATTCCCTCAACGATATTAAATGGAATAACC 18rs21_a12.seq
1151 AGAAAACAATCGAAATTAAAATTCCCTCAACGATATTAAATGGAATAACC 515_a12.seq
1151 AGAAAACAATCGAAATTAAAATTCCCTCAACGATATTAAATGGAATAACC cjb111_a12.seq
1151 AGAAAACAATCGAAATTAAAATTCCCTCAACGATATTAAATGGAATAACC h36b_a12.seq

ATTGTTAAAAGGTAATTGCCTACACCAATAAATGTTCTGATATCAAAGTT Majority
         1210      1220      1230      1240      1250
1201 ATTGTTAAAAGGTAATTGCCTACACCAATAAATGTTCTGATATCAAAGTT
1201 ATTGTTAAAAGGTAATTGCCTACACCAATAAATGTTCTGATATCAAAGTT 2603_a12.seq
1201 ATTGTTAAAAGGTAATTGCCTACACCAATAAATGTTCTGATATCAAAGTT 18rs21_a12.seq
1201 ATTGTTAAAAGGTAATTGCCTACACCAATAAATGTTCTGATATCAAAGTT 515_a12.seq
1201 ATTGTTAAAAGGTAATTGCCTACACCAATAAATGTTCTGATATCAAAGTT cjb111_a12.seq
1201 ATTGTTAAAAGGTAATTGCCTACACCAATAAATGTTCTGATATCAAAGTT h36b_a12.seq

AGCAAATATAGCATACAAAGGAATCGCAAAGACATAGTTGAGAGCTACCA Majority
         1260      1270      1280      1290      1300
1251 AGCAAATATAGCATACAAAGGAATCGCAAAGACATAGTTGAGAGCTACCA
1251 AGCAAATATAGCATACAAAGGAATCGCAAAGACATAGTTGAGAGCTACCA 2603_a12.seq
1251 AGCAAATATAGCATACAAAGGAATCGCAAAGACATAGTTGAGAGCTACCA 18rs21_a12.seq
1251 AGCAAATATAGCATACAAAGGAATCGCAAAGACATAGTTGAGAGCTACCA 515_a12.seq
1251 AGCAAATATAGCATACAAAGGAATCGCAAAGACATAGTTGAGAGCTACCA cjb111_a12.seq
1251 AGCAAATATAGCATACAAAGGAATCGCAAAGACATAGTTGAGAGCTACCA h36b_a12.seq

TAGATACGGTCAAGCTAACTGTACCAAATAGACTAGCTTTAATAAAATCT Majority
         1310      1320      1330      1340      1350
1301 TAGATACGGTCAAGCTAACTGTACCAAATAGACTAGCTTTAATAAAATCT
1301 TAGATACGGTCAAGCTAACTGTACCAAATAGACTAGCTTTAATAAAATCT 2603_a12.seq
1301 TAGATACAGTCAAGCTAACTGTACCAAATAGACTAGCTTTAATAAAATCT 18rs21_a12.seq
1301 TAGATACGGTCAAGCTAACTGTACCAAATAGACTAGCTTTAATAAAATCT 515_a12.seq
1301 TAGATACGGTCAAGCTAACTGTACCAAATAGACTAGCTTTAATAAAATCT cjb111_a12.seq
1301 TAGATACGGTCAAGCTAACTGTACCAAATAGACTAGCTTTAATAAAATCT h36b_a12.seq
```

FIGURE 19B

```
TTTGCACTCTCTCTATTTTTCCAGAAAATAGCGAAACTTGCTAAAAATAA Majority
        1360      1370      1380      1390      1400
1351 TTTGCACTCTCTCTATTTTTCCAGAAAATAGCGAAACTTGCTAAAAATAA
1351 TTTGCACTCTCTCTATTTTTCCAGAAAATAGCGAAACTTGCTAAAAATAA 2603_a12.seq
1351 TTTGCACTCTCTCTATTTTTCCAGAAAATAGCGAAACTTGCTAAAAATAA 18rs21_a12.seq
1351 TTTGCACTCTCTCTATTTTTCCAGAAAATAGCGAAACTTGCTAAAAATAA 515_a12.seq
1351 TTTGCACTCTCTCTATTTTTCCAGAAAATAGCGAAACTTGCTAAAAATAA cjb111_a12.seq
1351 TTTGCACTCTCTCTATTTTTCCAGAAAATAGCGAAACTTGCTAAAAATAA h36b_a12.seq

AGCTAGAGCAACCATATTCATCGGTAAACCAATAAACGTTTCTGGACCAC Majority
        1410      1420      1430      1440      1450
1401 AGCTAGAGCAACCATATTCATCGGTAAACCAATAAACGTTTCTGGACCAC
1401 AGCTAGAGCAACCATATTCATCGGTAAACCAATAAACGTTTCTGGACCAC 2603_a12.seq
1401 AGCTAGAGCAACCATATTCATCGGTAAACCAATAAACGTTTCTGGACCAC 18rs21_a12.seq
1401 AGCTAGAGCAACCATATTCATCGGTAAACCGATAAAGGTTTCTGGACCAC 515_a12.seq
1401 AGCTAGAGCAACCATATTCATCGGTAAACCAATAAACGTTTCTGGACCAC cjb111_a12.seq
1401 AGCTAGAGCAACCATATTCATCGGTAAACCAATAAACGTTTCTGGACCAC h36b_a12.seq

GATTAGCAAGTATAACTTTTAAAAGTGATCTTAATAAGAGTACACCATAA Majority
        1460      1470      1480      1490      1500
1451 GATTAGCAAGTATAACTTTTAAAAGTGATCTTAATAAGAGTACACCATAA
1451 GATTAGCAAGTATAACTTTTAAAAGTGATCTTAATAAGAGTACACCATAA 2603_a12.seq
1451 GATTAGCAAGTATAACTTTTAAAAGTGATCTTAATAAGAGTACACCATAA 18rs21_a12.seq
1451 GATTAGCAAGTATAACTTTTAAAAGTGATCTTAATAAGAGTACACCATAA 515_a12.seq
1451 GATTAGCAAGTATAACTTTTAAAAGTGATCTTAATAAGAGTACACCATAA cjb111_a12.seq
1451 GATTAGCAAGTATAACTTTTAAAAGTGATCTTAATAAGAGTACACCATAA h36b_a12.seq

CTTGATTTCAAATCAAATAAAATAAAAAGCAACTAACATCGGAAGGATTGA Majority
        1510      1520      1530      1540      1550
1501 CTTGATTTCAAATCAAATAAAATAAAAAGCAACTAACATCGGAAGGATTGA
1501 CTTGATTTCAAATCAAATAAAATAAAAAGCAACTAACATCGGAAGGATTGA 2603_a12.seq
1501 CTTGATTTCAAATCAAATAAAATAAAAAGTAACTAACATCGGAAGGATTGA 18rs21_a12.seq
1501 CTTGATTTCAAATCAAATAAAATAAAAAGCAACTAACATCGGAAGGATTGA 515_a12.seq
1501 CTTGATTTCAAATCAAATAAAATAAAAAGCAACTAACATCGGAAGGATTGA cjb111_a12.seq
1501 CTTGATTTCAAATCAAATAAAATAAAAAGCAACTAACATCGGAAGCGATTGA h36b_a12.seq

AAAATCAACCTTTAAAAATTCTGCTCCTGGTATTAATGGAAATGAAACCA Majority
        1560      1570      1580      1590      1600
1551 AAAATCAACCTTTAAAAATTCTGCTCCTGGTATTAATGGAAATGAAACCA
1551 AAAATCAACCTTTAAAAATTCTGCTCCTGGTATTAATGGAAATGAAACCA 2603_a12.seq
1551 AAAATCAACCTTTAAAAATTCTGCTCCTGGTATTAATGGAAATGAAACCA 18rs21_a12.seq
1551 AAAATCAACCTTTAAAAATTCTGCTCCTGGTATTAATGGAAATGAAACCA 515_a12.seq
1551 AAAATCAACCTTTAAAAATTCTGCTCCTGGTATTAATGGAAATGAAACCA cjb111_a12.seq
1551 AAAATCAACCTTTAAAAATTCTGCTCCTGGTATTAATGGAAATGAAACCA h36b_a12.seq

TCATCAATACAAAAGATAAGGCAGAAAGAATGGCGATTGTCACCATTTTA Majority
        1610      1620      1630      1640      1650
1601 TCATCAATACAAAAGATAAGGCAGAAAGAATGGCGATTGTCACCATTTTA
1601 TCATCAATACAAAAGATAAGGCAGAAAGAATGGCGATTGTCACCATTTTA 2603_a12.seq
1601 TCATCAATACAAAAGATAAGGCAGAAAGAATGGCGATTGTCACCATTTTA 18rs21_a12.seq
1601 TCATCAATACAAAAGATAAGGCAGAAAGAATGGCGATTGTCACCATTTTA 515_a12.seq
1601 TCATCAATACAAAAGATAAGGCAGAAAGAATGGCGATTGTCACCATTTTA cjb111_a12.seq
1601 TCATCAATACAAAAGATAAGGCAGAAAGAATGGCGATTGTCACCATTTTA h36b_a12.seq

CGTGTATTTGTCATAAAAAAATTCCTCCAATTTAAATAAATTGAAAGAAG Majority
        1660      1670      1680      1690      1700
1651 CGTGTATTTGTCATAAAAAAATTCCTCCAATTTAAATAAATTGAAAGAAG
1651 CGTGTATTTGTCATAAAAAAATTCCTCCAATTTAAATAAATTGAAAGAAG 2603_a12.seq
1651 CGTGTATTTGTCATAAAAAAATTCCTCCAATTTAAATAAATTGAAAGAAG 18rs21_a12.seq
1651 CGTGTATTTGTCATAAAAAAATTCCTCCAATTTAAATAAATTGAAAGAAG 515_a12.seq
1651 CGTGTATTTGTCATAAAAAAATTCCTCCAATTTAAATAAATTGAAAGAAG cjb111_a12.seq
1651 CGTGTATTTGTCATAAAAAAATTCCTCCAATTTAAATAAATTGAAAGAAG h36b_a12.seq

CTCCAAAGGTAAGCGTATGTACGCGAAAAAAA-CCTTTGTCTTCTCCCATC Majority
        1710      1720      1730      1740      1750
1701 CTCCAAAGGTAAGCGTATGTACGCGAAAAAAA-CCTTTGTCTTCTCCCATC
1701 CTCCAAAGGTAAGCGTATGTACGCGAAAAAAA-CCTTTGTCTTCTCCCATC 2603_a12.seq
1701 CTCCAAAGGTAAGCGTATGTACGCGAAAAAAA-CCTTTGTCTTCTCCCATC 18rs21_a12.seq
1701 CTCCAAAGGTAAGCGTAGGTACGCGAAAAAAACCTTTGTCTTCTCCCATC 515_a12.seq
1701 CTCCAAAGGTAAGCGTATGTACGCGAAAAAAA-CCTTTGTCTTCTCCCATC cjb111_a12.seq
1701 CTCCAAAGGTAAGCGTATGTACGCGAAAAAAA-CCTTTGTCTTCTCCCATC h36b_a12.seq

CAGACTTTACTGTCGGTTGTGGAATCTCACCACATCAGCTTTCGCTCGCG Majority
        1760      1770      1780      1790      1800
1750 CAGACTTTACTGTCGGTTGTGGAATCTCACCACATCAGCTTTCGCTCGCG
1750 CAGACTTTACTGTCGGTTGTGGAATCTCACCACATCAGCTTTCGCTCGCG 2603_a12.seq
1751 CAGACTTTACTGTCGGTTGTGGAATCTCACCACATCAGCTTTCGCTCGCG 18rs21_a12.seq
1750 CAGACTTTACTGTCGGTTGTGGAATCTCACCACATCAGCTTTCGCTCGCG 515_a12.seq
1750 CAGACTTTACTGTCGGTTGTGGAATCTCACCACATCAGCTTTCGCTCGCG cjb111_a12.seq
1750 CAGACTTTACTGTCGGTTGTGGAATCTCACCACATCAGCTTTCGCTCGCG h36b_a12.seq
```

FIGURE 19C

GACTGATGCTTCACAACTGACAAATAAGTTGGAAGCGATTACCGCCGGTC Majority

1810 1820 1830 1840 1850

1800 GACTGATGCTTCACAACTGACAAATAAGTTGGAAGCGATTACCGCCGGTC 2603_a12.seq
1800 GACTGATGCTTCACAACTGACAAATAAGTTGGAAGCGATTACCGCCGGTC 18rs21_a12.seq
1801 GACTGATGCTTCACAACTGACAAATAAGTTGGAAGCGATTACCGCCGGTC 515_a12.seq
1800 GACTGATGCTTCACAACTGACAAATAAGTTGGAAGCGATTACCGCCGGTC cjb111_a12.seq
1800 GACTGATGCTTCACAACTGACAAATAAGTTGGAAGCGATTACCGCCGGTC h36b_a12.seq

GGGAATTACACCCTGCCCTGAAGACACCTATAGCATAACAAAAAAAACTT Majority

1860 1870 1880 1890 1900

1850 GGGAATTACACCCTGCCCTGAAGACACCTATAGCATAACAAAAAAAACTT 2603_a12.seq
1850 GGGAATTACACCCTGCCCTGAAGACACCTATAGCATAACAAAAAAAACTT 18rs21_a12.seq
1851 GGGAATTACACCCTGCCCTGAAGACACCTATAGCATAACAAAAAAAACTT 515_a12.seq
1850 GGGAATTACACCCTGCCCTGAAGACACCTATAGCATAACAAAAAAAACTT cjb111_a12.seq
1850 GGGAATTACACCCTGCCCTGAAGACACCTATAGCATAACAAAAAAAACTT h36b_a12.seq

GCAATTGCAAGTTTTTTAATTACTAATTAGTAGTAGTGATTAAAAATCAT Majority

1910 1920 1930 1940 1950

1900 GCAATTGCAAGTTTTTTAATTACTAATTAGTAGTAGTGATTAAAAATCAT 2603_a12.seq
1900 GCAATTGCAAGTTTTTTAATTACTAATTAGTAGTAGTGATTAAAAATCAT 18rs21_a12.seq
1901 GCAATTGCAAGTTTTTTAATTACTAATTAGTAGTAGTGATTAAAAATCAT 515_a12.seq
1900 GCAATTGCAAGTTTTTTAATTACTAATTAGTAGTAGTGATTAAAAATCAT cjb111_a12.seq
1900 GCAATTGCAAGTTTTTTAATTACTAATTAGTAGTAGTGATTAAAAATCAT h36b_a12.seq

ATTAATACCAAATTACTATACTGTATCGTTTCTTTCAGATTTGCTATTTT Majority

1960 1970 1980 1990 2000

1950 ATTAATACCAAATTACTATACTGTATCGTTTCTTTCAGATTTGCTATTTT 2603_a12.seq
1950 ATTAATACCAAATTACTATACTGTATCGTTTCTTTCAGATTTGCTATTTT 18rs21_a12.seq
1951 ATTAATACCAAATTACTATACTGTATCGTTTCTTTCAGATTTGCTATTTT 515_a12.seq
1950 ATTAATACCAAATTACTATACTGTATCGTTTCTTTCAGATTTGCTATTTT cjb111_a12.seq
1950 ATTAATACCAAATTACTATACTGTATCGTTTCTTTCAGATTTGCTATTTT h36b_a12.seq

TAGTTTTTCTTAAAAAGATAAACAAAATTCCCAAAATAATACAACCAAGA Majority

2010 2020 2030 2040 2050

2000 TAGTTTTTCTTAAAAAGATAAACAAAATTCCCAAAATAATACAACCAAGA 2603_a12.seq
2000 TAGTTTTTCTTAAAAAGATAAACAAAATTCCCAAAATAATACAACCAAGA 18rs21_a12.seq
2001 TAGTTTTTCTTAAAAAGATAAACAAAATTCCCAAAATAATACAACCAAGA 515_a12.seq
2000 TAGTTTTTCTTAAAAAGATAAACAAAATTCCCAAAATAATACAACCAAGA cjb111_a12.seq
2000 TAGTTTTTCTTAAAAAGATAAACAAAATTCCCAAAATAATACAACCAAGA h36b_a12.seq

ATTGTCAGTCCTCCACCAATAATCATTCCTGTTTTAGGAAGAAATGATTG Majority

2060 2070 2080 2090 2100

2050 ATTGTCAGTCCTCCACCAATAATCATTCCTGTTTTAGGAAGAAATGATTG 2603_a12.seq
2050 ATTGTCAGTCCTCCACCAATAATCATTCCTGTTTTAGGAAGAAATGATTG 18rs21_a12.seq
2051 ATTGTCAGTCCTCCACCAATAATCATTCCTGTTTTAGGAAGAAATGATTG 515_a12.seq
2050 ATTGTCAGTCCTCCACCAATAATCATTCCTGTTTTAGGAAGAAATGATTG cjb111_a12.seq
2050 ATTGTCAGTCCTCCGCCAATAATCATTCCTGTTTTAGGAAGAAATGATTG h36b_a12.seq

TGGAAAAAGCGGTTGTGATGGTTTAGGATTTGTTGGTGGAGGAGTTTCTT Majority

2110 2120 2130 2140 2150

2100 TGGAAAAAGCGGTTGTGATGGTTTAGGATTTGTTGGTGGAGGAGTTTCTT 2603_a12.seq
2100 TGGAAAAAGCGGTTGTGATGGTTTAGGATTTGTTGGTGGAGGAGTTTCTT 18rs21_a12.seq
2101 TGGAAAAAGCGGTTGTGATGGTTTAGGATTTGTTGGTGGAGGAGTTTCTT 515_a12.seq
2100 TGGAAAAAGCGGTTGTGATGGTTTAGGATTTGTTGGTGGAGGAGTTTCTT cjb111_a12.seq
2100 TGGAAAAAGCGGTTGTGATGGTTTAGGATTTGTTGGTGGAGGAGTTTCTT h36b_a12.seq

TTTCGTTTTCTACCTCTACTTCCTGTGTTTTATTAGCAACTACAGCAACT Majority

2160 2170 2180 2190 2200

2150 TTTCGTTTTCTACCTCTACTTCCTGTGTTTTATTAGCAACTACAGCAACT 2603_a12.seq
2150 TTTCGTTTTCTACCTCTACTTCCTGTGTTTTATTAGCAACTACAGCAACT 18rs21_a12.seq
2151 TTTCGTTTTCTACCTCTACTTCCTGTGTTTTATTAGCAACTACAGCAACT 515_a12.seq
2150 TTTCGTTTTCTACCTCTACTTCCTGTGTTTTATTAGCAACTACAGCAACT cjb111_a12.seq
2150 TTTCGTTTTCTACCTCTACTTCCTGTGTTTTATTAGCAACTACAGCAACT h36b_a12.seq

ACAGCATCCTTCATAGATATACGGTAACCAGTTAGTGCTTTTGCTTCTCG Majority

2210 2220 2230 2240 2250

2200 ACAGCATCCTTCATAGATATACGGTAACCAGTTAGTGCTTTTGCTTCTCG 2603_a12.seq
2200 ACAGCATCCTTCATAGATATACGGTAACCAGTTAGTGCTTTTGCTTCTCG 18rs21_a12.seq
2201 ACAGCATCCTTCATAGATATACGGTAACCAGTTAGTGCTTTTACTTCTCG 515_a12.seq
2200 ACAGCATCCTTCATAGATATACGGTAACCAGTTAGTGCTTTTGCTTCTCG cjb111_a12.seq
2200 ACAGCATCCTTCATAGATATACGGTAACCAGTTAGTGCTTTTGCTTCTCG h36b_a12.seq

FIGURE 19D

```
                                                
```

```
                                                
```

```
                                                
```

```
                TGATTATCAAGTTAGTTACTTGATTATTAGCTTTAATATTATAGTACCA Majority
                        3610      3620      3630      3640      3650
3600 TGATTATCA-AGTTAGTTACTTGATTATTAGCTTTAATATTATAGTACCA
3600 TGATTATCAAAGTTAGTTACTTGATTATTAGCTTTAATATTATAGTACCA 2603_al2.seq
3601 TGATTATCAAAGTTAGTTACTTGATTATTAGCTTTAATATTATAGTACCA 18rs21_al2.seq
3600 TGATTATCAAAGTTAGTTACTTGATTATTAGCTTTAATATTATAGTACCA 515_al2.seq
3600 TGATTATCAAAGTTAGTTACTTGATTATTAGCTTTAATATTATAGTACCA cjb111_al2.seq
3600 TGATTATCAAAGTTAGTTACTTGATTATTAGCTTTAATATTATAGTACCA h36b_al2.seq

                ATTTGAAATAAAAGGATATGAGGTTATCAAAAGACCAACTAAGAACAATA Majority
                        3660      3670      3680      3690      3700
3650 ATTTGAAATAAAAGGATATGAGGTTATCAAAAGACCAACTAAGAACAATA
3650 ATTTGAAATAAAAGGATATGAGGTTATCAAAAGACCAACTAAGAACAATA 2603_al2.seq
3651 ATTTGAAATAAAAGGATATGAGGTTATCAAAAGACCAACTAAGAACAATA 18rs21_al2.seq
3650 ATTTGAAATAAAAGGATATGAGGTTATCAAAAGACCNACTAAGAANAATA 515_al2.seq
3650 ATTTGAAATAAAAGGATATGAGGTTATCAAAAGACCAACTAAGAACAATA cjb111_al2.seq
3650 ATTTGAAATAAAAGGATATGAGGTTATCAAAAGACCAACTAAGAACAATA h36b_al2.seq

                GTATCAAGCCTACATTCATCCATCGATTTAAAACGACCGATTTCTTAAGG Majority
                        3710      3720      3730      3740      3750
3700 GTATCAAGCCTACATTCATCCATCGATTTAAAACGACCGATTTCTTAAGG
3700 GTATCAAGCCTACATTCATCCATCGATTTAAAACGACCGATTTCTTAAGG 2603_al2.seq
3701 NTATCAANCCTANTTCATCCATCGATTTAAAACGACCGATTTCTTAAGG 18rs21_al2.seq
3700 GTATCANGCCTACATTCATCCATCGATTTAAAACGACCGATTTCTTAAGG 515_al2.seq
3700 GTATCAAGCCTACATTCATCCATCGATTTAAAACGACCGATTTCTTAAGG cjb111_al2.seq
3700 GTATCAAGCCTACATTCATCCATCGATTTAAAACGACCGATTTCTTAAGG h36b_al2.seq

                TTTTTCTGAAATTTTCCTCCCATTATGATTCAATTCCTTTTCTAACACTT Majority
                        3760      3770      3780      3790      3800
3750 TTTTTCTGAAATTTTCCTCCCATTATGATTCAATTCCTTTTCTAACACTT
3750 TTTTTCTGAAATTTTCCTCCCATTATGATTCAATTCCTTTTCTAACACTT 2603_al2.seq
3751 TTTTTCTGAAATTTTCCTCNCATTATGATTCAATTCCTTTTCTAACACTT 18rs21_al2.seq
3750 TTTTTCTGAAATTTTCCTCCCATTATGATTCAATTCCTTTTCTAACACTT 515_al2.seq
3750 TTTTTCTGAAATTTTCCTCCCATTATGATTCAATTCCTTTTCTAACACTT cjb111_al2.seq
3750 TTTTTCTGAAATNTTCCTCCCATTATGATTCAATTCCTTTTCTAACACTT h36b_al2.seq

                GCTAAACGATTTTTTTGACGTTGACGTTTTATTAACCAAAGTAACCAAGC Majority
                        3810      3820      3830      3840      3850
3800 GCTAAACGATTTTTTTGACGTTGACGTTTTATTAACCAAAGTAACCAAGC
3800 GCTAAACGATTTTTTTGACGTTGACGTTTTATTAACCAAAGTAACCAAGC 2603_al2.seq
3801 GNTAAACGATTTTTTTGACGTTGACGTTTTATTAACCAAAGTAACCAAGC 18rs21_al2.seq
3800 GCTAAACGATTTTTTTGACGTTGACGTTTTATTAACCAAAGTAACCAAGC 515_al2.seq
3800 GCTAAACGATTTTTTTGACGTTGACGTTTTATTAACCAAAGTAACCAAGC cjb111_al2.seq
3800 GCTAAACGATTTTTTTGACGTTGACGTTTTATTAACCAAAGTAACCAAGC h36b_al2.seq

                AATAATAACTAAAGATATATAGAAAAGATATCTATAAATCGTGTTTAAAT Majority
                        3860      3870      3880      3890      3900
3850 AATAATAACTAAAGATATATAGAANAGATATCTATAAATCGTGTTTAAAT
3850 AATAATAACTAAAGATATATAGAANAGATATCTATAAATCGTGTTTAAAT 2603_al2.seq
3851 AATAATAACTAAAGATATATAGAAAAGATATCTATAAATCGTGTTTAAAT 18rs21_al2.seq
3850 AATAATAACTAAAGATATATAGAAAAGATATCTATAAATCGTGTTTAAAT 515_al2.seq
3850 AATAATAACTAAAGATATATAGAAAAGATATCTATAAATCGTGTTTAAAT cjb111_al2.seq
3850 AATAATAACTAAAGATATATAGAAAAGATATCTATAAATCGTGTTTAAAT h36b_al2.seq

                GACCGTCTTTTATTAATTTTTCATCAATAGGACCTTTATAAGGGATACGA Majority
                        3910      3920      3930      3940      3950
3900 GACCGTCTTTTATTAATTTTTCATCAATAGGACCTTTATAAGGGATACGA
3900 GACCGTCTTTTATTAATTTTTCATCAATAGGACCTTTATAAGGGATACGA 2603_al2.seq
3901 GACCGTCTTTTATTAATTTTTCATCAATAGGACCTTTATAAGGGATACGA 18rs21_al2.seq
3900 GACCNTCTTTTATTAATTTTTCATCAATAGGACCTTTATAAGGGATACGA 515_al2.seq
3900 GACCNTCTTTTANTAATTTTTCATCAATANGACCTTTATAAGGNATACGA cjb111_al2.seq
3900 GACCNTCTTTTANTAATTTTTCATCAATAGGACCTTTATAAGGNATACGA h36b_al2.seq

                TGTCCCCTTACTAAAAGTCTGTGTGTATTGACCATAATCGGGGTGCAAGT Majority
                        3960      3970      3980      3990      4000
3950 TGTCCCCTTACTAAAAGTCTGTGTGTATTGANCATAATCGGGGTGCAAGT
3950 TGTCCCCTTACTAAAAGTCTGTGTGTATTGANCATAATCGGGGTGCAAGT 2603_al2.seq
3951 TGNCCCCTTACTAAAAGTCTGTGTGTATTGACCATAATCGGGGTGCAAGT 18rs21_al2.seq
3950 TGTCCCCTTACTAAAAGTCTGTGTGTATTGACCATAATCGGGGTGCAAGT 515_al2.seq
3950 TGTCCCCTTACTAAAAGTCTGTGTGTATTGACCATAATCGGGGTGCAAGT cjb111_al2.seq
3950 TGTCCCCTTACTAAAAGTCTGTGTGTATTGACCATAATCGGGGTGCAAGT h36b_al2.seq

                CAATAAGGTTGCATAATCATGTCCAGGAACAACCAACAAATCTGAAAAGT Majority
                        4010      4020      4030      4040      4050
4000 CAATAAGGTTGCATAATCATGTCCAGGAACAACCAACAAATCTGAAAAGT
4000 CAATAAGGTTGCATAATCATGTCCAGGAACAACCAACAAATCTGAAAAGT 2603_al2.seq
4001 CAATAAGGTTGCATAATCATGTCCAGGAACAACCAACAAATCTGAAAAGT 18rs21_al2.seq
4000 NAATAAGGTTGCATAATCATGTCCAGGAACAACCAACAAATCTGAAAAGT 515_al2.seq
4000 CAATAAGGTTGCATAATCATGTCCAGGAACAACCAACANATCTGAAAAGT cjb111_al2.seq
4000 CAATAAGGTTGCATAATCATGTCCAGGAACAACCAANAAATCTGAAAAGT h36b_al2.seq
```

FIGURE 19H

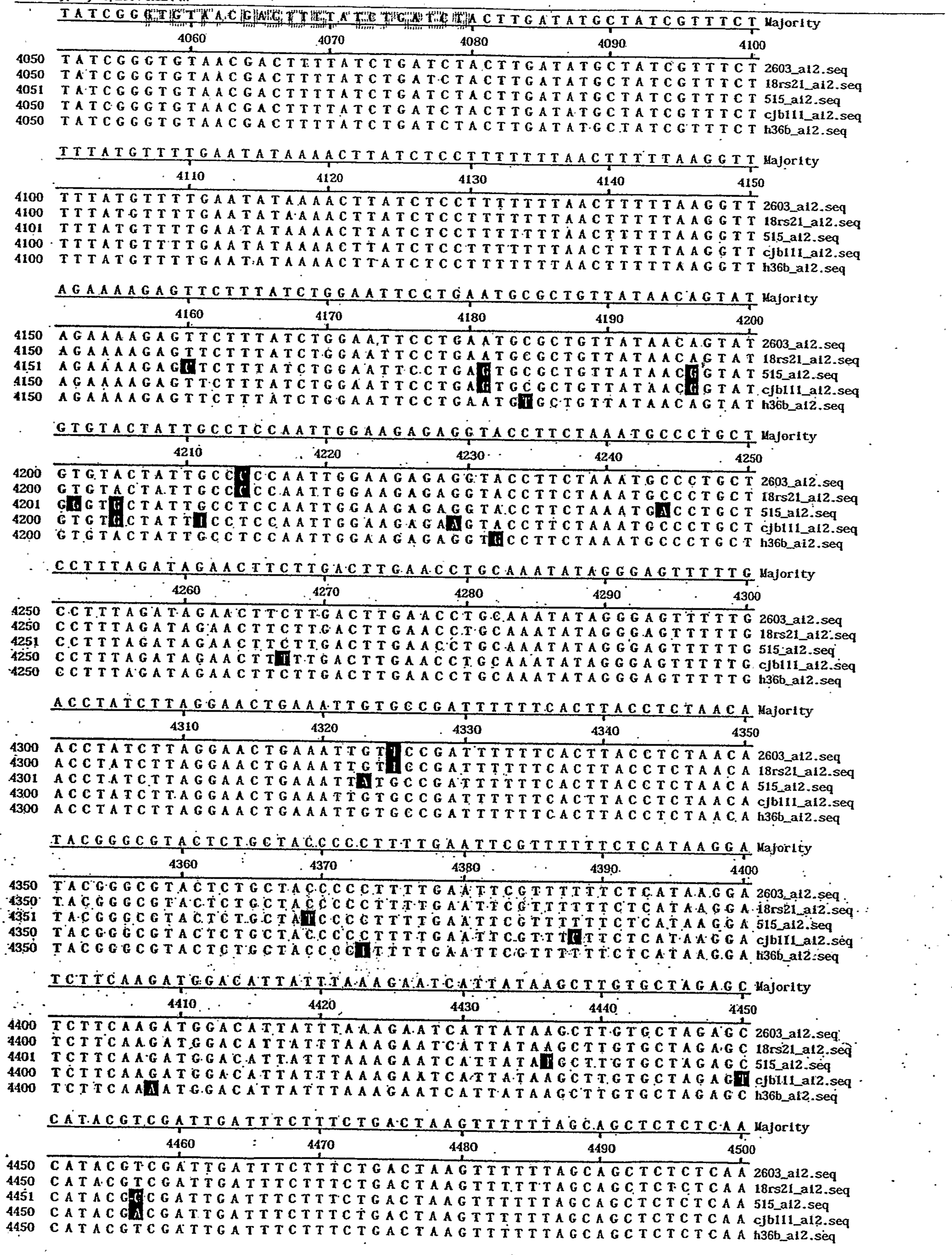

Majority
2603_a12.seq
18rs21_a12.seq
515_a12.seq
cjb111_a12.seq
h36b_a12.seq

FIGURE 191

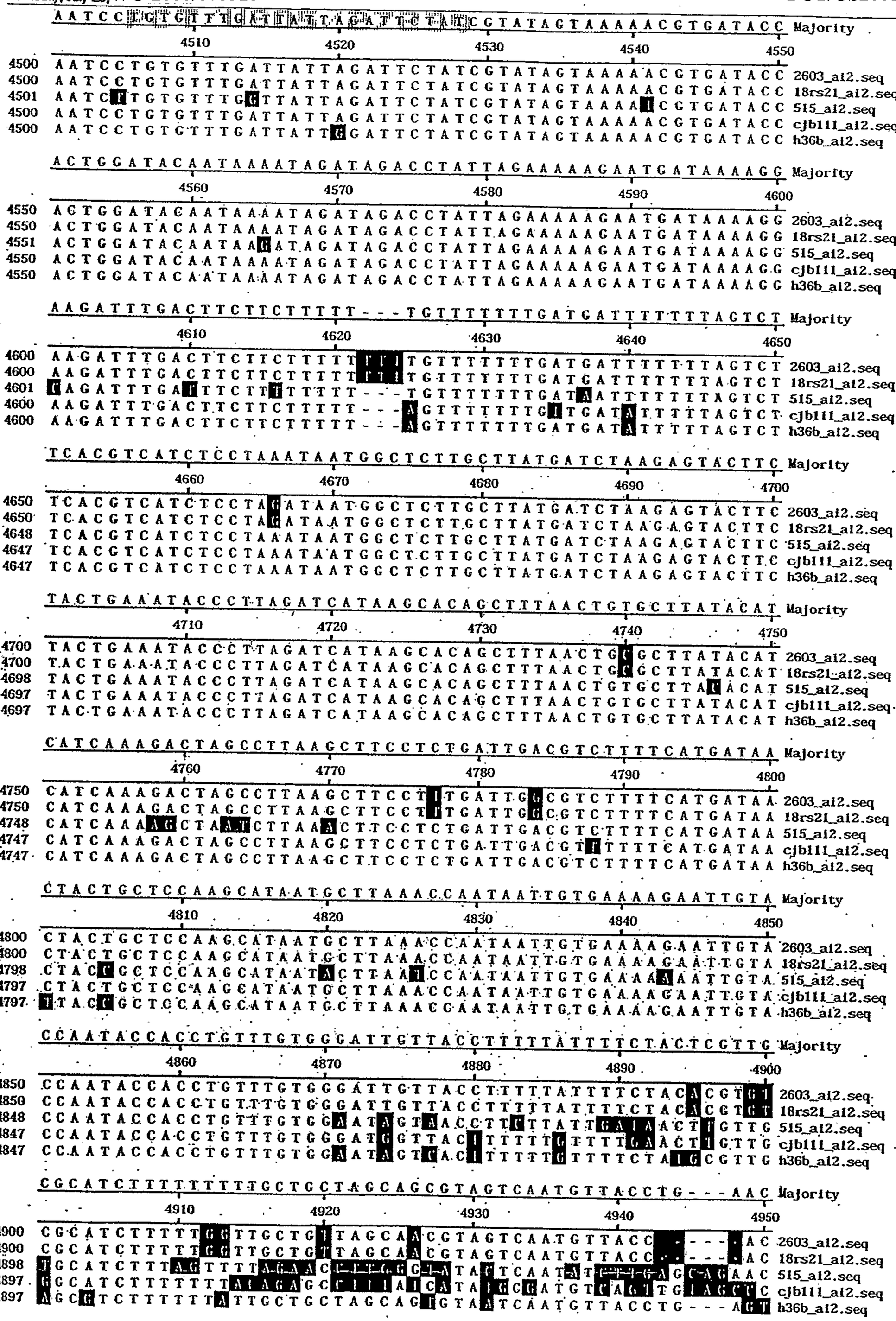

AATCCTGTGTTTGATTATTAGATTCTATCGTATAGTAAAAACGTGATACC Majority
4510 4520 4530 4540 4550
4500 AATCCTGTGTTTGATTATTAGATTCTATCGTATAGTAAAAACGTGATACC 2603_a12.seq
4500 AATCCTGTGTTTGATTATTAGATTCTATCGTATAGTAAAAACGTGATACC 18rs21_a12.seq
4501 AATCTTGTGTTTGTTATTAGATTCTATCGTATAGTAAAATCGTGATACC 515_a12.seq
4500 AATCCTGTGTTTGATTATTAGATTCTATCGTATAGTAAAAACGTGATACC cjb111_a12.seq
4500 AATCCTGTGTTTGATTATTGGATTCTATCGTATAGTAAAAACGTGATACC h36b_a12.seq
ACTGGATACAATAAAATAGATAGACCTATTAGAAAAAGAATGATAAAAGG Majority
4560 4570 4580 4590 4600
4550 ACTGGATACAATAAAATAGATAGACCTATTAGAAAAAGAATGATAAAAGG 2603_a12.seq
4550 ACTGGATACAATAAAATAGATAGACCTATTAGAAAAAGAATGATAAAAGG 18rs21_a12.seq
4551 ACTGGATACAATAAGATAGATAGACCTATTAGAAAAAGAATGATAAAAGG 515_a12.seq
4550 ACTGGATACAATAAAATAGATAGACCTATTAGAAAAAGAATGATAAAAGG cjb111_a12.seq
4550 ACTGGATACAATAAAATAGATAGACCTATTAGAAAAAGAATGATAAAAGG h36b_a12.seq
AAGATTTGACTTCTTCTTTTT---TGTTTTTTTGATGATTTTTTTAGTCT Majority
4610 4620 4630 4640 4650
4600 AAGATTTGACTTCTTCTTTTTTTTTGTTTTTTTTGATGATTTTTTTAGTCT 2603_a12.seq
4600 AAGATTTGACTTCTTCTTTTTTTTTGTTTTTTTTGATGATTTTTTTAGTCT 18rs21_a12.seq
4601 GAGATTTGATTTCTTTTTTTT---TGTTTTTTTTGATAATTTTTTTAGTCT 515_a12.seq
4600 AAGATTTGACTTCTTCTTTTT---AGTTTTTTTGTTGATATTTTTTAGTCT cjb111_a12.seq
4600 AAGATTTGACTTCTTCTTTTT---AGTTTTTTTGATGATATTTTTAGTCT h36b_a12.seq
TCACGTCATCTCCTAAATAATGGCTCTTGCTTATGATCTAAGAGTACTTC Majority
4660 4670 4680 4690 4700
4650 TCACGTCATCTCCTAGATAATGGCTCTTGCTTATGATCTAAGAGTACTTC 2603_a12.seq
4650 TCACGTCATCTCCTAGATAATGGCTCTTGCTTATGATCTAAGAGTACTTC 18rs21_a12.seq
4648 TCACGTCATCTCCTAAATAATGGCTCTTGCTTATGATCTAAGAGTACTTC 515_a12.seq
4647 TCACGTCATCTCCTAAATAATGGCTCTTGCTTATGATCTAAGAGTACTTC cjb111_a12.seq
4647 TCACGTCATCTCCTAAATAATGGCTCTTGCTTATGATCTAAGAGTACTTC h36b_a12.seq
TACTGAAATACCCTTAGATCATAAGCACAGCTTTAACTGTGCTTATACAT Majority
4710 4720 4730 4740 4750
4700 TACTGAAATACCCTTAGATCATAAGCACAGCTTTAACTGCGCTTATACAT 2603_a12.seq
4700 TACTGAAATACCCTTAGATCATAAGCACAGCTTTAACTGCGCTTATACAT 18rs21_a12.seq
4698 TACTGAAATACCCTTAGATCATAAGCACAGCTTTAACTGTGCTTACACAT 515_a12.seq
4697 TACTGAAATACCCTTAGATCATAAGCACAGCTTTAACTGTGCTTATACAT cjb111_a12.seq
4697 TACTGAAATACCCTTAGATCATAAGCACAGCTTTAACTGTGCTTATACAT h36b_a12.seq
CATCAAAGACTAGCCTTAAGCTTCCTCTGATTGACGTCTTTTCATGATAA Majority
4760 4770 4780 4790 4800
4750 CATCAAAGACTAGCCTTAAGCTTCCTTTGATTGGCGTCTTTTCATGATAA 2603_a12.seq
4750 CATCAAAGACTAGCCTTAAGCTTCCTTTGATTGGCGTCTTTTCATGATAA 18rs21_a12.seq
4748 CATCAAAAGCTAATCTTAAACTTCCTCTGATTGACGTCTTTTCATGATAA 515_a12.seq
4747 CATCAAAGACTAGCCTTAAGCTTCCTCTGATTGACGTTTTTCATGATAA cjb111_a12.seq
4747 CATCAAAGACTAGCCTTAAGCTTCCTCTGATTGACGTCTTTTCATGATAA h36b_a12.seq
CTACTGCTCCAAGCATAATGCTTAAACCAATAATTGTGAAAAGAATTGTA Majority
4810 4820 4830 4840 4850
4800 CTACTGCTCCAAGCATAATGCTTAAACCAATAATTGTGAAAAGAATTGTA 2603_a12.seq
4800 CTACTGCTCCAAGCATAATGCTTAAACCAATAATTGTGAAAAGAATTGTA 18rs21_a12.seq
4798 CTACTGCTCCAAGCATAATACTTAATCCAATAATTGTGAAAAAAATTGTA 515_a12.seq
4797 CTACTGCTCCAAGCATAATGCTTAAACCAATAATTGTGAAAAGAATTGTA cjb111_a12.seq
4797 TTACTGCTCCAAGCATAATGCTTAAACCAATAATTGTGAAAAGAATTGTA h36b_a12.seq
CCAATACCACCTGTTTGTGGGATTGTTACCTTTTTATTTTCTACTCGTTG Majority
4860 4870 4880 4890 4900
4850 CCAATACCACCTGTTTGTGGGATTGTTACCTTTTTATTTTCTACACGTGT 2603_a12.seq
4850 CCAATACCACCTGTTTGTGGGATTGTTACCTTTTTATTTTCTACACGTGT 18rs21_a12.seq
4848 CCAATACCACCTGTTTGTGGAATAGTAACCTTCTTATTGAAACTTGTTG 515_a12.seq
4847 CCAATACCACCTGTTTGTGGATGGTTACTTTTTTGTTTTGAACTTGTTG cjb111_a12.seq
4847 CCAATACCACCTGTTTGTGGAATAGTCACTTTTTTGTTTTCTAGCGTTG h36b_a12.seq
CGCATCTTTTTTTTTGCTGCTAGCAGCGTAGTCAATGTTACCTG---AAC Majority
4910 4920 4930 4940 4950
4900 CGCATCTTTTTTGGTTGCTGTAGCAACGTAGTCAATGTTACC-----AC 2603_a12.seq
4900 CGCATCTTTTTTGGTTGCTGTAGCAACGTAGTCAATGTTACC-----AC 18rs21_a12.seq
4898 TGCATCTTTAGTTTTAGAACTTTGGGTATCTCAATATTGCAGCAGAAC 515_a12.seq
4897 GGCATCTTTTTTTTAAAGAGCCTTTAACATAGCGATGTCAGTTGAGCTC cjb111_a12.seq
4897 AGCGTCTTTTTTTATTGCTGCTAGCAGTGTAATCAATGTTACCTG---AGT h36b_a12.seq

FIGURE 19J

CTTTTCTGTTCACCTTTGATTAACTTCAAACTTAACATCACCTGCCAAT Majority
4960 4970 4980 4990 5000
4944 CTGTTATGTATGACCCTTGATTAACTACAAACTTAAIATIACCTGCCAAC 2603_a12.seq
4944 CTGTTATGTATGACCCTTGATTAACTACAAACTTAAIATTACCTGCCAAG 18rs21_a12.seq
4948 CTTTACTGTATGAAGTAGCATTAACATIAAACGAAACATCACCTGAIAAT 515_a12.seq
4947 CTTTGCTATATGATGTGGCAGTIACTTCAAAGTTTIACATCACCTGACAAT cjb111_a12.seq
4944 CGTCTGCTTTIGAGCTGTIICCAACTTIGAAATCAACATCACCTGCTAAT h36b_a12.seq
TTAGCATAACCTGCTGGAGCTTGTGTTTCTTCAAGGTTGTAAGTACCTTC Majority
5010 5020 5030 5040 5050
4994 TTAGCAAATCCTGCTGGAGCAAGTGTTTCTTCAAGGTTGTAAGTACCGTC 2603_a12.seq
4994 TTAGCAAAICCTGCTGGAGCAAGTGTTTCTTCAAGGTTGTAAGTACCGTC 18rs21_a12.seq
4998 TTAGCATAACCAGITGGIGCTTGTGTTTCTTCCAAIGAGTAITGACCTTC 515_a12.seq
4997 GTGGCATAACCTGCTGGIGCTTGAGTTTCTTCCAAGCCATAAGTGCCTTI cjb111_a12.seq
4994 TTAGCGTAACCAGCTGGAGCTTGTGTTTCTTCAAGITTATAAICACCAGT h36b_a12.seq
TGCAAGACCTGTAACTTCAAATTGACCTTGATCGTTTGAAGTCAATTTAA Majority
5060 5070 5080 5090 5100
5044 TGCAAGACCTGTAACTTCAAATTGACCTTGATCGTTTGAAGTGTAGGTAA 2603_a12.seq
5044 TGCAAGACCTGTAACTTCAAATTGACCTTGATCGTTTGAAGTGTAGGTAA 18rs21_a12.seq
5048 AGTCAAGCCAGTAATTTCAAATTGACCTTIATCGTTAGAAATCAATTTAA 515_a12.seq
5047 ATCCAAACCAGTAATTTCAAATTGACCACCGGCGTTAGAGATCAATTTAA cjb111_a12.seq
5044 TGCAAGACCGCTAACTTCAAATTGACCGTCAGCGTTTGAAGTCATTTTAA h36b_a12.seq
CAACTTTAGCAGC-TTTTTATCTGTTACCCACTCATAAGCTGTACGAGCT Majority
5110 5120 5130 5140 5150
5094 IGGCTCTAGC---CTTATCTGTTAICCACTCATAAGCTGTACGAGCC 2603_a12.seq
5094 IGGCTCTAGC---CTTATCTGTTAICCACTCATAAGCTGTACGAGCC 18rs21_a12.seq
5098 CAACATTCIIAGCATTTTTATCTICAACCCAITCGTAIGAGTAGTIAGCT 515_a12.seq
5097 CAACATTAICAGCCTTTTTATCTGCAACCCAITCATAIGAGTAGTTAGCT cjb111_a12.seq
5094 CAACGTTAGCAICTTCTTTATTAGTTACCCACTCGTAAGCTGTACGAGCT h36b_a12.seq
TCAATGAAGGCTGCATTGTAAGCTTCTTGTTTAGTTTTGATAGTTGCTTT Majority
5160 5170 5180 5190 5200
5138 TCAATGAAGGCTGCATCGTAATCTGCTTGTTTAGTTTTGATAAGTICTTT 2603_a12.seq
5138 TCAATGAAGGCTGCATCGTAATCTGCTTGTTTAGTTTTGATAAGTICTTT 18rs21_a12.seq
5148 TIAACAAAGGCATCATTGTAAGCTTTTGTTTIICACTAACGGTAGCCAA 515_a12.seq
5147 TTAACAAAAGCGTCATTGTAAGCTTTTGTTTIIGATCAACAGTAGCCAA cjb111_a12.seq
5144 ACGATGAAGGCTGCATTGTAAGCGTCTTGAGCAGTTTTGATTICTGCTTT h36b_a12.seq
TGCTGTTTTACCTTTTTGACCTTTTTGTTCTTCTGCAGACAACTTGTTAT Majority
5210 5220 5230 5240 5250
5188 TGCAGTAATICCTTTTTTCACCTTTTTTGGTCTGITGCAGACAACTTGTTAT 2603_a12.seq
5188 TGCAGTAATICCTTTTTTCACCTTTTTTGGTCTGITGCAGACAACTTGTTAT 18rs21_a12.seq
5198 TGCTGATTTACCATCTTGACCTTCTTGTTIITTCTTIAGICAAATIATTGT 515_a12.seq
5197 TGCTGTTTTACCTTCTTGGCCTTCTTGTTIITTCTIIAGICAAGTCGTTAT cjb111_a12.seq
5194 AGCTGTTTTCACCTTTAGIACCTTTTTTGATCGTCTGCTGAAGIITGTTGT h36b_a12.seq
AAGCTGCGATAGCTTCATCTAAAGCTATTTTAGTAGTAGCTAAAGCTGTT Majority
5260 5270 5280 5290 5300
5238 AAGCAGCGATAGCTTCATCTAAAGCTATTTTCTTAGCAGCTAAAG---TT 2603_a12.seq
5238 AAGCAGCGATAGCTTCATCTAAAGCTATTTTCITAGCAGCTAAAG---TT 18rs21_a12.seq
5248 AAGCTIIAACAGCAGCATCCAAIGCTGATTTAGTTGAAICTACAGCAGCCT 515_a12.seq
5247 AAGCTIIAACAGCTTCATCCAATGCTAGTTTAGCAGTTIITACAGCTGCC cjb111_a12.seq
5244 AAGCTGCGATIIIAGCGTCAAGAGCAGTTTTAGCAITTGIGIAAGCTGTT h36b_a12.seq
TTTTGAGCTGCTTTTGCTTCATCTGCTTTAAGTGCAAGGTATTTACCTGC Majority
5310 5320 5330 5340 5350
5285 TTTTGACCT--TCTGATTGATCTGCTTTAAGAGCAAGGTATTTACCTGC 2603_a12.seq
5285 TTTTGAICT--TCTGATTGATCTGCTTTAAGAGCAAGGTATTTACCTGC 18rs21_a12.seq
5298 TTTTCTGCAICTGTTGCAACACCTGATTTACGTGCCAAGTACTTLCCAIC 515_a12.seq
5297 TTTGCTTICAGCAGTTGCTGCACGTGCTTTACGTGCCAAGTATTTGCCTIC cjb111_a12.seq
5294 GTTTGAGCTGATATTGICGCAGATGATTTTCAATGCAAGGTATTTACCAGC h36b_a12.seq
TGAGTTTTTAACAACGAATTGTGCACCTGCCAAGCGTTCAGTTTCAGGTT Majority
5360 5370 5380 5390 5400
5332 TGAGTTTTTTCACAACGAATTGTGCACCAGCCAAACGGTCACCTT--GTT 2603_a12.seq
5332 TGAGTTTTTTCACAACGAATTGTGCACCAGCCAAACGGTCACCTT--GTT 18rs21_a12.seq
5348 T---TTCCTTAACAAGGAAGGTAGCIICCTGCCAAGCGTTCIITTCCAIGTT 515_a12.seq
5347 T---TTCGTTAACGAGGAAGGIAGCIICCTGCCAAGCGTTCAGTGTIAGCTT cjb111_a12.seq
5344 TGAATGTTTAACAACGAAGTGTGCACCTIGCAAGCGTTCIGTTTCAGICG h36b_a12.seq

FIGURE 19K

Majority
2603_a12.seq
18rs21_a12.seq
515_a12.seq
cjb111_a12.seq
h36b_a12.seq

FIGURE 19L


TGGTTTCTGGATTTTCAACAATAGCTGGACCGTTAACAGTAGCTCTATAA Majority
5860 5870 5880 5890 5900
5817 TGGATTCTGGATTATCAAIAATIGCTIGACCATTAACAGTAGCACTATAA 2603_a12.seq
5817 TGGATTCTGGATTATCAAIAATIGCTIGACCATTAACAGTAGCACTATAA 18rs21_a12.seq
5818 TGGTTTCTAGAACTTCAACGACAGCAGAACCGTTCAAAGTAGCTGAGTAA 515_a12.seq
5817 TGGTTTCTGGAATTTCAACAGTAGIGGAGCCGTTCACGTAGCTGAGTAA cjb111_a12.seq
5838 TGCTTTCTGGGTTGTCATCAATAGCCGCACCGTTAACTGTTGCGCTATAT h36b_a12.seq
GTCAATGTGAATTCAACATCTTTTGTTTTTGCTGCTTTTTCCACTTTTGC Majority
5910 5920 5930 5940 5950
5867 GTCAATGTAAATTCAAIATCAGCTGTTTTAGCTGCTTTTTCCAATTTGCC 2603_a12.seq
5867 GTCAATGTAAATTCAAIATCAGCTGTTTTAGCTGCTTTTTCCAATTTGCC 18rs21_a12.seq
5868 GTGAICITGAITTCAACATCTTTTGTTTTTGCGGCTTTTGCCACTGCTIC 515_a12.seq
5867 GTGAICITGATTTCAACATCTTTGIITTTTGCAGCTGCTGCIACTGCTGC cjb111_a12.seq
5888 GTCAATGTGAATTCAACAGCTTGTGTTTCIGCTICTTTTTGTAGCTTAGI h36b_a12.seq
CAAACCTTTAGCTGTGAATTTTAATGTGAAACCACGGTCATCAGTTCCAA Majority
5960 5970 5980 5990 6000
5917 CAAICCATCAGCTGTGAATTTTAATGTGAAACCACGGGCATCAATGCTAA 2603_a12.seq
5917 CAATCCATCAGCTGTGAATTTTAATGTGAAACCACGGGCATCAATGCTAA 18rs21_a12.seq
5918 AAGACCTTTGTCAGTCAAGACAAGGCGGAAACCIIGGTCATCIGCTACAA 515_a12.seq
5917 AAGACCTGTTGCATTCAALGCAALGACGGAAACCTTGGTCATCTGTTACGA cjb111_a12.seq
5938 CAAACCTGTAICTGTGAATTTGAAAGTGAAACCACGGTCATCATGTICAA h36b_a12.seq
GTTTGTAGTCTGTATCCTTAACAAAAGTTCCTG------CCGATGCTTCA Majority
6010 6020 6030 6040 6050
5967 GTTCATAGTCTGTATCCTTAGCAAAAGTTICTGIAGIICCIGAAGCTTIA 2603_a12.seq
5967 GTTCATAGTCTGTATCCTTAGCAAAAGTTICTGIAGIICCIGAAGCTTIA 18rs21_a12.seq
5968 GTTTGTAATITGTAGCATCAAGAGIIGCACCAT------CCAATGITACA 515_a12.seq
5967 GTTTGTAGTITAAAACAGGAAAAICIIICACCAT------CCAATGTTACI cjb111_a12.seq
5988 CATTGTAATCCTGTACCTTCTACAAAAGAGCCTG------ACGATGCTTCT h36b_a12.seq
AGGTTAACAGTGTTACCCATTGTCAAACCATTTGACATGCTATCTGTCCA Majority
6060 6070 6080 6090 6100
6017 AGGCTAACAGTIGAACCCATTGTCAAACCATTTGACATIATATCTGTCCA 2603_a12.seq
6017 AGGCTAACAGTIGAACCCATTGTCAAACCATTTGACATIATATCTGTCCA 18rs21_a12.seq
6012 ---GCAAIAICGTTGTIGAAAGTCAAACCIITTGGICATGCTATCGGTCCA 515_a12.seq
6011 ---TTAACGTTGTTGIIGAACGTCAAACCIITTAGICATGCTATCAGTCCA cjb111_a12.seq
6032 AAGTTAACAGTGTTACCCATTGTCAAACCATTTGACATGGTATCGACCCA h36b_a12.seq
AACCAAGTTTTTGTATTTAGAACCTTTGTGAATTTTTGTTTTAACTTCGT Majority
6110 6120 6130 6140 6150
6067 AACCAAGTTTTTGGTATTTAGAACCTTTGTGAATTTTTGTTTTAACTTCAT 2603_a12.seq
6067 AACCAAGTTTTTGGTATTTAGAACCTTTGTGAATTTTTGTTTTAACTTCAT 18rs21_a12.seq
6059 GAITAAITTCTTATAGTCIGAACCTTTTAAGGATTTTTGTTCCAACATGGT 515_a12.seq
6058 AACCAGITTCTTATAGTCIGAGCCTTTTAAGAATTTTTGTTCCAACTATGT cjb111_a12.seq
6082 AGCAAGGTTTTTGTATTTAGACCTTTTIIGGATTTTTGTTTTAACTTCGT h36b_a12.seq
AAGGTATAACTTTACCGATTTCAGCAGTAGCAGTTGCTTTGTCTTTTTTGA Majority
6160 6170 6180 6190 6200
6117 AAGGAACAACTTTACCGATTTCAGCAGTAGCAGTTGCTTTGTCACGTGCA 2603_a12.seq
6117 AAGGAACAACTTTACCGATTTCAGCAGTAGCAGTTGCTTTGTCACGTGCA 18rs21_a12.seq
6109 AIIITITAACATCACCAACAGAIGCICAGACAGTCICTTTGTCTTTTTIG 515_a12.seq
6108 AIICTITIIGTCACCAAITGICGCICAGACAACACCTTTGTCTTTTICG cjb111_a12.seq
6132 AAGGATTICTTGACCGATTTCAGCAGTAGCIITTGCTTTIICTGCCTGA h36b_a12.seq
TAGTTAGCATAATCTGCGCCAGCTGTCAAAAGTCTATTAACA-CTGTCAA Majority
6210 6220 6230 6240 6250
6167 TAATTAGCATAATTTGCGCCAGCTGTCAAAAGTCTATTAACAICTGTCAA 2603_a12.seq
6167 TAATTAGCATAATTTGCGCCAGCTGTCAAAAGTCTATTAACAICTGTCAA 18rs21_a12.seq
6159 TIGTTGGCATAATC---------- 515_a12.seq
6158 TIGTCAGTATAATC---------- cjb111_a12.seq
6182 TATTTGTCGTAGTCTGCGCCAGCTGTCAATAGGTTGAATTAICA-CTGTCAI h36b_a12.seq
T--TGTCAAGTTGTTTGTTTTAGCAAAGTTTTTATCTATTTGTGGTTTTT Majority
6260 6270 6280 6290 6300
6217 TGGTGTCAAATCGTTTGTTTTAGCAAAGTTTTTATCAATTTCTGGTTTTT 2603_a12.seq
6217 TGGTGTCAAATCGTTTGTTTTAGCAAAGTTTTTATCAATTTCTGGTTTTT 18rs21_a12.seq
6173 ---AAGTTGTTTTATCT--GCGAAGTTCTTATCTACTTGTGGTTTTG 515_a12.seq
6172 ---AAGATCTTTTATCT--GCAAAGTTCTTATCTACTTGTGGTTTTG cjb111_a12.seq
6231 T--CAACAATTTATTTGCTTTAGCAAAGTTTTTGTCGATTTGTGGTGCTT h36b_a12.seq

FIGURE 19M

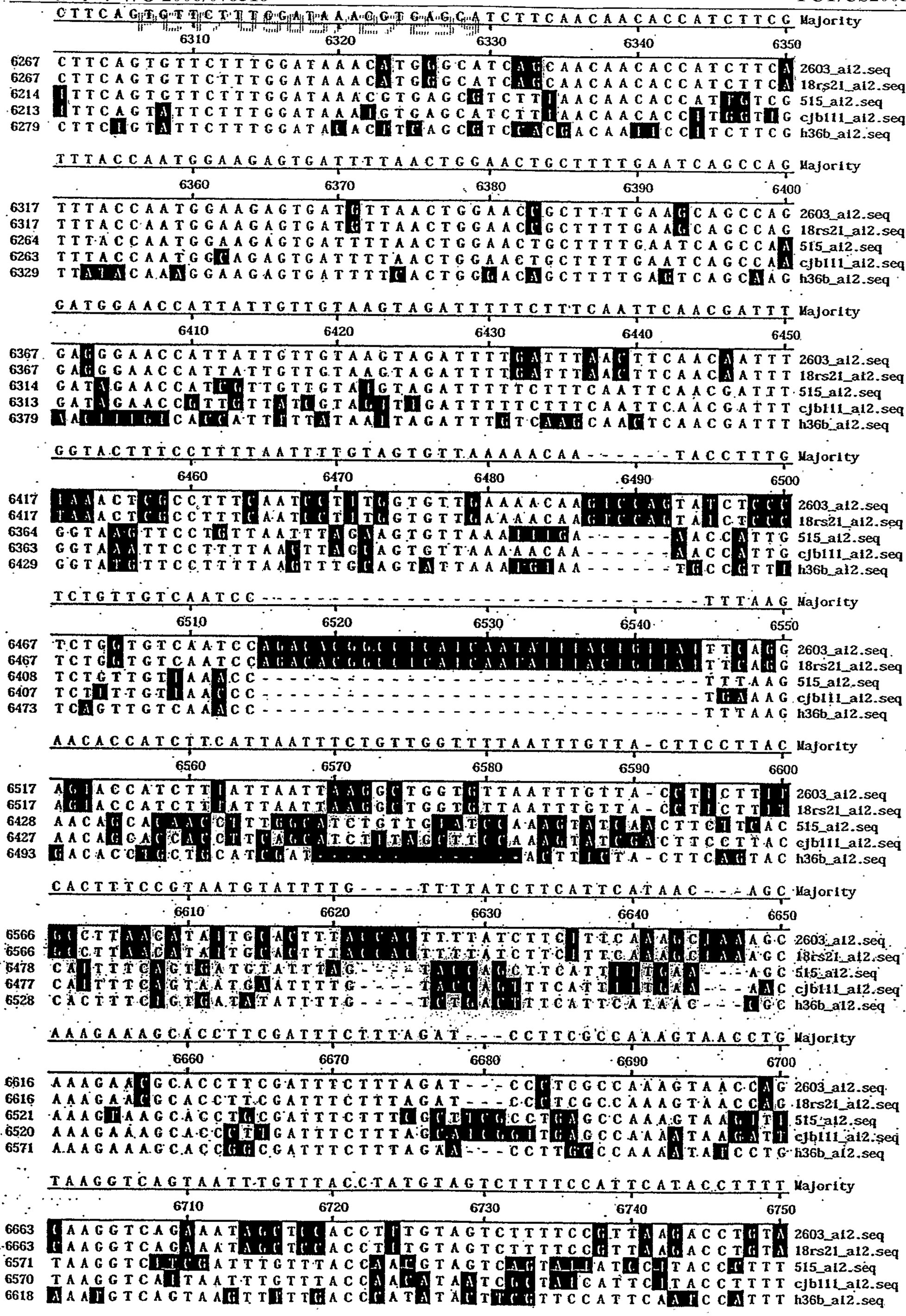

CTTCAGTGTTCTTTGGATAAACGTGAGCATCTTCAACAACACCATCTTCG Majority
6310 6320 6330 6340 6350
6267 CTTCAGTGTTCTTTGGATAAACATGGGCATCAGCAACAACACCATCTTCA 2603_a12.seq
6267 CTTCAGTGTTCTTTGGATAAACATGGGCATCAGCAACAACACCATCTTCA 18rs21_a12.seq
6214 ITTCAGTGTTCTTTGGATAAACGTGAGCGTCTTIAACAACACCATTGTCG 515_a12.seq
6213 ITTCAGTATTCTTTGGATAAAIGTGAGCATCTTIAACAACACCITGGTIG cjb111_a12.seq
6279 CTTCIGTATTCTTTGGATAIACITCAGCGTCCACGACAAIICCITCTTCG h36b_a12.seq
TTTACCAATGGAAGAGTGATTTTAACTGGAACTGCTTTTGAATCAGCCAG Majority
6360 6370 6380 6390 6400
6317 TTTACCAATGGAAGAGTGATGTTAACTGGAACCGCTTTTGAAGCAGCCAG 2603_a12.seq
6317 TTTACCAATGGAAGAGTGATGTTAACTGGAACTGCTTTTGAAGCAGCCAG 18rs21_a12.seq
6264 TTTACCAATGGAAGAGTGATTTTAACTGGAACTGCTTTTGAATCAGCCAA 515_a12.seq
6263 TTTACCAATGGCAGAGTGATTTTAACTGGAACTGCTTTTGAATCAGCCAA cjb111_a12.seq
6329 TTAIACAAAGGAAGAGTGATTTTCACTGGGACAGCTTTTGAGTCAGCAAG h36b_a12.seq
GATGGAACCATTATTGTTGTAAGTAGATTTTTCTTTCAATTCAACGATTT Majority
6410 6420 6430 6440 6450
6367 GAGGGAACCATTATTGTTGTAAGTAGATTTTGATTTAACTTCAACAATTT 2603_a12.seq
6367 GAGGGAACCATTATTGTTGTAAGTAGATTTTGATTTAACTTCAACAATTT 18rs21_a12.seq
6314 GATAGAACCATCGTTGTTGTAIGTAGATTTTTCTTTCAATTCAACGATTT 515_a12.seq
6313 GATAGAACCGTTGTTATIGTALITIGATTTTTCTTTCAATTCAACGATTT cjb111_a12.seq
6379 AACIIIGICACGATTTTTATAAITAGATTTGTCAAGCAACTCAACGATTT h36b_a12.seq
GGTACTTTCCTTTTAATTTTGTAGTGTTAAAAACAA------TACCTTTG Majority
6460 6470 6480 6490 6500
6417 IAAACTCGCCTTTCAATCCTITGGTGTTGAAAACAAGICCAGTAICTCCC 2603_a12.seq
6417 TAAACTCGCCTTTCAATICTITGGTGTTGAAAACAAGTCCAGTAICTCCC 18rs21_a12.seq
6364 GGTAAGTTCCTGTTAATTTAGAAGTGTTAAAIIIGA------AACCATTG 515_a12.seq
6363 GGTAAATTCCTTTTAAGTTAGTAGTGTTAAAAAACAA------AACCATTG cjb111_a12.seq
6429 GGTATGTTCCTTTTAAGTTTGCAGTATTAAAIIGIAA------TGCCGTTI h36b_a12.seq
TCTGTTGTCAATCC----------TTTAAG Majority
6510 6520 6530 6540 6550
6467 TCTGGTGTCAATCCAGACACGGCCICAICAATAIIIACIGIIAITTCAGG 2603_a12.seq
6467 TCTGGTGTCAATCCAGACACGGCCICAICAATAIIIACIGIIAITTCAGG 18rs21_a12.seq
6408 TCTGTTGTIAAACC----------TTTAAG 515_a12.seq
6407 TCTITTGTIAACC----------TGAAAG cjb111_a12.seq
6473 TCAGTTGTCAAACC----------TTTAAG h36b_a12.seq
AACACCATCTTCATTAATTTCTGTTGGTTTTAATTTGTTA-CTTCCTTAC Majority
6560 6570 6580 6590 6600
6517 AGIACCATCTTIATTAATTAAGGCTGGTGTTAATTTGTTA-CCTICTTII 2603_a12.seq
6517 AGIACCATCTTIATTAATTAAGGCTGGTGTTAATTTGTTA-CCTICTTII 18rs21_a12.seq
6428 AACAGCACAACCTITGGCATCTGTTGIATCCAAAGTATCAACTTCITCAC 515_a12.seq
6427 AACAGGACCACCTTCAGCATCTITAGCTTCCAAAGTATCGACTTCCTTAC cjb111_a12.seq
6493 GACACCTGCTGCATCGAT------------ACTTICTA-CTTCAGTAC h36b_a12.seq
CACTTTCCGTAATGTATTTTG----TTTTATCTTCATTCATAAC---AGC Majority
6610 6620 6630 6640 6650
6566 GCCTTAACATAITGCACTTTAICACTTTTATCTTCITTCAAAGCIAAAGCC 2603_a12.seq
6566 GCCTTAACATAITGCACTTTAICACTTTTATCTTCITTCAAAGCIAAAGCC 18rs21_a12.seq
6478 CAITTTCAGTGATGTATTTAG----TACCAGCTTCATTIITGAA---AGC 515_a12.seq
6477 CAITTTCAGTAATGAATTTTG----TACCAGTTTCATTIITGAA---AAC cjb111_a12.seq
6528 CACTTTCIGTGATATATTTTG----TCTGACTTTCATTCATAAC---CGC h36b_a12.seq
AAAGAAAGCACCTTCGATTTCTTTAGAT---CCTTCGCCAAAGTAACCTG Majority
6660 6670 6680 6690 6700
6616 AAAGAACGCACCTTCGATTTCTTTAGAT---CCCTCGCCAAAGTAACCAG 2603_a12.seq
6616 AAAGAACGCACCTTCGATTTCTTTAGAT---CCCTCGCCAAAGTAACCAG 18rs21_a12.seq
6521 AAAGTAAGCACCTGCGATTTCTTTTGCTTCGCCTGAGCCAAAGTAAGITI 515_a12.seq
6520 AAAGAAAGCACCCTIGATTTCTTTAGCAICGGITGAGCCAAAATAAGATI cjb111_a12.seq
6571 AAAGAAAGCACCGGCGATTTCTTTAGAA---CCTTGCCCAAAATAICCTG h36b_a12.seq
TAAGGTCAGTAATTTGTTTACCTATGTAGTCTTTTCCATTCATACCTTTT Majority
6710 6720 6730 6740 6750
6663 CAAGGTCAGAAATAGCTICACCTITGTAGTCTTTTCCGTTAAGACCTGTA 2603_a12.seq
6663 CAAGGTCAGAAATAGCTICACCTITGTAGTCTTTTCCGTTAAGACCTGTA 18rs21_a12.seq
6571 TAAGGTCITCGATTTGTTTACCAATGTAGTCAGTAIIATGCCITACCGTTT 515_a12.seq
6570 TAAGGTCAITAATTTGTTTACCAACATAATCGCTAICATTCITACCTTTT cjb111_a12.seq
6618 AAATGTCAGTAAGTTTTTGACCCATATATTCTTCGTTCCATTCAATCCATTT h36b_a12.seq

FIGURE 19N

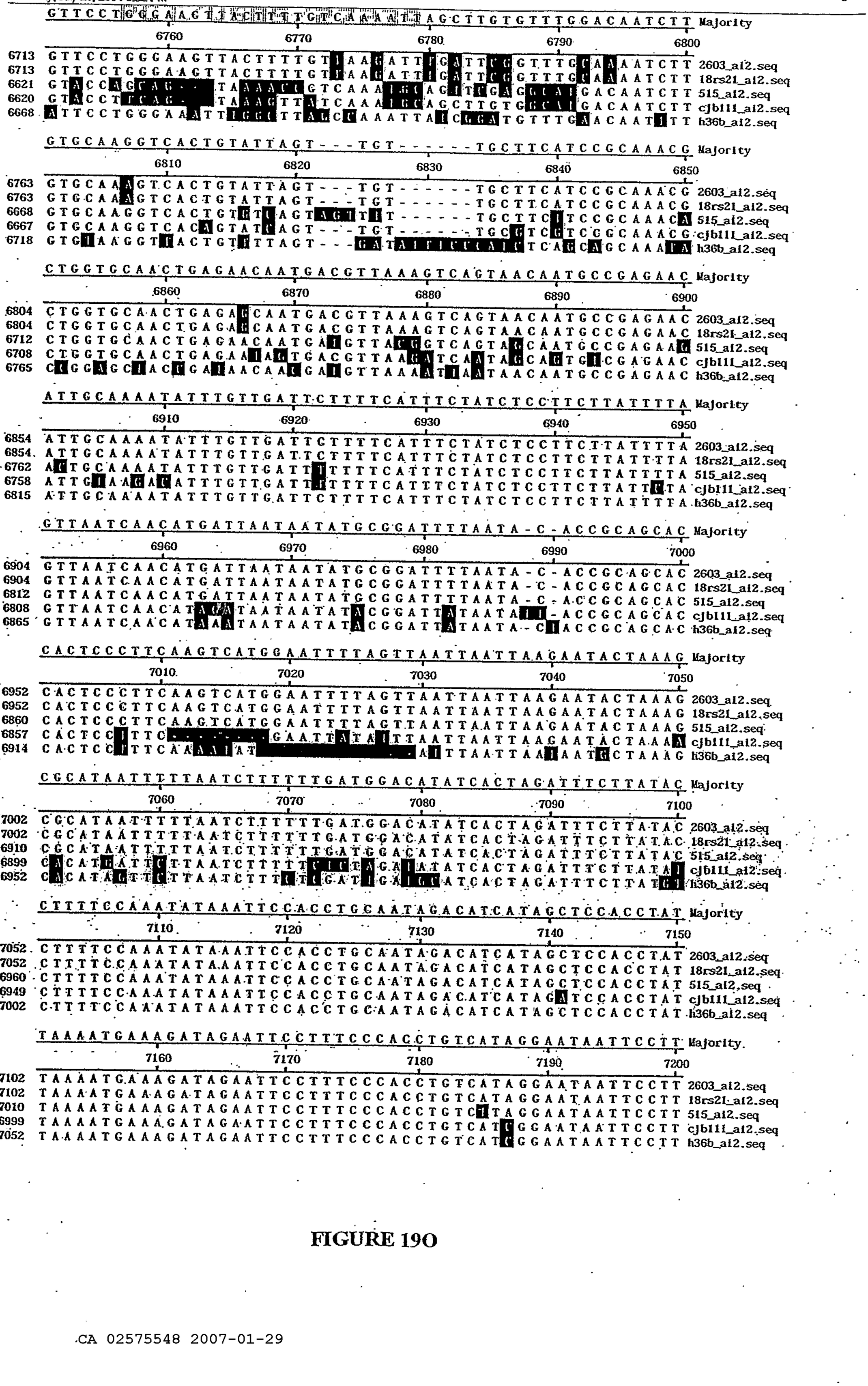

GTTCCTGGGAAGTTACTTTTGTCAAAATAGCTTGTGTTTGGACAATCTT Majority
6760 6770 6780 6790 6800
6713 GTTCCTGGGAAGTTACTTTTGTTAAGATTTGATTCGGTTTGCAAAATCTT 2603_a12.seq
6713 GTTCCTGGGAAGTTACTTTTGTTAAGATTTGATTCGGTTTGCAAAATCTT 18rs21_a12.seq
6621 GTACCAGCAG--TAAAACTGTCAAATGCAGTTCGAGGCATGACAATCTT 515_a12.seq
6620 GTACCTTCAG--TAAAGTTATCAAATGCAGCTTGTGGCATGACAATCTT cjb111_a12.seq
6668 ATTCCTGGGAAATTIGGCTTAGCTAAATTAICGGATGTTTGAACAATTTT h36b_a12.seq
GTGCAAGGTCACTGTATTAGT---TGT------TGCTTCATCCGCAAACG Majority
6810 6820 6830 6840 6850
6763 GTGCAAAGTCACTGTATTAGT---TGT------TGCTTCATCCGCAAACG 2603_a12.seq
6763 GTGCAAAGTCACTGTATTAGT---TGT------TGCTTCATCCGCAAACG 18rs21_a12.seq
6668 GTGCAAGGTCACTGTGTCAGTAGTTT------TGCTTCITCCGCAAACA 515_a12.seq
6667 GTGCAAGGTCACAGTATCAGT---TGT------TGCGTCGTCCGCAAACG cjb111_a12.seq
6718 GTGIAAGGTTACTGTTTTAGT---GATAIIICCAICTCAGCAGCAAAIA h36b_a12.seq
CTGGTGCAACTGAGAACAATGACGTTAAAGTCAGTAACAATGCCGAGAAC Majority
6860 6870 6880 6890 6900
6804 CTGGTGCAACTGAGAGCAATGACGTTAAAGTCAGTAACAATGCCGAGAAC 2603_a12.seq
6804 CTGGTGCAACTGAGAGCAATGACGTTAAAGTCAGTAACAATGCCGAGAAC 18rs21_a12.seq
6712 CTGGTGCAACTGAGAACAATGAIGTTACGGTCAGTAGCAATGCCGAGAAC 515_a12.seq
6708 CTGGTGCAACTGAGAAIAGTGACGTTAAGATCAATAGCAGTGICGAGAAG cjb111_a12.seq
6765 CGGAGCIACTGAIAACAACGAIGTTAAAATIAATAACAATGCCGAGAAC h36b_a12.seq
ATTGCAAAATATTTGTTGATTCTTTTCATTTCTATCTCCTTCTTATTTTA Majority
6910 6920 6930 6940 6950
6854 ATTGCAAAATATTTGTTGATTCTTTTCATTTCTATCTCCTTCTTATTTTA 2603_a12.seq
6854 ATTGCAAAATATTTGTTGATTCTTTTCATTTCTATCTCCTTCTTATTTTA 18rs21_a12.seq
6762 AGTGCAAAATATTTGTTGATTTTTTTCATTTCTATCTCCTTCTTATTTTA 515_a12.seq
6758 ATTGIAAGACATTTGTTGATTTTTTTCATTTCTATCTCCTTCTTATTTTA cjb111_a12.seq
6815 ATTGCAAAATATTTGTTGATTCTTTTCATTTCTATCTCCTTCTTATTTTA h36b_a12.seq
GTTAATCAACATGATTAATAATATGCGGATTTTAATA-C-ACCGCAGCAC Majority
6960 6970 6980 6990 7000
6904 GTTAATCAACATGATTAATAATATGCGGATTTTAATA-C-ACCGCAGCAC 2603_a12.seq
6904 GTTAATCAACATGATTAATAATATGCGGATTTTAATA-C-ACCGCAGCAC 18rs21_a12.seq
6812 GTTAATCAACATGATTAATAATATGCGGATTTTAATA-C-ACCGCAGCAC 515_a12.seq
6808 GTTAATCAACATAGAATAATAATATACGGATTATAATAII-ACCGCAGCAC cjb111_a12.seq
6865 GTTAATCAACATAAATAATAATATACGGATTATAATA-CIACCGCAGCAC h36b_a12.seq
CACTCCCTTCAAGTCATGGAATTTTAGTTAATTAATTAAGAATACTAAAG Majority
7010 7020 7030 7040 7050
6952 CACTCCCTTCAAGTCATGGAATTTTAGTTAATTAATTAAGAATACTAAAG 2603_a12.seq
6952 CACTCCCTTCAAGTCATGGAATTTTAGTTAATTAATTAAGAATACTAAAG 18rs21_a12.seq
6860 CACTCCCTTCAAGTCATGGAATTTTAGTTAATTAATTAAGAATACTAAAG 515_a12.seq
6857 CACTCCITTC--------GAATTATAITTAATTAATTAAGAATACTAAAA cjb111_a12.seq
6914 CACTCCITTCAAAAIAT----------AITTAATTAAIAATGCTAAAG h36b_a12.seq
CGCATAATTTTTAATCTTTTTTGATGGACATATCACTAGATTTCTTATAC Majority
7060 7070 7080 7090 7100
7002 CGCATAATTTTTAATCTTTTTTGATGGACATATCACTAGATTTCTTATAC 2603_a12.seq
7002 CGCATAATTTTTAATCTTTTTTGATGGACATATCACTAGATTTCTTATAC 18rs21_a12.seq
6910 CGCATAATTTTTAATCTTTTTTGATGGACATATCACTAGATTTCTTATAC 515_a12.seq
6899 CACATAATTCTTAATCTTTTTCICTAGAIATATCACTAGATTTCTTATAI cjb111_a12.seq
6952 CACATAGTTCTTAATCTTTTTTGATIGAIGTATCACTAGATTTCTTATGI h36b_a12.seq
CTTTTCCAAATATAAATTCCACCTGCAATAGACATCATAGCTCCACCTAT Majority
7110 7120 7130 7140 7150
7052 CTTTTCCAAATATAAATTCCACCTGCAATAGACATCATAGCTCCACCTAT 2603_a12.seq
7052 CTTTTCCAAATATAAATTCCACCTGCAATAGACATCATAGCTCCACCTAT 18rs21_a12.seq
6960 CTTTTCCAAATATAAATTCCACCTGCAATAGACATCATAGCTCCACCTAT 515_a12.seq
6949 CTTTTCCAAATATAAATTCCACCTGCAATAGACATCATAGATCCACCTAT cjb111_a12.seq
7002 CTTTTCCAAATATAAATTCCACCTGCAATAGACATCATAGCTCCACCTAT h36b_a12.seq
TAAAATGAAAGATAGAATTCCTTTCCCACCTGTCATAGGAATAATTCCTT Majority
7160 7170 7180 7190 7200
7102 TAAAATGAAAGATAGAATTCCTTTCCCACCTGTCATAGGAATAATTCCTT 2603_a12.seq
7102 TAAAATGAAAGATAGAATTCCTTTCCCACCTGTCATAGGAATAATTCCTT 18rs21_a12.seq
7010 TAAAATGAAAGATAGAATTCCTTTCCCACCTGTCITAGGAATAATTCCTT 515_a12.seq
6999 TAAAATGAAAGATAGAATTCCTTTCCCACCTGTCATTGGAATAATTCCTT cjb111_a12.seq
7052 TAAAATGAAAGATAGAATTCCTTTCCCACCTGTCATGGGAATAATTCCTT h36b_a12.seq

FIGURE 19O

```
                TTGGTGGAATATCGCGTGTTGGTAATTAAATGCTTGTCACCTTCCTCATGA Majority
                        7210      7220      7230      7240      7250
7152  TTGGTGGAATATGCGTGTTGGTAATTAAATGCTTGTCACCTTCCTCATGA 2603_a12.seq
7152  TTGGTGGAATATGCCGTGTTGGTAATTAAATGCTTGTCACCTTCCTCATGA 18rs21_a12.seq
7060  TTGCTGGAATATGCGTGTTGGTAATTAAATGCTTGTCACCTTCCTCATGA 515_a12.seq
7049  TTGGTGGAATATGCGTGTTGGTAATTAAATGCTTGTCACCTTCCTCATGA cjb111_a12.seq
7102  TTGGTGGAATATGCGTGTTGGTAATTAAATGCTTGTCACCTTCCTCATGA h36b_a12.seq

                TATTCAGAAATCTGTTTATTAACAGCTATTATATTTTTTATCGATCCTTT Majority
                        7260      7270      7280      7290      7300
7202  TATTCAGAAATCTGTTTATTAACAGCTATTATATTTTTTATCGATCCTTT 2603_a12.seq
7202  TATTCAGAAATCTGTTTATTAACAGCTATTATATTTTTTATCGATCCTTT 18rs21_a12.seq
7110  TATTCAGAAATCTGTTTATTAACAGCTATTATATTTTTTATCGATCCTTT 515_a12.seq
7099  TATTCAGAAATCTGTTTATTAACAGCTATTATATTTTGTATCGATCCTTT cjb111_a12.seq
7152  TATTCAGAAATCTGTTTATTAACAGCTATTATATTTTGTATCGATCCTTT h36b_a12.seq

                AACCACTTCAAAAGTTAAAATTGGTTTATTAGTAATTTTTTGATAATCCT Majority
                        7310      7320      7330      7340      7350
7252  AACCACTTCAAAAGTTAAAATTGGTTTATTAGTAATTTTTTGATAATCCT 2603_a12.seq
7252  AACCACTTCAAAAGTTAAAATTGGTTTATTAGTAATTTTTTGATAATCCT 18rs21_a12.seq
7160  AACCACTTCAAAAGTTAAAATTGGTTTATTAGTAATTTTTTGATAATCCT 515_a12.seq
7149  AACAACTTCAAAAGTTAAAATTGGTTTATTAGTAATTTTTTGATAATCCT cjb111_a12.seq
7202  AACAACTTCAAAAGTTAAAATTGGTTTATTAGTAATTTTTTGATAATCCT h36b_a12.seq

                CCGGCGAAACTGCTTCTATTAACTGATATTTGCCATCTTTCAAATCTTTG Majority
                        7360      7370      7380      7390      7400
7302  CCGGCGAAACTGCTTCTATTAACTGATATTTGCCATCTTTCAAATCTTTG 2603_a12.seq
7302  CCGGCGAAACTGCTTCTATTAACTGATATTTGCCATCTTTCAAATCTTTG 18rs21_a12.seq
7210  CCGGCGAAACTGCTTCTATTAACTGATATTTGCCATCTTTCAAATCTTTG 515_a12.seq
7199  CCGGCGAAAGTGCTTCTATTAACTGATATTTGCCATCTTTCAAATCTTTG cjb111_a12.seq
7252  CCGGCGAAACTGCTTCTATTAACTGATATTTGCCATCTTTCAAATCTTTG h36b_a12.seq

                TAAGAAATTTTGCCGTTTTCTCCCGTCACTACTTTTGAATTATTATTTTTT Majority
                        7410      7420      7430      7440      7450
7352  TAAGAAATTTTGCCGTTTTCTCCCGTCACTACTTTTGAATTATTATTTTTT 2603_a12.seq
7352  TAAGAAATTTTGCCGTTTTCTCCCGTCACTACTTTTGAATTATTATTTTTT 18rs21_a12.seq
7260  TAAGAAATTTTGCCGTTTTCTCCCGTCACTACTTTTGAATTATTATTTTTT 515_a12.seq
7249  TAAGAAATTTTGCCGTTTTCTCCCGTCACTACTTTTGAATTATTATTTTTT cjb111_a12.seq
7302  TAAGAAATTTTGCCGTTTTCTCCCGTCACTACTTTTGAATTATTATTTTTT h36b_a12.seq

                TATTGGTAAATAAAGTTTATAATCTTCATTAAATTCTTGAAGTTCAAACG Majority
                        7460      7470      7480      7490      7500
7402  TATTGGTAAATAAAGTTTATAATCTTCATTAAATTCTTGAAGTTCAAACG 2603_a12.seq
7402  TATTGGTAAATAAAGTTTATAATCTTCATTAAATTCTTGAAGTTCAAACG 18rs21_a12.seq
7310  TATTGGTAAATAAAGTTTATAATCTTCATTAAATTCTTGAAGTTCAAACG 515_a12.seq
7299  TATTGGTAAATAAAGTTTATAATCTTCATTAAATTCTTGAAGTTCAAACG cjb111_a12.seq
7352  TATTGGTAAATAAAGTTTATAATCTTCATTAAATTCTTGAAGTTCAAAYG h36b_a12.seq

                TAGCTCCTTTGAGAAGCAACTTATTATTATCTTTATCAACTTTTATAAAT Majority
                        7510      7520      7530      7540      7550
7452  TAGCTCCTTTGAGAAGCAACTTATTATTATCTTTATCAACTTTTATAAAT 2603_a12.seq
7452  TAGCTCCTTTGAGAAGCAAYTTATTATTATCTTTATCAACTTTTATAAAT 18rs21_a12.seq
7360  TAGCTCCTTTGAGRAGCAACTTATTATTATCTTTATCAACTTTTATAAAT 515_a12.seq
7349  TAGCTCCTTTGAGAAGCAACTTATTATTATCTTTATCAACTTTTGTAAAT cjb111_a12.seq
7402  TAGCTCCTTTGAGAAGCAACTTATTATTATCTTTATCAACTTTTGTAAAT h36b_a12.seq

                TCAATTTCACCTAACTTCTTCTCGTTTTTAATCGTTATTGTAGGATATTC Majority
                        7560      7570      7580      7590      7600
7502  TCAATTTCACCTAACTTCTTCTCGTTTTTAATCGTTATTGTAGGATATTC 2603_a12.seq
7502  TCAATTTCACCTAACTTCTTCTCGTTTTTAATCGTTATTGTAGGATATTC 18rs21_a12.seq
7410  TCAATTTCACCTAACTTCTTCTCGTTTTTAATCGTTATTGTAGGATATTC 515_a12.seq
7399  TCAATTTCACCTAACTTCTTCTCGTTTTTAATCGTTATTGTAGGATATTC cjb111_a12.seq
7452  TCAATTTCACCTAACTTCTTCTCGTTTTTAATCGTTATTGTAGGATATTC h36b_a12.seq

                TCTCACATCACGAATTTTAGGGATTGGAAAATCTCTAAGTGTATTAGGAT Majority
                        7610      7620      7630      7640      7650
7552  TCTCACATCACGAATTTTAGGGATTGGAAAATCTCTAAGTGTATTAGGAT 2603_a12.seq
7552  TCTCACATCACGAATTTTAGGGATTGGAAAATCTCTAAGTGTATTAGGAT 18rs21_a12.seq
7460  TCTCACATCACGAATTTTAGGGATTGGAAAATCTCTAAGTGTATTAGGAT 515_a12.seq
7449  TCTCACATCACGAATTTTAGGGATTGGAAAATCTCTAAGTGTATTAGGAT cjb111_a12.seq
7502  TCTCACATCACGAATTTTAGGGATTGGAAAATCTCTAAGTGTATYRGRYT h36b_a12.seq
```

FIGURE 19P

```
                                                
```

ACGACCTCATTAGTTCTGTGATTTCTCCATCTGAAAGTTTAAAAGCTTCC Majority
8110 8120 8130 8140 8150
8052 ACGACCTCATTAGTTCTGTGATTTCTCCATCTGAAAGTTTAAAAGCTTCC 2603_a12.seq
8052 ACGACCTCATTAGTTCTGTGATTTCTCCATCTGAAAGTTTAAAAGCTTCC 18rs21_a12.seq
7960 ACGACCTCATTAGTTCTGTGATTTCTCCATCTGAAAGTTTAAAAGCTTCC 515_a12.seq
7949 AYGACYTCATTAGTTCTGTYATTTCYCCATCTGAAAGTTYAAAAGCTTCC cjb111_a12.seq
8002 AYGAATTCATTAGTTCTGTYATTTCYCCATCTGAAAGTTYAAAAGCTTCC h36b_a12.seq
TCTTTCAATTTTTGAAAAGTACCATCTTGATTTTTCTTATACTCCTCATT Majority
8160 8170 8180 8190 8200
8102 TCTTTCAATTTTTGAAAAGTACCATCTTGATTTTTCTTATACTCCTCATT 2603_a12.seq
8102 TCTTTCAATTTTTGAAAAGTACCATCTTGATTTTTCTTATACTCCTCATT 18rs21_a12.seq
8010 TCTTTCAATTTTTGAAAAGTACCATCTTGATTTTTCTTATACTCCTCATT 515_a12.seq
7999 TCTTTCAATTTTTGAAAAGTACCATCTTGATTTTTCTTATAATCCTCATT cjb111_a12.seq
8052 TCYTTCAATTTTTGAAAAGTACCATCTTGATTTTTCTTATAATCCTCATT h36b_a12.seq
ATAAACTTGTCTAAAACCAGATATATCGATACCAAAATTAAAAATGTCAT Majority
8210 8220 8230 8240 8250
8152 ATAAACTTGTCTAAAACCAGATATATCGATACCAAAATTAAAAATGTCAT 2603_a12.seq
8152 ATAAACTTGTCTAAAACCAGATATATCGATACCAAAATTAAAAATGTCAT 18rs21_a12.seq
8060 ATAAACTTGTCTAAAACCAGATATATCGATACCAAAATTAAAAATGTCAT 515_a12.seq
8049 ATAAACTTGTCTAAAACCAGATATATCYATACCAAAATTAAAGATGTCAT cjb111_a12.seq
8102 ATAAACTTGTCTAAAACCAGATATATCTATACCAAAATTAAAGATGTCAT h36b_a12.seq
AATTTTTCTGTTTTAAACTATTTATATAAAGTTTGGTTGGTGTTCCATGT Majority
8260 8270 8280 8290 8300
8202 AATTTTTCTGTTTTAAACTATTTATATAAAGTTTGGTTGGTGTTCCATGT 2603_a12.seq
8202 AATTTTTCTGTTTTAAACTATTTATATAAAGTTTGGTTGGTGTTCCATGT 18rs21_a12.seq
8110 AATTTTTCTGTTTTAAACTATTTATATAAAGTTTGGTTGGTGTTCCATGT 515_a12.seq
8099 AATTTYTCTGTTTTAAACTATTTATATAAAGTTTGGTTGGTGTTCCATGT cjb111_a12.seq
8152 AATTTTTCTGTTTTAAACTATTTATATAAAGTTTGGTTGGTGTTCCATGT h36b_a12.seq
TCTTTCACTGGTCCATTTCGATAAATTGTACCTTTAGGGTAATTAAGATT Majority
8310 8320 8330 8340 8350
8252 TCTTTCACTGGTCCATTTCGATAAATTGTACCTTTAGGGTAATTAAGATT 2603_a12.seq
8252 TCTTTCACTGGTCCATTTCGATAAATTGTACCTTTAGGGTAATTAAGATT 18rs21_a12.seq
8160 TCTTTCACTGGTCCATTTCGATAAATTGTACCTTTAGGGTAATTAAGATT 515_a12.seq
8149 TCYTYTACTGGTCCATTTCGATAAAATGTACCTTTAGGGTAATTAAGATT cjb111_a12.seq
8202 TCYTYTACTGGTCCATTTCTATAAATTGTACCTTTAGGGTAATTAAGATT h36b_a12.seq
TAAATCTAAATAATGAAGTTTTTGTAAGTTTCCAGAGATTATCTGTGTTT Majority
8360 8370 8380 8390 8400
8302 TAAATCTAAATAATGAAGTTTTTGTAAGTTTCCAGAGATTATCTGTGTTT 2603_a12.seq
8302 TAAATCTAAATAATGAAGTTTTTGTAAGTTTCCAGAGATTATCTGTGTTT 18rs21_a12.seq
8210 TAAATCTAAATAATGAAGTTTTTGTAAGTTTCCAGAGATTATCTGTGTTT 515_a12.seq
8199 TAAATCTAAATAATGAAGTTTTTGTAAGTTTCCAGAGATTATCTGTGTTT cjb111_a12.seq
8252 TAAATCTAAATAATGAAGTTTTTGTAAGTTTCCAGAAATTATCTGTGTTT h36b_a12.seq
GATAACTATCTAAGGGAAACAAAAAGTAACTCTCCCCATTTCCTTTTATA Majority
8410 8420 8430 8440 8450
8352 GATAACTATCTAAGGGAAACAAAAAGTAACTCTCCCCATTTCCTTTTATA 2603_a12.seq
8352 GATAACTATCTAAGGGAAACAAAAAGTAACTCTCCCCATTTCCTTTTATA 18rs21_a12.seq
8260 GATAACTATCTAAGGGAAATAAAAAGTAACTCTCCCCATTTCCTTTTATA 515_a12.seq
8249 GATAACTATCTAAGGGAAACAAAAAGTAACTCTCCCCATTYCCTTTTATA cjb111_a12.seq
8302 GATAACTATCTAAGGGAAACAAAAAGTAACTCTCCCCATTGCCTTTGATC h36b_a12.seq
TCCTCGGGCTTATCAGTAAGTAGAAAATTACTTTTATTTAGATATCCATT Majority
8460 8470 8480 8490 8500
8402 TCCTCGGGCTTATCAGTAAGTAGAAAATTACTTTTATTTAGATATCCATT 2603_a12.seq
8402 TCCTCGGGCTTATCAGTAAGTAGAAAATTACTTTTATTTAGATATCCATT 18rs21_a12.seq
8310 TCATCGGGCTTATCAGTAAGTAGAAAATTACTTTTATTTAGATATCCATT 515_a12.seq
8299 TCCTCGGGCTTATCAGTAAGTAGAAAATTACTTTTATTTAGATATCCATT cjb111_a12.seq
8352 TYYTCYGGATCATCAGTYAYAAAYTACTTTTATTTAGATATCCATT h36b_a12.seq
TTTTTTCATTTGTTCAAATTGGCTTTCATATGATGCACCCAGTTTAAAAT Majority
8510 8520 8530 8540 8550
8452 TTTTTTCATTTGTTCAAATTGGCTTTCATATGATGCACCCAGTTTAAAAT 2603_a12.seq
8452 TTTTTTCATTTGTTCAAATTGGCTTTCATATGATGCACCCAGTTTAAAAT 18rs21_a12.seq
8360 TTTTTTCATTTGTTCAAATTGGCTTTCATATGATGCACCCAGTTTAAAAT 515_a12.seq
8349 TTTTTTCATTTGTTCAAATTGGCTTTCATATGATGCACCCAGTTTAAAAT cjb111_a12.seq
8402 TTCTTTTYATTYCYCTCAATTGAYTTGCGTATGYTGAACCCYTTTACAAAAC h36b_a12.seq

FIGURE 19R

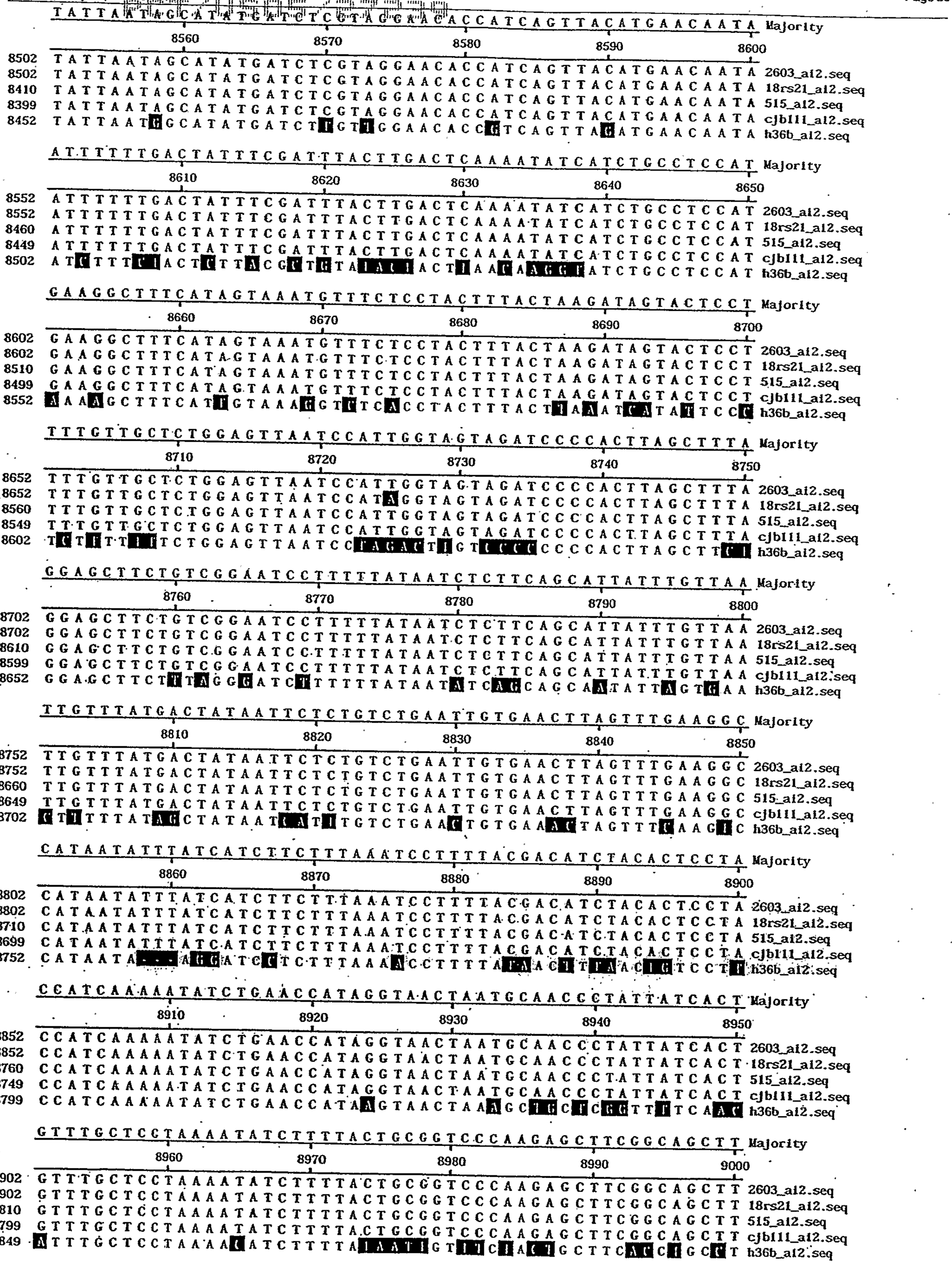

TATTAATAGCATATGATCTCGTAGGAACACCATCAGTTACATGAACAATA Majority
8560 8570 8580 8590 8600
8502 TATTAATAGCATATGATCTCGTAGGAACACCATCAGTTACATGAACAATA 2603_a12.seq
8502 TATTAATAGCATATGATCTCGTAGGAACACCATCAGTTACATGAACAATA 18rs21_a12.seq
8410 TATTAATAGCATATGATCTCGTAGGAACACCATCAGTTACATGAACAATA 515_a12.seq
8399 TATTAATAGCATATGATCTCGTAGGAACACCATCAGTTACATGAACAATA cjb111_a12.seq
8452 TATTAATGGCATATGATCTTGTTGGAACACCGTCAGTTAGATGAACAATA h36b_a12.seq
ATTTTTTGACTATTTCGATTTACTTGACTCAAAATATCATCTGCCTCCAT Majority
8610 8620 8630 8640 8650
8552 ATTTTTTGACTATTTCGATTTACTTGACTCAAAATATCATCTGCCTCCAT 2603_a12.seq
8552 ATTTTTTGACTATTTCGATTTACTTGACTCAAAATATCATCTGCCTCCAT 18rs21_a12.seq
8460 ATTTTTTGACTATTTCGATTTACTTGACTCAAAATATCATCTGCCTCCAT 515_a12.seq
8449 ATTTTTTGACTATTTCGATTTACTTGACTCAAAATATCATCTGCCTCCAT cjb111_a12.seq
8502 ATCTTTTCTACTCTTACGCTGTAIACTIAAGAAGGTATCTGCCTCCAT h36b_a12.seq
GAAGGCTTTCATAGTAAATGTTTCTCCTACTTTACTAAGATAGTACTCCT Majority
8660 8670 8680 8690 8700
8602 GAAGGCTTTCATAGTAAATGTTTCTCCTACTTTACTAAGATAGTACTCCT 2603_a12.seq
8602 GAAGGCTTTCATAGTAAATGTTTCTCCTACTTTACTAAGATAGTACTCCT 18rs21_a12.seq
8510 GAAGGCTTTCATAGTAAATGTTTCTCCTACTTTACTAAGATAGTACTCCT 515_a12.seq
8499 GAAGGCTTTCATAGTAAATGTTTCTCCTACTTTACTAAGATAGTACTCCT cjb111_a12.seq
8552 AAAAGCTTTCATTGTAAAGGTTTCACCTACTTTACTTAATCATATTCCC h36b_a12.seq
TTTGTTGCTCTGGAGTTAATCCATTGGTAGTAGATCCCCACTTAGCTTTA Majority
8710 8720 8730 8740 8750
8652 TTTGTTGCTCTGGAGTTAATCCATTGGTAGTAGATCCCCACTTAGCTTTA 2603_a12.seq
8652 TTTGTTGCTCTGGAGTTAATCCATAGGTAGTAGATCCCCACTTAGCTTTA 18rs21_a12.seq
8560 TTTGTTGCTCTGGAGTTAATCCATTGGTAGTAGATCCCCACTTAGCTTTA 515_a12.seq
8549 TTTGTTGCTCTGGAGTTAATCCATTGGTAGTAGATCCCCACTTAGCTTTA cjb111_a12.seq
8602 TCTTTTTTCTGGAGTTAATCCTAGACTTGTCCCCCCCACTTAGCTTTC h36b_a12.seq
GGAGCTTCTGTCGGAATCCTTTTTATAATCTCTTCAGCATTATTTGTTAA Majority
8760 8770 8780 8790 8800
8702 GGAGCTTCTGTCGGAATCCTTTTTATAATCTCTTCAGCATTATTTGTTAA 2603_a12.seq
8702 GGAGCTTCTGTCGGAATCCTTTTTATAATCTCTTCAGCATTATTTGTTAA 18rs21_a12.seq
8610 GGAGCTTCTGTCGGAATCCTTTTTATAATCTCTTCAGCATTATTTGTTAA 515_a12.seq
8599 GGAGCTTCTGTCGGAATCCTTTTTATAATCTCTTCAGCATTATTTGTTAA cjb111_a12.seq
8652 GGAGCTTCTTTAGGGATCTTTTTTATAATACTCAGCAGCAATATTAGTGAA h36b_a12.seq
TTGTTTATGACTATAATTCTCTGTCTGAATTGTGAACTTAGTTTGAAGGC Majority
8810 8820 8830 8840 8850
8752 TTGTTTATGACTATAATTCTCTGTCTGAATTGTGAACTTAGTTTGAAGGC 2603_a12.seq
8752 TTGTTTATGACTATAATTCTCTGTCTGAATTGTGAACTTAGTTTGAAGGC 18rs21_a12.seq
8660 TTGTTTATGACTATAATTCTCTGTCTGAATTGTGAACTTAGTTTGAAGGC 515_a12.seq
8649 TTGTTTATGACTATAATTCTCTGTCTGAATTGTGAACTTAGTTTGAAGGC cjb111_a12.seq
8702 CTTTTTATAGCTATAATCATTGTCTGAACTGTGAAACTAGTTTCAAGIC h36b_a12.seq
CATAATATTTATCATCTTCTTTAAATCCTTTTACGACATCTACACTCCTA Majority
8860 8870 8880 8890 8900
8802 CATAATATTTATCATCTTCTTTAAATCCTTTTACGACATCTACACTCCTA 2603_a12.seq
8802 CATAATATTTATCATCTTCTTTAAATCCTTTTACGACATCTACACTCCTA 18rs21_a12.seq
8710 CATAATATTTATCATCTTCTTTAAATCCTTTTACGACATCTACACTCCTA 515_a12.seq
8699 CATAATATTTATCATCTTCTTTAAATCCTTTTACGACATCTACACTCCTA cjb111_a12.seq
8752 CATAATA---AGGATCGCTCTTTAAAACCTTTTACAACTTAACIGTCCTT h36b_a12.seq
CCATCAAAAATATCTGAACCATAGGTAACTAATGCAACCCTATTATCACT Majority
8910 8920 8930 8940 8950
8852 CCATCAAAAATATCTGAACCATAGGTAACTAATGCAACCCTATTATCACT 2603_a12.seq
8852 CCATCAAAAATATCTGAACCATAGGTAACTAATGCAACCCTATTATCACT 18rs21_a12.seq
8760 CCATCAAAAATATCTGAACCATAGGTAACTAATGCAACCCTATTATCACT 515_a12.seq
8749 CCATCAAAAATATCTGAACCATAGGTAACTAATGCAACCCTATTATCACT cjb111_a12.seq
8799 CCATCAAAAATATCTGAACCATAAGTAACTAAAGCTGCTCGGTTTTCAAC h36b_a12.seq
GTTTGCTCCTAAAATATCTTTTACTGCGGTCCCAAGAGCTTCGGCAGCTT Majority
8960 8970 8980 8990 9000
8902 GTTTGCTCCTAAAATATCTTTTACTGCGGTCCCAAGAGCTTCGGCAGCTT 2603_a12.seq
8902 GTTTGCTCCTAAAATATCTTTTACTGCGGTCCCAAGAGCTTCGGCAGCTT 18rs21_a12.seq
8810 GTTTGCTCCTAAAATATCTTTTACTGCGGTCCCAAGAGCTTCGGCAGCTT 515_a12.seq
8799 GTTTGCTCCTAAAATATCTTTTACTGCGGTCCCAAGAGCTTCGGCAGCTT cjb111_a12.seq
8849 ATTTGCTCCTAAAACATCTTTTAIAATIGTITCIACTGCTTCACCIGCCT h36b_a12.seq

FIGURE 19S

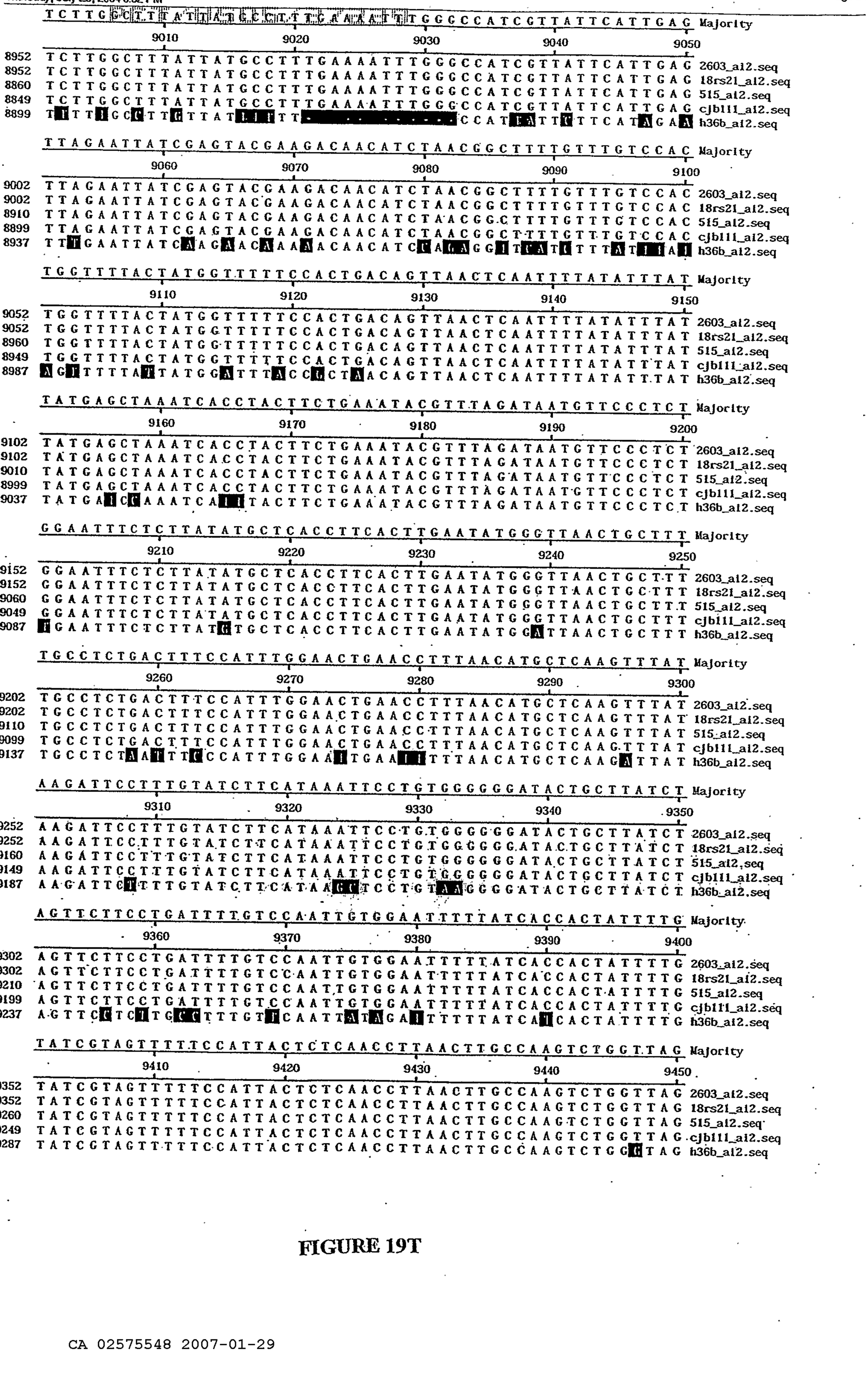

TCTTGGCTTTATTATGCCTTTGAAAATTTGGGCCATCGTTATTCATTGAG Majority
9010 9020 9030 9040 9050
8952 TCTTGGCTTTATTATGCCTTTGAAAATTTGGGCCATCGTTATTCATTGAG 2603_a12.seq
8952 TCTTGGCTTTATTATGCCTTTGAAAATTTGGGCCATCGTTATTCATTGAG 18rs21_a12.seq
8860 TCTTGGCTTTATTATGCCTTTGAAAATTTGGGCCATCGTTATTCATTGAG 515_a12.seq
8849 TCTTGGCTTTATTATGCCTTTGAAAATTTGGGCCATCGTTATTCATTGAG cjb111_a12.seq
8899 TTTTGCCTTGTTATTTT CCATTGTTCATAGAA h36b_a12.seq
TTAGAATTATCGAGTACGAAGACAACATCTAACGGCTTTTGTTTGTCCAC Majority
9060 9070 9080 9090 9100
9002 TTAGAATTATCGAGTACGAAGACAACATCTAACGGCTTTTGTTTGTCCAC 2603_a12.seq
9002 TTAGAATTATCGAGTACGAAGACAACATCTAACGGCTTTTGTTTGTCCAC 18rs21_a12.seq
8910 TTAGAATTATCGAGTACGAAGACAACATCTAACGGCTTTTGTTTGTCCAC 515_a12.seq
8899 TTAGAATTATCGAGTACGAAGACAACATCTAACGGCTTTTGTTTGTCCAC cjb111_a12.seq
8937 TTTGAATTATCAAGAACAAAACAACATCCAGGGTTAATTTTATTTAT h36b_a12.seq
TGGTTTTACTATGGTTTTTCCACTGACAGTTAACTCAATTTTATATTTAT Majority
9110 9120 9130 9140 9150
9052 TGGTTTTACTATGGTTTTTCCACTGACAGTTAACTCAATTTTATATTTAT 2603_a12.seq
9052 TGGTTTTACTATGGTTTTTCCACTGACAGTTAACTCAATTTTATATTTAT 18rs21_a12.seq
8960 TGGTTTTACTATGGTTTTTCCACTGACAGTTAACTCAATTTTATATTTAT 515_a12.seq
8949 TGGTTTTACTATGGTTTTTCCACTGACAGTTAACTCAATTTTATATTTAT cjb111_a12.seq
8987 AGTTTTTATTATGGATTTACCCTACACAGTTAACTCAATTTTATATTTAT h36b_a12.seq
TATGAGCTAAATCACCTACTTCTGAAATACGTTTAGATAATGTTCCCTCT Majority
9160 9170 9180 9190 9200
9102 TATGAGCTAAATCACCTACTTCTGAAATACGTTTAGATAATGTTCCCTCT 2603_a12.seq
9102 TATGAGCTAAATCACCTACTTCTGAAATACGTTTAGATAATGTTCCCTCT 18rs21_a12.seq
9010 TATGAGCTAAATCACCTACTTCTGAAATACGTTTAGATAATGTTCCCTCT 515_a12.seq
8999 TATGAGCTAAATCACCTACTTCTGAAATACGTTTAGATAATGTTCCCTCT cjb111_a12.seq
9037 TATGAICCAAATCAITACTTCTGAAATACGTTTAGATAATGTTCCCTCT h36b_a12.seq
GGAATTTCTCTTATATGCTCACCTTCACTTGAATATGGGTTAACTGCTTT Majority
9210 9220 9230 9240 9250
9152 GGAATTTCTCTTATATGCTCACCTTCACTTGAATATGGGTTAACTGCTT 2603_a12.seq
9152 GGAATTTCTCTTATATGCTCACCTTCACTTGAATATGGGTTAACTGCTTT 18rs21_a12.seq
9060 GGAATTTCTCTTATATGCTCACCTTCACTTGAATATGGGTTAACTGCTTT 515_a12.seq
9049 GGAATTTCTCTTATATGCTCACCTTCACTTGAATATGGGTTAACTGCTTT cjb111_a12.seq
9087 TGAATTTCTCTTATGTGCTCACCTTCACTTGAATATGGATTAACTGCTTT h36b_a12.seq
TGCCTCTGACTTTCCATTTGGAACTGAACCTTTAACATGCTCAAGTTTAT Majority
9260 9270 9280 9290 9300
9202 TGCCTCTGACTTTCCATTTGGAACTGAACCTTTAACATGCTCAAGTTTAT 2603_a12.seq
9202 TGCCTCTGACTTTCCATTTGGAACTGAACCTTTAACATGCTCAAGTTTAT 18rs21_a12.seq
9110 TGCCTCTGACTTTCCATTTGGAACTGAACCTTTAACATGCTCAAGTTTAT 515_a12.seq
9099 TGCCTCTGACTTTCCATTTGGAACTGAACCTTTAACATGCTCAAGTTTAT cjb111_a12.seq
9137 TGCCTCTAATTTGCCCATTTGGAAITGAAIITTTAACATGCTCAAGATTAT h36b_a12.seq
AAGATTCCTTTGTATCTTCATAAATTCCTGTGGGGGGATACTGCTTATCT Majority
9310 9320 9330 9340 9350
9252 AAGATTCCTTTGTATCTTCATAAATTCCTGTGGGGGGATACTGCTTATCT 2603_a12.seq
9252 AAGATTCCTTTGTATCTTCATAAATTCCTGTGGGGGGATACTGCTTATCT 18rs21_a12.seq
9160 AAGATTCCTTTGTATCTTCATAAATTCCTGTGGGGGGATACTGCTTATCT 515_a12.seq
9149 AAGATTCCTTTGTATCTTCATAAATTCCTGTGGGGGGATACTGCTTATCT cjb111_a12.seq
9187 AAGATTCTTTTGTATCTTCATAAGCTCCTGTAAGGGGATACTGCTTATCT h36b_a12.seq
AGTTCTTCCTGATTTTGTCCAATTGTGGAATTTTTATCACCACTATTTTG Majority
9360 9370 9380 9390 9400
9302 AGTTCTTCCTGATTTTGTCCAATTGTGGAATTTTTATCACCACTATTTTG 2603_a12.seq
9302 AGTTCTTCCTGATTTTGTCCAATTGTGGAATTTTTATCACCACTATTTTG 18rs21_a12.seq
9210 AGTTCTTCCTGATTTTGTCCAATTGTGGAATTTTTATCACCACTATTTTG 515_a12.seq
9199 AGTTCTTCCTGATTTTGTCCAATTGTGGAATTTTTATCACCACTATTTTG cjb111_a12.seq
9237 AGTTCGTCTTGCCTTTGTTCAATTATAGAITTTTTATCAICACTATTTTG h36b_a12.seq
TATCGTAGTTTTTCCATTACTCTCAACCTTAACTTGCCAAGTCTGGTTAG Majority
9410 9420 9430 9440 9450
9352 TATCGTAGTTTTTCCATTACTCTCAACCTTAACTTGCCAAGTCTGGTTAG 2603_a12.seq
9352 TATCGTAGTTTTTCCATTACTCTCAACCTTAACTTGCCAAGTCTGGTTAG 18rs21_a12.seq
9260 TATCGTAGTTTTTCCATTACTCTCAACCTTAACTTGCCAAGTCTGGTTAG 515_a12.seq
9249 TATCGTAGTTTTTCCATTACTCTCAACCTTAACTTGCCAAGTCTGGTTAG cjb111_a12.seq
9287 TATCGTAGTTTTTCCATTACTCTCAACCTTAACTTGCCAAGTCTGGGTAG h36b_a12.seq

FIGURE 19T

```
                                                
```

```
TCCCAATTTTTGAATAATTTTCTTTTTATTTGATAATCTTATTTTTTTAT Majority
          9910      9920      9930      9940      9950
9852 TCCCAATTTTTGAATAATTTTCTTTTTATTTGATAATCTTATTTTTTTAT 2603_ai2.seq
9852 TCCCAATTTTTGAATAATTTTCTTTTTATTTGATAATCTTATTTTTTTAT 18rs21_ai2.seq
9760 TCCCAATTTTTGAATGATTTTTCTTTTTATTTGATA--CATTATTGTCTC 515_ai2.seq
9749 TCCCAATTTTTGAATAATTTTTCTTTTTATTTGATAATCTTATTTTTTTAT cjb111_ai2.seq
9787 TCCCAATTTTTGAATGATTTTTCTTTTTATTTGATAATCTTATTTTTTTAT h36b_ai2.seq

TATCTTAGAAATATTTCAATGAGCTTAAGTAGTTGATTTTTCTTTTTTTTA Majority
          9960      9970      9980      9990     10000
9902 TATCTTAGAAATATTTCAATGAGCTTAAGTAGTTGATTTTTCTTTTTTTTA 2603_ai2.seq
9902 TATCTTAGAAATATTTCAATGAGCTTAAGTAGTTGATTTTTCTTTTTTTTA 18rs21_ai2.seq
9808 TATCATGGAAACGTATTAATIA--TAATGTAGTTGTCCCTTACTTACCCC 515_ai2.seq
9799 TATCTTAGAAATATTTCAATGAGCTTAAGTAGTTGATTTTTCTTTTTTTTA cjb111_ai2.seq
9837 TATCTTAGAAATACTTCAATIAGCTTAAGTAGTTGATTTTTCTTTTTTTTA h36b_ai2.seq

TGTTTTAAAATATTGCTTAAAAATAATGTTTGAGAGAG--TTTACTGAAT Majority
         10010     10020     10030     10040     10050
9952 TGTTTTAAAATATTGCTTAAAAATAATGTTTGAGAGAG--TTTACTGAAT 2603_ai2.seq
9952 TGTTTTAAAATATTGCTTAAAAATAATGTTTGAGAGAG--TTTACTGAAT 18rs21_ai2.seq
9856 T-TTTGAITGAAGCTACAAAAIATAITTATCTAIGATACG--ATTACAATAT 515_ai2.seq
9849 TGTTTTAAAATATTGCTTAAAAATAATGTTTGAGAGAG--TTTACTGAAT cjb111_ai2.seq
9887 TGTTTTAAAATATTGCTTAAAAATAATGTTTGAGAGAGAGTTTACTGAAT h36b_ai2.seq

TGATTGAAAATTATTTAGAAAAAGACATCCTTAATCAAATAAAACTTCTA Majority
         10060     10070     10080     10090     10100
10000 TGATTGAAAATTATTTAGAAAAAGACATCCTTAATCAAATAAAACTTCTA 2603_ai2.seq
10000 TGATTGAAAATTATTTAGAAAAAGACATCCTTAATCAAATAAAACTTCTA 18rs21_ai2.seq
9903 TGITGTCAAAAATGTTTAIIAIAGAIICGTIIAAAATIAATTIAAAAAAGAC 515_ai2.seq
9897 TGATTGAAAATTATTTAGAAAAAGACATCCTTAATCAAATAAAACTTCTA cjb111_ai2.seq
9937 TGATTGAAAATTATTTAGAAAAAGACATCCTTAATCAAATAAAACTTCTA h36b_ai2.seq

ACTTTATGCTATGATTACTACCCTTCCATTACTCTAGACAAATCATGTCA Majority
         10110     10120     10130     10140     10150
10050 ACTTTATGCTATGATTACTACCCTTCCATTACTCTAGACAAATCATGTCA 2603_ai2.seq
10050 ACTTTATGCTATGATTACTACCCTTCCATTACTCTAGACAAATCATGTCA 18rs21_ai2.seq
9953 GATATAGACTATATTTTICITGCTTACTTACTTCCCATAAGTCTTTC 515_ai2.seq
9947 ACTTTATGCTATGATTACTACCCTTCCATTACTCTAGACAAATCATGTCA cjb111_ai2.seq
9987 ACTTTATGCTATGATTACTACCCTTCCATTACTCTAGACAAATCATGTCA h36b_ai2.seq

TCAACTTGGTTTATCTGAACTACTTATTAGGAA-ATATTGTCATGATTTA Majority
         10160     10170     10180     10190     10200
10100 TCAACTTGGTTTATCTGAACTACTTATTAGGAA-ATATTGTCATGATTTA 2603_ai2.seq
10100 TCAACTTGGTTTATCTGAACTACTTATTAGGAA-ATATTGTCATGATTTA 18rs21_ai2.seq
10003 TAAICCAAATTGGACTGAGAAGCGTATCGAIAATGTAATAGCIAITTTCG 515_ai2.seq
9997 TCAACTTGGTTTATCTGAACTACTTATTAGGAA-ATATTGTCATGATTTA cjb111_ai2.seq
10037 TCAACTTGGTTTATCTGAACTACTTATTAGGAA-ATATTGTCATGATTTA h36b_ai2.seq

ACAACTTTATTTAACAGTCAACTCTCTCTGAATATCGAAAAGAGTACAAT Majority
         10210     10220     10230     10240     10250
10149 ACAACTTTATTTAACAGTCAACTCTCTCTGAATATCGAAAAGAGTACAAT 2603_ai2.seq
10149 ACAACTTTATTTAACAGTCAACTCTCTCTGAATATCGAAAAGAGTACAAT 18rs21_ai2.seq
10053 AAAATTATCTCAAAAIICAAAAATTATIAGTCACIAAAGATGCCTT 515_ai2.seq
10046 ACAACTTTATTTAACAGTCAACTCTCTCTGAATATCGAAAAGAGTACAAT cjb111_ai2.seq
10086 ACAACTTTATTTAACAGTCAACTCTCTCTGAATATCGAAAAGAGTACAAT h36b_ai2.seq

AGTCTACCAATCTAATGGTGTAACTAGAGAACAAGCTTTCAAATATATTT Majority
         10260     10270     10280     10290     10300
10199 AGTCTACCAATCTAATGGTGTAACTAGAGAACAAGCTTTCAAATATATTT 2603_ai2.seq
10199 AGTCTACCAATCTAATGGTGTAACTAGAGAACAAGCTTTCAAATATATTT 18rs21_ai2.seq
10103 CCCTTATCIGACTCCTAICATGACGAGIIGGTAAAAGTTGCTAT-TATTT 515_ai2.seq
10096 AGTCTACCAATCTAATGGTGTAACTAGAGAACAAGCTTTCAAATATATTT cjb111_ai2.seq
10136 AGTCTACCAATCTAATGGTGTAACTAGAGAACAAGCTTTCAAATATATTT h36b_ai2.seq

ATCATCAATCACACGTTTTACAACTTTTAAAATTTTTGATCACGAATGAT Majority
         10310     10320     10330     10340     10350
10249 ATCATCAATCACACGTTTTACAACTTTTAAAATTTTTGATCACGAATGAT 2603_ai2.seq
10249 ATCATCAATCACACGTTTTACAACTTTTAAAATTTTTGATCACGAATGAT 18rs21_ai2.seq
10152 TTTCCGAACA----TTTTAGG----TTA-AAATCGATTAATCC----TGAA 515_ai2.seq
10146 ATCATCAATCACACGTTTTACAACTTTTAAAATTTTTGATCACGAATGAT cjb111_ai2.seq
10186 ATCATCAATCACACGTTTTACAACTTTTAAAATTTTTGATCACGAATGAT h36b_ai2.seq
```

FIGURE 19V

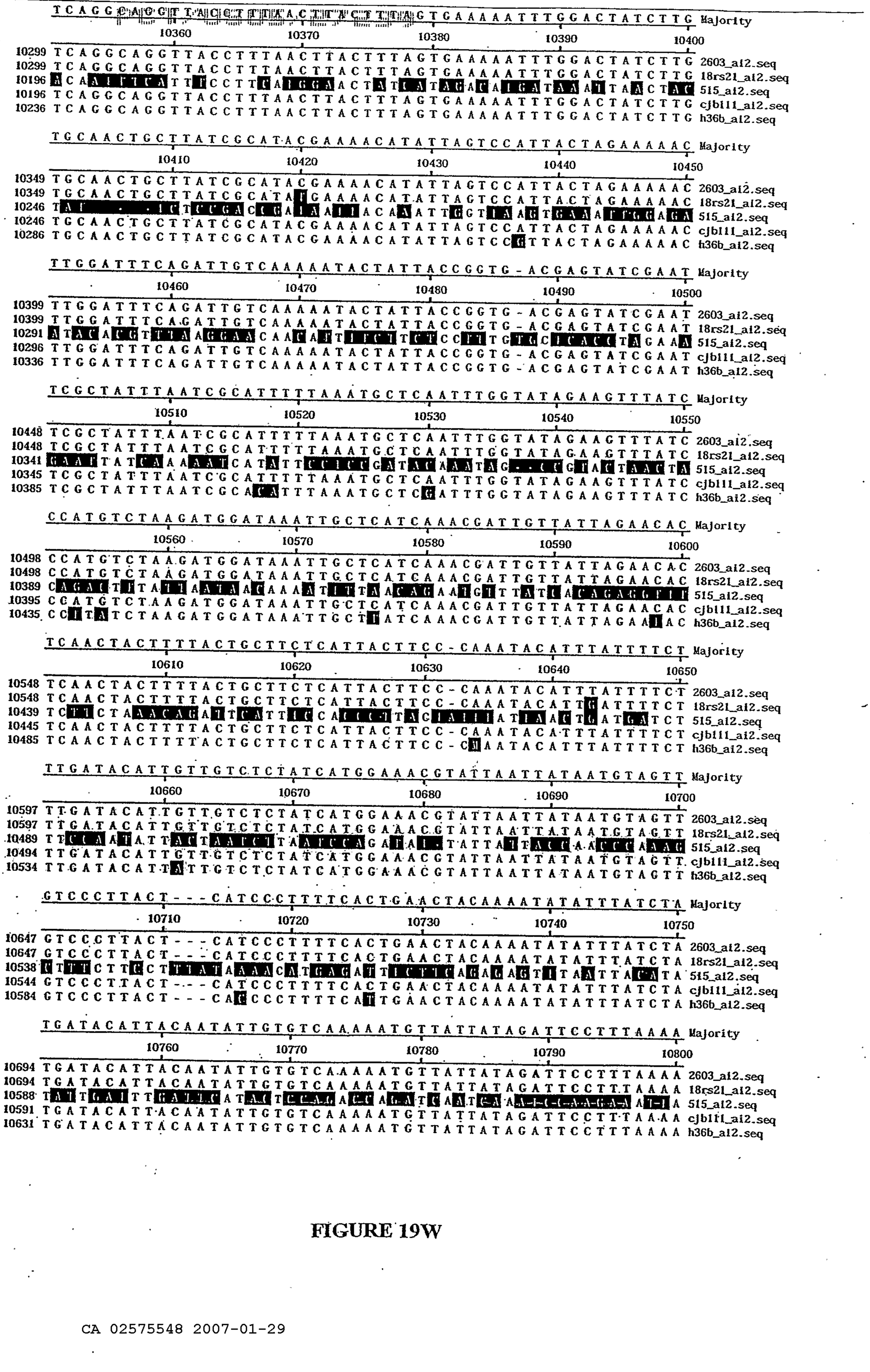

TCAGGCAGGTTACCTTTAACTTACTTTAGTGAAAAATTTGGACTATCTTG Majority
10360 10370 10380 10390 10400
10299 TCAGGCAGGTTACCTTTAACTTACTTTAGTGAAAAATTTGGACTATCTTG 2603_a12.seq
10299 TCAGGCAGGTTACCTTTAACTTACTTTAGTGAAAAATTTGGACTATCTTG 2603_a12.seq
10196 ACAAIITCATTTCCTTGAIGGAACTATIATAGACAIGATAAATTACTAC 18rs21_a12.seq
10196 TCAGGCAGGTTACCTTTAACTTACTTTAGTGAAAAATTTGGACTATCTTG 515_a12.seq
10236 TCAGGCAGGTTACCTTTAACTTACTTTAGTGAAAAATTTGGACTATCTTG cjb111_a12.seq
10236 TCAGGCAGGTTACCTTTAACTTACTTTAGTGAAAAATTTGGACTATCTTG h36b_a12.seq
TGCAACTGCTTATCGCATACGAAAACATATTAGTCCATTACTAGAAAAAC Majority
10410 10420 10430 10440 10450
10349 TGCAACTGCTTATCGCATACGAAAACATATTAGTCCATTACTAGAAAAAC 2603_a12.seq
10349 TGCAACTGCTTATCGCATACGAAAACATATTAGTCCATTACTAGAAAAAC 2603_a12.seq
10246 TAI..ICTCTGACIGAIAAIIACAATTGGTIAGTGAAATTGGAGA 18rs21_a12.seq
10246 TGCAACTGCTTATCGCATACGAAAACATATTAGTCCATTACTAGAAAAAC 515_a12.seq
10286 TGCAACTGCTTATCGCATACGAAAACATATTAGTCCATTACTAGAAAAAC cjb111_a12.seq
10286 TGCAACTGCTTATCGCATACGAAAACATATTAGTCCGTTACTAGAAAAAC h36b_a12.seq
TTGGATTTCAGATTGTCAAAAATACTATTACCGGTG-ACGAGTATCGAAT Majority
10460 10470 10480 10490 10500
10399 TTGGATTTCAGATTGTCAAAAATACTATTACCGGTG-ACGAGTATCGAAT 2603_a12.seq
10399 TTGGATTTCAGATTGTCAAAAATACTATTACCGGTG-ACGAGTATCGAAT 2603_a12.seq
10291 ATACACGTTTAAGGAACAACATTICTTCTCTTGTGCICACCTAGAA 18rs21_a12.seq
10296 TTGGATTTCAGATTGTCAAAAATACTATTACCGGTG-ACGAGTATCGAAT 515_a12.seq
10336 TTGGATTTCAGATTGTCAAAAATACTATTACCGGTG-ACGAGTATCGAAT cjb111_a12.seq
10336 TTGGATTTCAGATTGTCAAAAATACTATTACCGGTG-ACGAGTATCGAAT h36b_a12.seq
TCGCTATTTAATCGCATTTTTAAATGCTCAATTTGGTATAGAAGTTTATC Majority
10510 10520 10530 10540 10550
10448 TCGCTATTTAATCGCATTTTTAAATGCTCAATTTGGTATAGAAGTTTATC 2603_a12.seq
10448 TCGCTATTTAATCGCATTTTTAAATGCTCAATTTGGTATAGAAGTTTATC 2603_a12.seq
10341 GAATTATCAAAAATCATATTCCICTGATACAAATAG..CTGTACTAACTA 18rs21_a12.seq
10345 TCGCTATTTAATCGCATTTTTAAATGCTCAATTTGGTATAGAAGTTTATC 515_a12.seq
10385 TCGCTATTTAATCGCACATTTAAATGCTCGATTTGGTATAGAAGTTTATC h36b_a12.seq
CCATGTCTAAGATGGATAAATTGCTCATCAAACGATTGTTATTAGAACAC Majority
10560 10570 10580 10590 10600
10498 CCATGTCTAAGATGGATAAATTGCTCATCAAACGATTGTTATTAGAACAC 2603_a12.seq
10498 CCATGTCTAAGATGGATAAATTGCTCATCAAACGATTGTTATTAGAACAC 2603_a12.seq
10389 CAGACTTTATTAATAACAAATTTTAACAGAATGTTTATCAGAGGTTT 18rs21_a12.seq
10395 CCATGTCTAAGATGGATAAATTGCTCATCAAACGATTGTTATTAGAACAC 515_a12.seq
10435 CCTTATCTAAGATGGATAAATTGCTTATCAAACGATTGTTATTAGAAIAC h36b_a12.seq
TCAACTACTTTTACTGCTTCTCATTACTTCC-CAAATACATTTATTTTCT Majority
10610 10620 10630 10640 10650
10548 TCAACTACTTTTACTGCTTCTCATTACTTCC-CAAATACATTTATTTTCT 2603_a12.seq
10548 TCAACTACTTTTACTGCTTCTCATTACTTCC-CAAATACATTGATTTTCT 2603_a12.seq
10439 TCTTCTAAACAGATTCATTICCACICCTTACGIAIIIATIAATGATCT 515_a12.seq
10485 TCAACTACTTTTACTGCTTCTCATTACTTCC-CAAATACATTTATTTTCT cjb111_a12.seq
10485 TCAACTACTTTTACTGCTTCTCATTACTTCC-CMAATACATTTATTTTCT h36b_a12.seq
TTGATACATTGTTGTCTCTATCATGGAAACGTATTAATTATAATGTAGTT Majority
10660 10670 10680 10690 10700
10597 TTGATACATTGTTGTCTCTATCATGGAAACGTATTAATTATAATGTAGTT 2603_a12.seq
10597 TTGATACATTGTTGTCTCTATCATGGAAACGTATTAATTATAATGTAGTT 2603_a12.seq
10489 TTCCAATATTACTAATCTTAATCCAGATAI.TATTATTACCAACCCAAAG 18rs21_a12.seq
10494 TTGATACATTGTTGTCTCTATCATGGAAACGTATTAATTATAATGTAGTT 515_a12.seq
10534 TTGATACATTATTGTCTCTATCATGGAAACGTATTAATTATAATGTAGTT h36b_a12.seq
GTCCCTTACT---CATCCCTTTTCACTGAACTACAAAATATATTTATCTA Majority
10710 10720 10730 10740 10750
10647 GTCCCTTACT---CATCCCTTTTCACTGAACTACAAAATATATTTATCTA 2603_a12.seq
10647 GTCCCTTACT---CATCCCTTTTCACTGAACTACAAAATATATTTATCTA 2603_a12.seq
10538 CTTTCTTGCTTTATAAAACATGAGATTICTTCAGAGAGTITAATTACATA 18rs21_a12.seq
10544 GTCCCTTACT---CATCCCTTTTCACTGAACTACAAAATATATTTATCTA 515_a12.seq
10584 GTCCCTTACT---CACCCTTTTCATTGAACTACAAAATATATTTATCTA h36b_a12.seq
TGATACATTACAATATTGTGTCAAAAATGTTATTATAGATTCCTTTAAAA Majority
10760 10770 10780 10790 10800
10694 TGATACATTACAATATTGTGTCAAAAATGTTATTATAGATTCCTTTAAAA 2603_a12.seq
10694 TGATACATTACAATATTGTGTCAAAAATGTTATTATAGATTCCTTTAAAA 2603_a12.seq
10588 TATTGAITTGATTCATCACTCTAAGACTCTAGATCAATGAAGAAGAAGAAA 18rs21_a12.seq
10591 TGATACATTACAATATTGTGTCAAAAATGTTATTATAGATTCCTTTAAAA 515_a12.seq
10631 TGATACATTACAATATTGTGTCAAAAATGTTATTATAGATTCCTTTAAAA cjb111_a12.seq
10631 TGATACATTACAATATTGTGTCAAAAATGTTATTATAGATTCCTTTAAAA h36b_a12.seq

FIGURE 19W

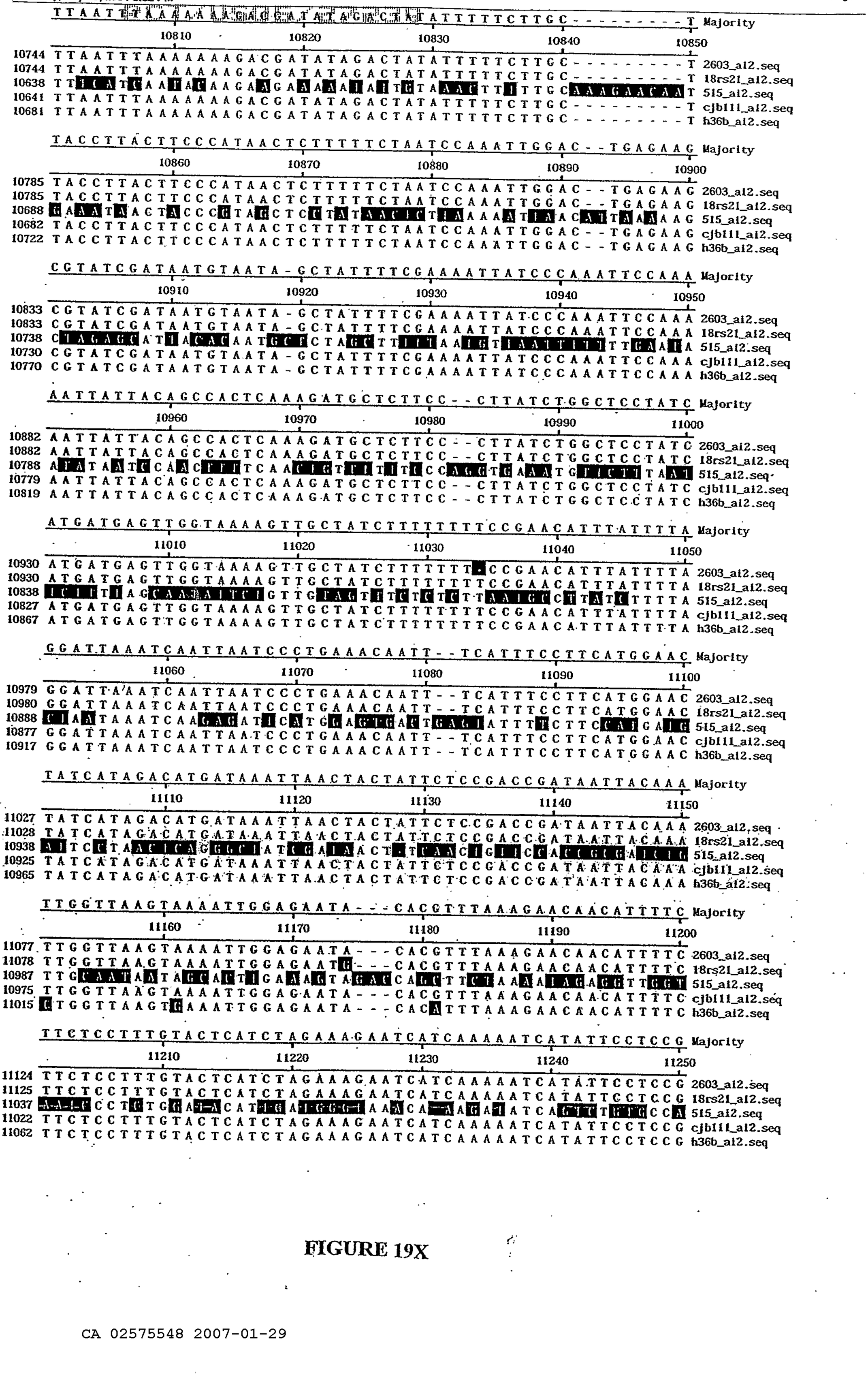

TTAATTTAAAAAAAAGACGATATAGACTATATTTTTCTTGC----------T Majority
10810 10820 10830 10840 10850
10744 TTAATTTAAAAAAAAGACGATATAGACTATATTTTTCTTGC----------T 2603_a12.seq
10744 TTAATTTAAAAAAAAGACGATATAGACTATATTTTTCTTGC----------T 18rs21_a12.seq
10638 TT[ICATCAATACAAGAGAAAAIATGTAAACTTITTGCAAAGAACAAT 515_a12.seq
10641 TTAATTTAAAAAAAAGACGATATAGACTATATTTTTCTTGC----------T cjb111_a12.seq
10681 TTAATTTAAAAAAAAGACGATATAGACTATATTTTTCTTGC----------T h36b_a12.seq
TACCTTACTTCCCATAACTCTTTTTCTAATCCAAATTGGAC--TGAGAAG Majority
10860 10870 10880 10890 10900
10785 TACCTTACTTCCCATAACTCTTTTTCTAATCCAAATTGGAC--TGAGAAG 2603_a12.seq
10785 TACCTTACTTCCCATAACTCTTTTTCTAATCCAAATTGGAC--TGAGAAG 18rs21_a12.seq
10688 GAAATAACTACCGTAGCTCCTATAACTCTAAAATATACTTAAAAG 515_a12.seq
10682 TACCTTACTTCCCATAACTCTTTTTCTAATCCAAATTGGAC--TGAGAAG cjb111_a12.seq
10722 TACCTTACTTCCCATAACTCTTTTTCTAATCCAAATTGGAC--TGAGAAG h36b_a12.seq
CGTATCGATAATGTAATA-GCTATTTTCGAAAATTATCCCAAATTCCAAA Majority
10910 10920 10930 10940 10950
10833 CGTATCGATAATGTAATA-GCTATTTTCGAAAATTATCCCAAATTCCAAA 2603_a12.seq
10833 CGTATCGATAATGTAATA-GCTATTTTCGAAAATTATCCCAAATTCCAAA 18rs21_a12.seq
10738 CTAGAGCATTACACAATGCTCTAGCTTTTAAIGTAATTTTTTGAAIA 515_a12.seq
10730 CGTATCGATAATGTAATA-GCTATTTTCGAAAATTATCCCAAATTCCAAA cjb111_a12.seq
10770 CGTATCGATAATGTAATA-GCTATTTTCGAAAATTATCCCAAATTCCAAA h36b_a12.seq
AATTATTACAGCCACTCAAAGATGCTCTTCC--CTTATCTGGCTCCTATC Majority
10960 10970 10980 10990 11000
10882 AATTATTACAGCCACTCAAAGATGCTCTTCC--CTTATCTGGCTCCTATC 2603_a12.seq
10882 AATTATTACAGCCACTCAAAGATGCTCTTCC--CTTATCTGGCTCCTATC 18rs21_a12.seq
10788 ATAATACTCAACTTTCAACIGTTITTCCCAGGTGAAATGTICTTAAT 515_a12.seq
10779 AATTATTACAGCCACTCAAAGATGCTCTTCC--CTTATCTGGCTCCTATC cjb111_a12.seq
10819 AATTATTACAGCCACTCAAAGATGCTCTTCC--CTTATCTGGCTCCTATC h36b_a12.seq
ATGATGAGTTGGTAAAAGTTGCTATCTTTTTTTTCCGAACATTTATTTTA Majority
11010 11020 11030 11040 11050
10930 ATGATGAGTTGGTAAAAGTTGCTATCTTTTTTT.CCGAACATTTATTTTA 2603_a12.seq
10930 ATGATGAGTTGGTAAAAGTTGCTATCTTTTTTTTCCGAACATTTATTTTA 18rs21_a12.seq
10838 ICITTIAGCAAAAITCIGTTGTAGTTCTCTCTTAAIGCCTTATCTTTTA 515_a12.seq
10827 ATGATGAGTTGGTAAAAGTTGCTATCTTTTTTTTCCGAACATTTATTTTA cjb111_a12.seq
10867 ATGATGAGTTGGTAAAAGTTGCTATCTTTTTTTTCCGAACATTTATTTTA h36b_a12.seq
GGATTAAATCAATTAATCCCTGAAACAATT--TCATTTCCTTCATGGAAC Majority
11060 11070 11080 11090 11100
10979 GGATTAAATCAATTAATCCCTGAAACAATT--TCATTTCCTTCATGGAAC 2603_a12.seq
10980 GGATTAAATCAATTAATCCCTGAAACAATT--TCATTTCCTTCATGGAAC 18rs21_a12.seq
10888 CIAATAAATCAAGAGATICATGGAGTGAITGAGIATTTTCTTCCAIGAIG 515_a12.seq
10877 GGATTAAATCAATTAATCCCTGAAACAATT--TCATTTCCTTCATGGAAC cjb111_a12.seq
10917 GGATTAAATCAATTAATCCCTGAAACAATT--TCATTTCCTTCATGGAAC h36b_a12.seq
TATCATAGACATGATAAATTAACTACTATTCTCCGACCGATAATTACAAA Majority
11110 11120 11130 11140 11150
11027 TATCATAGACATGATAAATTAACTACTATTCTCCGACCGATAATTACAAA 2603_a12.seq
11028 TATCATAGACATGATAAATTAACTACTATTCTCCGACCGATAATTACAAA 18rs21_a12.seq
10938 AITCTTAACICAGGGCIATCGAIAACT.TCAACIGIICCACCGGIGAICIG 515_a12.seq
10925 TATCATAGACATGATAAATTAACTACTATTCTCCGACCGATAATTACAAA cjb111_a12.seq
10965 TATCATAGACATGATAAATTAACTACTATTCTCCGACCGATAATTACAAA h36b_a12.seq
TTGGTTAAGTAAAATTGGAGAATA---CACGTTTAAAGAACAACATTTTC Majority
11160 11170 11180 11190 11200
11077 TTGGTTAAGTAAAATTGGAGAATA---CACGTTTAAAGAACAACATTTTC 2603_a12.seq
11078 TTGGTTAAGTAAAATTGGAGAATG---CACGTTTAAAGAACAACATTTTC 18rs21_a12.seq
10987 TTGCAATAATAGCATTIGAAAGTAGACCAGCTTCTAAAAIAGAGGTTGGT 515_a12.seq
10975 TTGGTTAAGTAAAATTGGAGAATA---CACGTTTAAAGAACAACATTTTC cjb111_a12.seq
11015 GTGGTTAAGTGAAATTGGAGAATA---CACATTTAAAGAACAACATTTTC h36b_a12.seq
TTCTCCTTTGTACTCATCTAGAAAGAATCATCAAAAATCATATTCCTCCG Majority
11210 11220 11230 11240 11250
11124 TTCTCCTTTGTACTCATCTAGAAAGAATCATCAAAAATCATATTCCTCCG 2603_a12.seq
11125 TTCTCCTTTGTACTCATCTAGAAAGAATCATCAAAAATCATATTCCTCCG 18rs21_a12.seq
11037 -A-A-I-GCCTGTGGA-I-ACATI-GAI-G-G-G-IAAACA--AGAIATCAGTCTGTGCCA 515_a12.seq
11022 TTCTCCTTTGTACTCATCTAGAAAGAATCATCAAAAATCATATTCCTCCG cjb111_a12.seq
11062 TTCTCCTTTGTACTCATCTAGAAAGAATCATCAAAAATCATATTCCTCCG h36b_a12.seq

FIGURE 19X

```
                ATACAGA--TAGCCGTACTAACTACAGACTTTATTAATAACCAAATTTTA Majority
                        11260       11270       11280       11290       11300
11174 ATACAGA--TAGCCGTACTAACTACAGACTTTATTAATAACCAAATTTTA 2603_a12.seq
11175 ATACAGA--TAGCCGTACTAACTACAGACTTTATTAATAACCAAATTTTA 18rs21_a12.seq
11086 ITAAAGAYATAGTCTGTTCAAAGIIIAIITTYCCCAAAAAGTTAATCTGT 515_a12.seq
11072 ATACAGA--TAGCCGTACTAACTACAGACTTTATTAATAACCAAATTTTA cjb111_a12.seq
11112 ATACAGA--TAGCCGTACTAACTACAGACTTTATTAATAACCAAATTTTA h36b_a12.seq

                ACAGAATGCTTATTACAGAGATTTTCTTCTAAAAAGATTCATTTCCACCC Majority
                        11310       11320       11330       11340       11350
11222 ACAGAATGCTTATTACAGAGATTTTCTTCTAAAAAGATTCATTTCCACCC 2603_a12.seq
11223 ACAGAATGCTTATTACAGAGATTTTCTTCTAAAAAGATTCATTTCCACCC 18rs21_a12.seq
11136 ITGGAGCTGATATTTCI.....CTTTCAAATGIGCIAATTCAGGTCCGICT 515_a12.seq
11120 ACAGAATGCTTATTACAGAGATTTTCTTCTAAAAAGATTCATTTCCACCC cjb111_a12.seq
11160 ACAGAATGCTTATTACAGAGATTTTCTTCTAAAAAGATTCATTTCCACCC h36b_a12.seq

                TTACTATTTATTAACTGATGATCTTTCCAATATTACTAATCTTAATCCAG Majority
                        11360       11370       11380       11390       11400
11272 TTACTATTTATTAACTGATGATCTTTCCAATATTACTAATCTTAATCCAG 2603_a12.seq
11273 TTACTATTTATTAACTGATGATCTTTCCAATATTACTAATCTTAATCCAG 18rs21_a12.seq
11181 CCTGCAATCTGTAAATAACATTTCAGAGTACTG-TGATCATCAA...AA 515_a12.seq
11170 TTACTATTTATTAACTGATGATCTTTCCAATATTACTAATCTTAATCCAG cjb111_a12.seq
11210 TTACTATTTATTAACTGATGATCTTTCCAATATTACTAATCTTAATCCAG h36b_a12.seq

                ACATTATTATTACCAATAAAAAGCTTTCTCCCTTTATCAAACATGAGATT Majority
                        11410       11420       11430       11440       11450
11322 ACATTATTATTACCAATAAAAAGCTTTCTCCCTTTATCAAACATGAGATT 2603_a12.seq
11323 ACATTATTATTACCAATAAAAAGCTTTCTCCCTTTATCAAACATGAGATT 18rs21_a12.seq
11227 AYGCTICTAAGAGCAATTCAAIGCCTTITICTTTAAT........AATTC 515_a12.seq
11220 ACATTATTATTACCAATAAAAAGCTTTCTCCCTTTATCAAACATGAGATT cjb111_a12.seq
11260 ACATTATTATTACCAATAAAAAGCTTTCTCCCTTTATCAAACATGAGATT h36b_a12.seq

                TCTTCAGAGAGTTTAATTACATATATTGATCTGATTAATACTTCAGACCA Majority
                        11460       11470       11480       11490       11500
11372 TCTTCAGAGAGTTTAATTACATATATTGATCTGATTAATACTTCAGACCA 2603_a12.seq
11373 TCTTCAGAGAGTTTAATTACATATATTGATCTGATTAATACTTCAGACCA 18rs21_a12.seq
11269 TACCAGCATAAGTGAIGAAATATCATCAGCAGATTIITCAAGGTAAGCC 515_a12.seq
11270 TCTTCAGAGAGTTTAATTACATATATTGATCTGATTAATACTTCAGACCA cjb111_a12.seq
11310 TCTTCAGAGAGTTTAATTACATATATTGATCTGATTAATACTTCAGACCA h36b_a12.seq

                AATCAATCAAATCCAAAAAATTATTTCATCAATACAGGAAGAAAAATATT Majority
                        11510       11520       11530       11540       11550
11422 AATCAATCAAATCCAAAAAATTATTTCATCAATACAGGAAGAAAAATATT 2603_a12.seq
11423 AATCAATCAAATCCAAAAAATTATTTCATCAATACAGGAAGAAAAATATT 18rs21_a12.seq
11319 GIGICAGCAAAATCAGAGCCTAGACTIICAGATACTGAATTAIAAATAAC 515_a12.seq
11320 AATCAATCAAATCCAAAAAATTATTTCATCAATACAGGAAGAAAAATATT cjb111_a12.seq
11360 AATCAATCAAATCCAAAAAATTATTTCATCAATACAGGAAGAAAAATATT h36b_a12.seq

                GCAAACTTTTTGCAAAAACTAATGAAATAACTAATCGTAGCTCCT------ Majority
                        11560       11570       11580       11590       11600
11472 GCAAACTTTTTGCAAAAACTAATGAAATAACTAATCGTAGCTCCT------ 2603_a12.seq
11473 GCAAACTTTTTGCAAAAACTAATGAAATAACTAATCGTAGCTCCT------ 18rs21_a12.seq
11369 ICCIIAGCTTCTAIATIAAAATGIITTAACGAITCAAGGCTICTCIIGG 515_a12.seq
11370 GCAAACTTTTTGCAAAAACTAATGAAATAACTAATCGTAGCTCCT------ cjb111_a12.seq
11410 GCAAACTTTTTGCAAAAACTAATGAAATAACTAATCGTAGCTCCT------ h36b_a12.seq

                ATAACTCTTAAAAATTAACATTAAAAAGCT-AGAGCATTGTGTAATGCTC Majority
                        11610       11620       11630       11640       11650
11517 ATAACTCTTAAAAATTAACATTAAAAAGCT-AGAGCATTGTGTAATGCTC 2603_a12.seq
11518 ATAACTCTTAAAAATTAACATTAAAAAGCT-AGAGCATTGTGTAATGCTC 18rs21_a12.seq
11419 ATACCGCATAAAAATCTGGACGATAATGCTTACACGCGCTGTGAGA... 515_a12.seq
11415 ATAACTCTTAAAAATTAACATTAAAAAGCT-AGAGCATTGTGTAATGCTC cjb111_a12.seq
11455 ATAACTCTTAAAAATTAACATTAAAAAGCT-AGAGCATTGTGTAATGCTC h36b_a12.seq

                TAGCTTTTTAATGTTAATTTTTTTGAATAATATAATCCAACTTTTCAACT Majority
                        11660       11670       11680       11690       11700
11566 TAGCTTTTTAATGTTAATTTTTTTGAATAATATAATCCAACTTTTCAACT 2603_a12.seq
11567 TAGCTTTTTAATGTTAATTTTTTTGAATAATATAATCCAACTTTTCAACT 18rs21_a12.seq
11466 NAGATGTTTCATAAGATAGCTCCAAAGAAAICTAAAAAACGA.TTATIGACA 515_a12.seq
11464 TAGCTTTTTAATGTTAATTTTTTTGAATAATATAATCCAACTTTTCAACT cjb111_a12.seq
11504 TAGCTTTTTAATGTTAATTTTTTTGAATAATATAATCCAACTTTTCAACT h36b_a12.seq
```

FIGURE 19Y

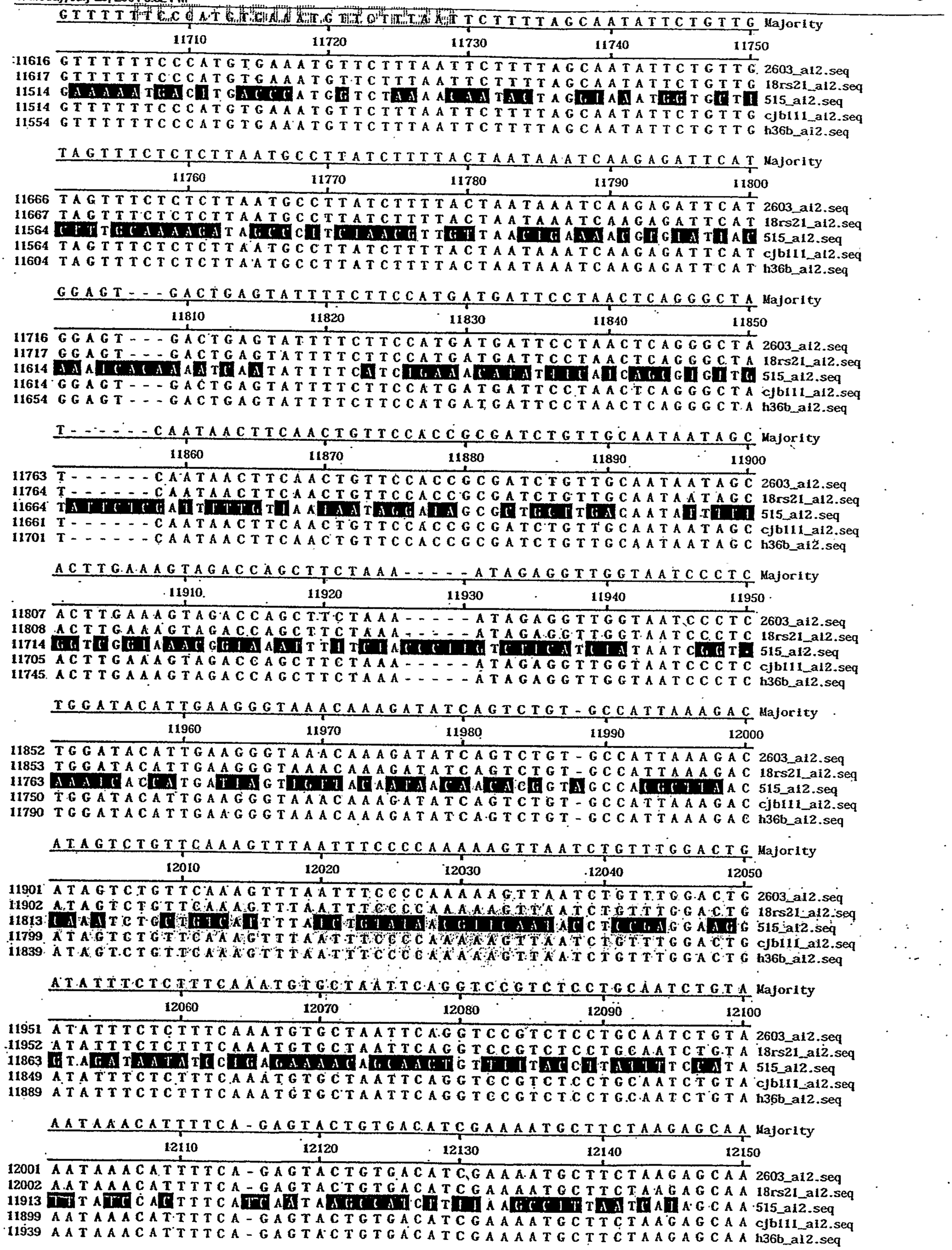

GTTTTTTCCCATGTGAAATGTTCTTTAATTCTTTTAGCAATATTCTGTTG Majority
11710 11720 11730 11740 11750
11616 GTTTTTTCCCATGTGAAATGTTCTTTAATTCTTTTAGCAATATTCTGTTG 2603_a12.seq
11617 GTTTTTTCCCATGTGAAATGTTCTTTAATTCTTTTAGCAATATTCTGTTG 18rs21_a12.seq
11514 515_a12.seq
11514 GTTTTTTCCCATGTGAAATGTTCTTTAATTCTTTTAGCAATATTCTGTTG cjb111_a12.seq
11554 GTTTTTTCCCATGTGAAATGTTCTTTAATTCTTTTAGCAATATTCTGTTG h36b_a12.seq
TAGTTTCTCTCTTAATGCCTTATCTTTTACTAATAAATCAAGAGATTCAT Majority
11760 11770 11780 11790 11800
11666 TAGTTTCTCTCTTAATGCCTTATCTTTTACTAATAAATCAAGAGATTCAT 2603_a12.seq
11667 TAGTTTCTCTCTTAATGCCTTATCTTTTACTAATAAATCAAGAGATTCAT 18rs21_a12.seq
11564 515_a12.seq
11564 TAGTTTCTCTCTTAATGCCTTATCTTTTACTAATAAATCAAGAGATTCAT cjb111_a12.seq
11604 TAGTTTCTCTCTTAATGCCTTATCTTTTACTAATAAATCAAGAGATTCAT h36b_a12.seq
GGAGT---GACTGAGTATTTTCTTCCATGATGATTCCTAACTCAGGGCTA Majority
11810 11820 11830 11840 11850
11716 GGAGT---GACTGAGTATTTTCTTCCATGATGATTCCTAACTCAGGGCTA 2603_a12.seq
11717 GGAGT---GACTGAGTATTTTCTTCCATGATGATTCCTAACTCAGGGCTA 18rs21_a12.seq
11614 515_a12.seq
11614 GGAGT---GACTGAGTATTTTCTTCCATGATGATTCCTAACTCAGGGCTA cjb111_a12.seq
11654 GGAGT---GACTGAGTATTTTCTTCCATGATGATTCCTAACTCAGGGCTA h36b_a12.seq
T------CAATAACTTCAACTGTTCCACCGCGATCTGTTGCAATAATAGC Majority
11860 11870 11880 11890 11900
11763 T------CAATAACTTCAACTGTTCCACCGCGATCTGTTGCAATAATAGC 2603_a12.seq
11764 T------CAATAACTTCAACTGTTCCACCGCGATCTGTTGCAATAATAGC 18rs21_a12.seq
11664 515_a12.seq
11661 T------CAATAACTTCAACTGTTCCACCGCGATCTGTTGCAATAATAGC cjb111_a12.seq
11701 T------CAATAACTTCAACTGTTCCACCGCGATCTGTTGCAATAATAGC h36b_a12.seq
ACTTGAAAGTAGACCAGCTTCTAAA-----ATAGAGGTTGGTAATCCCTC Majority
11910 11920 11930 11940 11950
11807 ACTTGAAAGTAGACCAGCTTCTAAA-----ATAGAGGTTGGTAATCCCTC 2603_a12.seq
11808 ACTTGAAAGTAGACCAGCTTCTAAA-----ATAGAGGTTGGTAATCCCTC 18rs21_a12.seq
11714 515_a12.seq
11705 ACTTGAAAGTAGACCAGCTTCTAAA-----ATAGAGGTTGGTAATCCCTC cjb111_a12.seq
11745 ACTTGAAAGTAGACCAGCTTCTAAA-----ATAGAGGTTGGTAATCCCTC h36b_a12.seq
TGGATACATTGAAGGGTAAACAAAGATATCAGTCTGT-GCCATTAAAGAC Majority
11960 11970 11980 11990 12000
11852 TGGATACATTGAAGGGTAAACAAAGATATCAGTCTGT-GCCATTAAAGAC 2603_a12.seq
11853 TGGATACATTGAAGGGTAAACAAAGATATCAGTCTGT-GCCATTAAAGAC 18rs21_a12.seq
11763 515_a12.seq
11750 TGGATACATTGAAGGGTAAACAAAGATATCAGTCTGT-GCCATTAAAGAC cjb111_a12.seq
11790 TGGATACATTGAAGGGTAAACAAAGATATCAGTCTGT-GCCATTAAAGAC h36b_a12.seq
ATAGTCTGTTCAAAGTTTAATTTCCCCAAAAAGTTAATCTGTTTGGACTG Majority
12010 12020 12030 12040 12050
11901 ATAGTCTGTTCAAAGTTTAATTTCCCCAAAAAGTTAATCTGTTTGGACTG 2603_a12.seq
11902 ATAGTCTGTTCAAAGTTTAATTTCCCCAAAAAGTTAATCTGTTTGGACTG 18rs21_a12.seq
11813 515_a12.seq
11799 ATAGTCTGTTCAAAGTTTAATTTCCCCAAAAAGTTAATCTGTTTGGACTG cjb111_a12.seq
11839 ATAGTCTGTTCAAAGTTTAATTTCCCCAAAAAGTTAATCTGTTTGGACTG h36b_a12.seq
ATATTTCTCTTTCAAATGTGCTAATTCAGGTCCGTCTCCTGCAATCTGTA Majority
12060 12070 12080 12090 12100
11951 ATATTTCTCTTTCAAATGTGCTAATTCAGGTCCGTCTCCTGCAATCTGTA 2603_a12.seq
11952 ATATTTCTCTTTCAAATGTGCTAATTCAGGTCCGTCTCCTGCAATCTGTA 18rs21_a12.seq
11863 515_a12.seq
11849 ATATTTCTCTTTCAAATGTGCTAATTCAGGTCCGTCTCCTGCAATCTGTA cjb111_a12.seq
11889 ATATTTCTCTTTCAAATGTGCTAATTCAGGTCCGTCTCCTGCAATCTGTA h36b_a12.seq
AATAAACATTTTCA-GAGTACTGTGACATCGAAAATGCTTCTAAGAGCAA Majority
12110 12120 12130 12140 12150
12001 AATAAACATTTTCA-GAGTACTGTGACATCGAAAATGCTTCTAAGAGCAA 2603_a12.seq
12002 AATAAACATTTTCA-GAGTACTGTGACATCGAAAATGCTTCTAAGAGCAA 18rs21_a12.seq
11913 515_a12.seq
11899 AATAAACATTTTCA-GAGTACTGTGACATCGAAAATGCTTCTAAGAGCAA cjb111_a12.seq
11939 AATAAACATTTTCA-GAGTACTGTGACATCGAAAATGCTTCTAAGAGCAA h36b_a12.seq

FIGURE 19Z

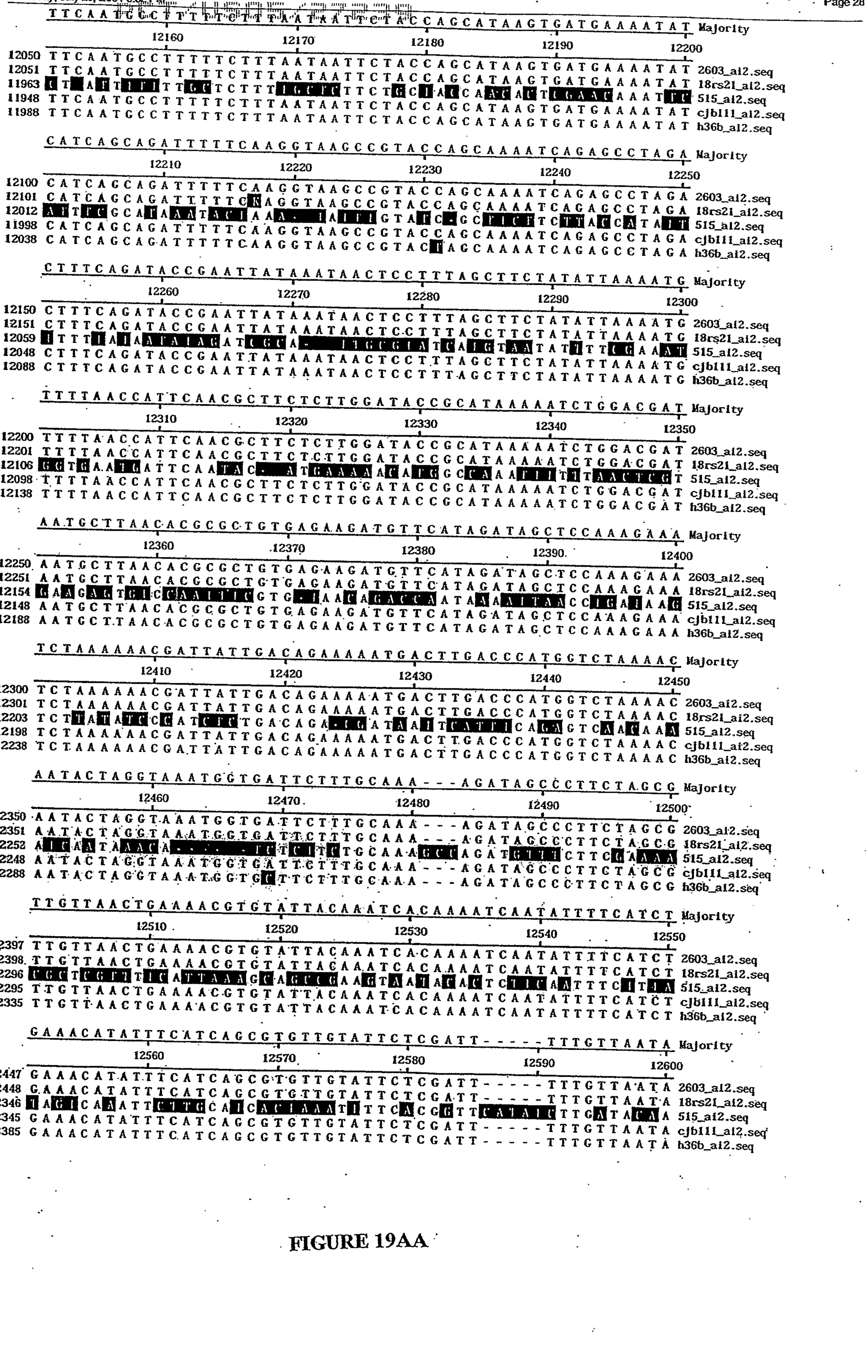

TTCAATGCCTTTTTCTTTAATAATTCTACCAGCATAAGTGATGAAAATAT Majority
12160 12170 12180 12190 12200
12050 TTCAATGCCTTTTTCTTTAATAATTCTACCAGCATAAGTGATGAAAATAT 2603_a12.seq
12051 TTCAATGCCTTTTTCTTTAATAATTCTACCAGCATAAGTGATGAAAATAT 18rs21_a12.seq
11963 [illegible] 515_a12.seq
11948 TTCAATGCCTTTTTCTTTAATAATTCTACCAGCATAAGTGATGAAAATAT cjb111_a12.seq
11988 TTCAATGCCTTTTTCTTTAATAATTCTACCAGCATAAGTGATGAAAATAT h36b_a12.seq
CATCAGCAGATTTTTCAAGGTAAGCCGTACCAGCAAAATCAGAGCCTAGA Majority
12210 12220 12230 12240 12250
12100 CATCAGCAGATTTTTCAAGGTAAGCCGTACCAGCAAAATCAGAGCCTAGA 2603_a12.seq
12101 CATCAGCAGATTTTTCNAGGTAAGCCGTACCAGCAAAATCAGAGCCTAGA 18rs21_a12.seq
12012 [illegible] 515_a12.seq
11998 CATCAGCAGATTTTTCAAGGTAAGCCGTACCAGCAAAATCAGAGCCTAGA cjb111_a12.seq
12038 CATCAGCAGATTTTTCAAGGTAAGCCGTACTAGCAAAATCAGAGCCTAGA h36b_a12.seq
CTTTCAGATACCGAATTATAAATAACTCCTTTAGCTTCTATATTAAAATG Majority
12260 12270 12280 12290 12300
12150 CTTTCAGATACCGAATTATAAATAACTCCTTTAGCTTCTATATTAAAATG 2603_a12.seq
12151 CTTTCAGATACCGAATTATAAATAACTCCTTTAGCTTCTATATTAAAATG 18rs21_a12.seq
12059 [illegible] 515_a12.seq
12048 CTTTCAGATACCGAATTATAAATAACTCCTTTAGCTTCTATATTAAAATG cjb111_a12.seq
12088 CTTTCAGATACCGAATTATAAATAACTCCTTTAGCTTCTATATTAAAATG h36b_a12.seq
TTTTAACCATTCAACGCTTCTCTTGGATACCGCATAAAAATCTGGACGAT Majority
12310 12320 12330 12340 12350
12200 TTTTAACCATTCAACGCTTCTCTTGGATACCGCATAAAAATCTGGACGAT 2603_a12.seq
12201 TTTTAACCATTCAACGCTTCTCTTGGATACCGCATAAAAATCTGGACGAT 18rs21_a12.seq
12106 [illegible] 515_a12.seq
12098 TTTTAACCATTCAACGCTTCTCTTGGATACCGCATAAAAATCTGGACGAT cjb111_a12.seq
12138 TTTTAACCATTCAACGCTTCTCTTGGATACCGCATAAAAATCTGGACGAT h36b_a12.seq
AATGCTTAACACGCGCTGTGAGAAGATGTTCATAGATAGCTCCAAAGAAA Majority
12360 12370 12380 12390 12400
12250 AATGCTTAACACGCGCTGTGAGAAGATGTTCATAGATAGCTCCAAAGAAA 2603_a12.seq
12251 AATGCTTAACACGCGCTGTGAGAAGATGTTCATAGATAGCTCCAAAGAAA 18rs21_a12.seq
12154 [illegible] 515_a12.seq
12148 AATGCTTAACACGCGCTGTGAGAAGATGTTCATAGATAGCTCCAAAGAAA cjb111_a12.seq
12188 AATGCTTAACACGCGCTGTGAGAAGATGTTCATAGATAGCTCCAAAGAAA h36b_a12.seq
TCTAAAAAACGATTATTGACAGAAAAATGACTTGACCCATGGTCTAAAAC Majority
12410 12420 12430 12440 12450
12300 TCTAAAAAACGATTATTGACAGAAAAATGACTTGACCCATGGTCTAAAAC 2603_a12.seq
12301 TCTAAAAAACGATTATTGACAGAAAAATGACTTGACCCATGGTCTAAAAC 18rs21_a12.seq
12203 [illegible] 515_a12.seq
12198 TCTAAAAAACGATTATTGACAGAAAAATGACTTGACCCATGGTCTAAAAC cjb111_a12.seq
12238 TCTAAAAAACGATTATTGACAGAAAAATGACTTGACCCATGGTCTAAAAC h36b_a12.seq
AATACTAGGTAAATGGTGATTCTTTGCAAA---AGATAGCCCTTCTAGCG Majority
12460 12470 12480 12490 12500
12350 AATACTAGGTAAATGGTGATTCTTTGCAAA---AGATAGCCCTTCTAGCG 2603_a12.seq
12351 AATACTAGGTAAATGGTGATTCTTTGCAAA---AGATAGCCCTTCTAGCG 18rs21_a12.seq
12252 [illegible] 515_a12.seq
12248 AATACTAGGTAAATGGTGATTCTTTGCAAA---AGATAGCCCTTCTAGCG cjb111_a12.seq
12288 AATACTAGGTAAATGGTGCTTCTTTGCAAA---AGATAGCCCTTCTAGCG h36b_a12.seq
TTGTTAACTGAAAACGTGTATTACAAATCACAAAATCAATATTTTCATCT Majority
12510 12520 12530 12540 12550
12397 TTGTTAACTGAAAACGTGTATTACAAATCACAAAATCAATATTTTCATCT 2603_a12.seq
12398 TTGTTAACTGAAAACGTGTATTACAAATCACAAAATCAATATTTTCATCT 18rs21_a12.seq
12296 [illegible] 515_a12.seq
12295 TTGTTAACTGAAAACGTGTATTACAAATCACAAAATCAATATTTTCATCT cjb111_a12.seq
12335 TTGTTAACTGAAAACGTGTATTACAAATCACAAAATCAATATTTTCATCT h36b_a12.seq
GAAACATATTTCATCAGCGTGTTGTATTCTCGATT-----TTTGTTAATA Majority
12560 12570 12580 12590 12600
12447 GAAACATATTTCATCAGCGTGTTGTATTCTCGATT-----TTTGTTAATA 2603_a12.seq
12448 GAAACATATTTCATCAGCGTGTTGTATTCTCGATT-----TTTGTTAATA 18rs21_a12.seq
12346 [illegible] 515_a12.seq
12345 GAAACATATTTCATCAGCGTGTTGTATTCTCGATT-----TTTGTTAATA cjb111_a12.seq
12385 GAAACATATTTCATCAGCGTGTTGTATTCTCGATT-----TTTGTTAATA h36b_a12.seq

FIGURE 19AA

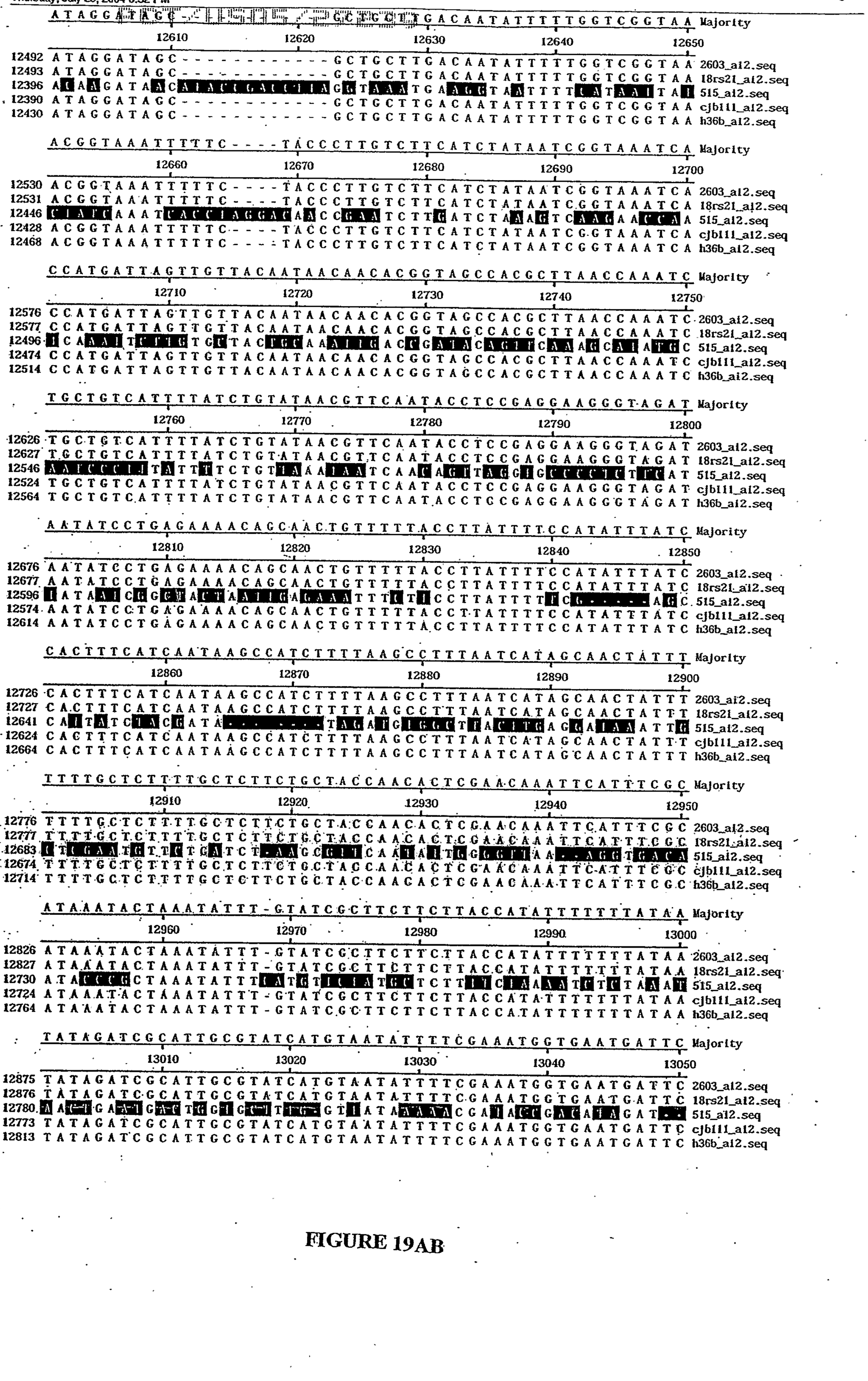

ATAGGATAGC-----GCTGCTTGACAATATTTTTGGTCGGTAA Majority
12610 12620 12630 12640 12650
12492 ATAGGATAGC----------GCTGCTTGACAATATTTTTGGTCGGTAA 2603_a12.seq
12493 ATAGGATAGC----------GCTGCTTGACAATATTTTTGGTCGGTAA 18rs21_a12.seq
12396 ATAGATACATACGATCTAGGTAAATGAAGGTATTTTTATAATATA 515_a12.seq
12390 ATAGGATAGC----------GCTGCTTGACAATATTTTTGGTCGGTAA cjb111_a12.seq
12430 ATAGGATAGC----------GCTGCTTGACAATATTTTTGGTCGGTAA h36b_a12.seq
ACGGTAAATTTTTC----TACCCTTGTCTTCATCTATAATCGGTAAATCA Majority
12660 12670 12680 12690 12700
12530 ACGGTAAATTTTTC----TACCCTTGTCTTCATCTATAATCGGTAAATCA 2603_a12.seq
12531 ACGGTAAATTTTTC----TACCCTTGTCTTCATCTATAATCGGTAAATCA 18rs21_a12.seq
12446 CTATCAAATCACCTAGGACAACCGAATCTTGATCTAAAGTCAAGAACCA 515_a12.seq
12428 ACGGTAAATTTTTC----TACCCTTGTCTTCATCTATAATCGGTAAATCA cjb111_a12.seq
12468 ACGGTAAATTTTTC----TACCCTTGTCTTCATCTATAATCGGTAAATCA h36b_a12.seq
CCATGATTAGTTGTTACAATAACAACACGGTAGCCACGCTTAACCAAATC Majority
12710 12720 12730 12740 12750
12576 CCATGATTAGTTGTTACAATAACAACACGGTAGCCACGCTTAACCAAATC 2603_a12.seq
12577 CCATGATTAGTTGTTACAATAACAACACGGTAGCCACGCTTAACCAAATC 18rs21_a12.seq
12496 ICAAATGTTGTGTACTGCAAAITGACGGATACAGTTCAAAGCATATGC 515_a12.seq
12474 CCATGATTAGTTGTTACAATAACAACACGGTAGCCACGCTTAACCAAATC cjb111_a12.seq
12514 CCATGATTAGTTGTTACAATAACAACACGGTAGCCACGCTTAACCAAATC h36b_a12.seq
TGCTGTCATTTTATCTGTATAACGTTCAATACCTCCGAGGAAGGGTAGAT Majority
12760 12770 12780 12790 12800
12626 TGCTGTCATTTTATCTGTATAACGTTCAATACCTCCGAGGAAGGGTAGAT 2603_a12.seq
12627 TGCTGTCATTTTATCTGTATAACGTTCAATACCTCCGAGGAAGGGTAGAT 18rs21_a12.seq
12546 AAICCCITTATTTCTGTAAAIAATCAACAGTTAGGIGCCCCTCTTCAT 515_a12.seq
12524 TGCTGTCATTTTATCTGTATAACGTTCAATACCTCCGAGGAAGGGTAGAT cjb111_a12.seq
12564 TGCTGTCATTTTATCTGTATAACGTTCAATACCTCCGAGGAAGGGTAGAT h36b_a12.seq
AATATCCTGAGAAAACAGCAACTGTTTTTACCTTATTTTCCATATTTATC Majority
12810 12820 12830 12840 12850
12676 AATATCCTGAGAAAACAGCAACTGTTTTTACCTTATTTTCCATATTTATC 2603_a12.seq
12677 AATATCCTGAGAAAACAGCAACTGTTTTTACCTTATTTTCCATATTTATC 18rs21_a12.seq
12596 IATAAICGGGTACTAAIGAGAAATTTCTICCTTATTTTICG------AGC 515_a12.seq
12574 AATATCCTGAGAAAACAGCAACTGTTTTTACCTTATTTTCCATATTTATC cjb111_a12.seq
12614 AATATCCTGAGAAAACAGCAACTGTTTTTACCTTATTTTCCATATTTATC h36b_a12.seq
CACTTTCATCAATAAGCCATCTTTTAAGCCTTTAATCATAGCAACTATTT Majority
12860 12870 12880 12890 12900
12726 CACTTTCATCAATAAGCCATCTTTTAAGCCTTTAATCATAGCAACTATTT 2603_a12.seq
12727 CACTTTCATCAATAAGCCATCTTTTAAGCCTTTAATCATAGCAACTATTT 18rs21_a12.seq
12641 CATITATCTACGATA------TAGATGIGGCTTACITGAGGAIAAATTG 515_a12.seq
12624 CACTTTCATCAATAAGCCATCTTTTAAGCCTTTAATCATAGCAACTATTT cjb111_a12.seq
12664 CACTTTCATCAATAAGCCATCTTTTAAGCCTTTAATCATAGCAACTATTT h36b_a12.seq
TTTTGCTCTTTTGCTCTTCTGCTACCAACACTCGAACAAATTCATTTCGC Majority
12910 12920 12930 12940 12950
12776 TTTTGCTCTTTTGCTCTTCTGCTACCAACACTCGAACAAATTCATTTCGC 2603_a12.seq
12777 TTTTGCTCTTTTGCTCTTCTGCTACCAACACTCGAACAAATTCATTTCGC 18rs21_a12.seq
12683 CTCGAATGTTCTGATCT-AAGCGITCAATAITGGGGITAA--AGGTGACA 515_a12.seq
12674 TTTTGCTCTTTTGCTCTTCTGCTACCAACACTCGAACAAATTCATTTCGC cjb111_a12.seq
12714 TTTTGCTCTTTTGCTCTTCTGCTACCAACACTCGAACAAATTCATTTCGC h36b_a12.seq
ATAAATACTAAATATTT-GTATCGCTTCTTCTTACCATATTTTTTTATAA Majority
12960 12970 12980 12990 13000
12826 ATAAATACTAAATATTT-GTATCGCTTCTTCTTACCATATTTTTTTATAA 2603_a12.seq
12827 ATAAATACTAAATATTT-GTATCGCTTCTTCTTACCATATTTTTTTATAA 18rs21_a12.seq
12730 ATACCCGCTAAATATTTIATGTICIATGCTCTTIICIAAATGTGTAAT 515_a12.seq
12724 ATAAATACTAAATATTT-GTATCGCTTCTTCTTACCATATTTTTTTATAA cjb111_a12.seq
12764 ATAAATACTAAATATTT-GTATCGCTTCTTCTTACCATATTTTTTTATAA h36b_a12.seq
TATAGATCGCATTGCGTATCATGTAATATTTTCGAAATGGTGAATGATTC Majority
13010 13020 13030 13040 13050
12875 TATAGATCGCATTGCGTATCATGTAATATTTTCGAAATGGTGAATGATTC 2603_a12.seq
12876 TATAGATCGCATTGCGTATCATGTAATATTTTCGAAATGGTGAATGATTC 18rs21_a12.seq
12780 AACTGACGTGCTGCTTGTAATAAAACGAICACGAIAGATC--- 515_a12.seq
12773 TATAGATCGCATTGCGTATCATGTAATATTTTCGAAATGGTGAATGATTC cjb111_a12.seq
12813 TATAGATCGCATTGCGTATCATGTAATATTTTCGAAATGGTGAATGATTC h36b_a12.seq

FIGURE 19AB

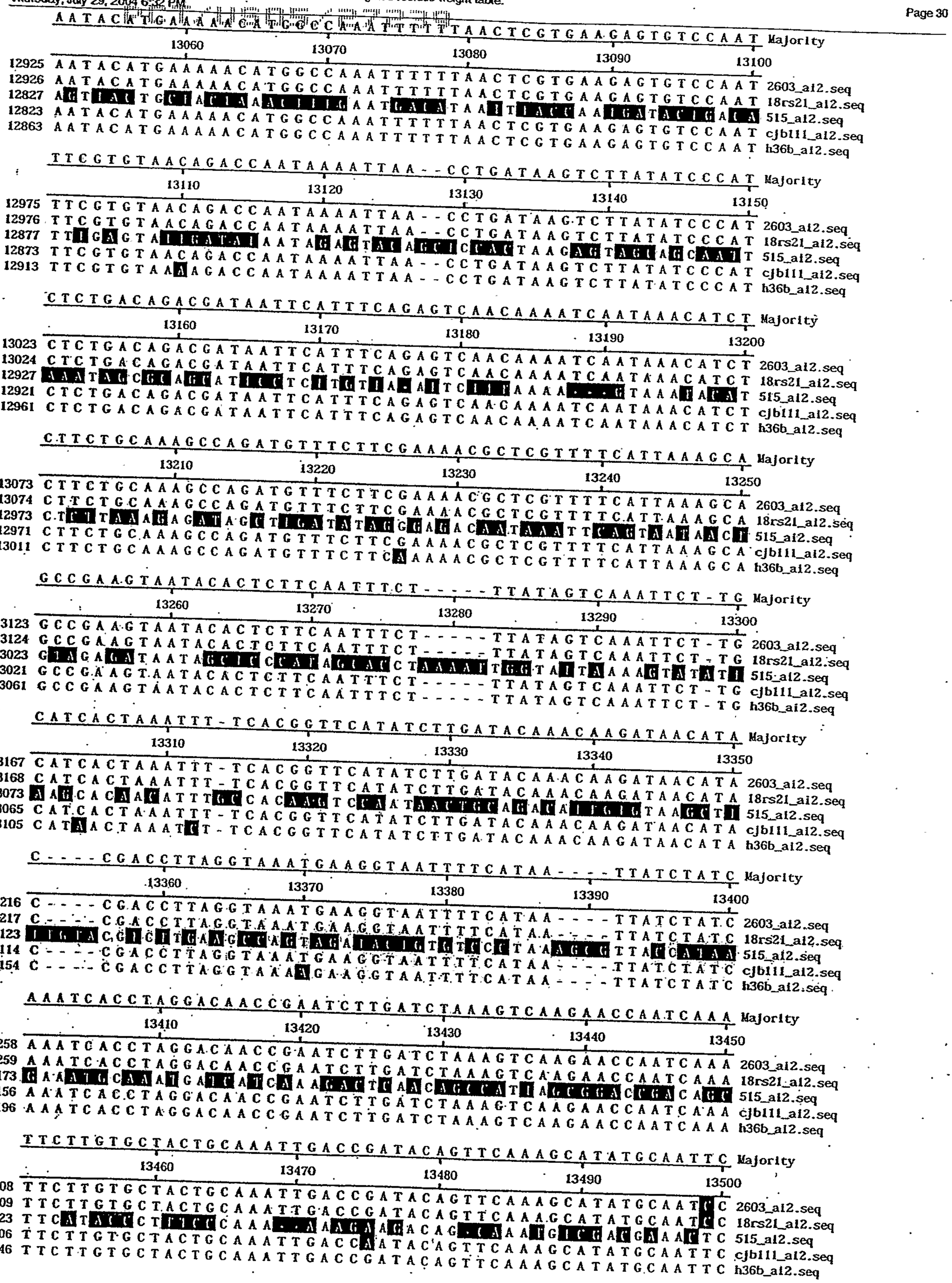

AATACATGAAAAACATGGCCAAATTTTTTAACTCGTGAAGAGTGTCCAAT Majority
13060 13070 13080 13090 13100
12925 AATACATGAAAAACATGGCCAAATTTTTTAACTCGTGAAGAGTGTCCAAT 2603_a12.seq
12926 AATACATGAAAAACATGGCCAAATTTTTTAACTCGTGAAGAGTGTCCAAT 18rs21_a12.seq
12827 AGTIACTGCIACIAAIIIGAATGAIATAAITIACGAIGATACIGACA 515_a12.seq
12823 AATACATGAAAAACATGGCCAAATTTTTTAACTCGTGAAGAGTGTCCAAT cjb111_a12.seq
12863 AATACATGAAAAACATGGCCAAATTTTTTAACTCGTGAAGAGTGTCCAAT h36b_a12.seq
TTCGTGTAACAGACCAATAAAATTAA--CCTGATAAGTCTTATATCCCAT Majority
13110 13120 13130 13140 13150
12975 TTCGTGTAACAGACCAATAAAATTAA--CCTGATAAGTCTTATATCCCAT
12976 TTCGTGTAACAGACCAATAAAATTAA--CCTGATAAGTCTTATATCCCAT 2603_a12.seq
12877 TTIGAGTAIIGATAIAATAGAGTAIAGCICACTAAGAGTAGIAGCAAIT 18rs21_a12.seq
12873 TTCGTGTAACAGACCAATAAAATTAA--CCTGATAAGTCTTATATCCCAT 515_a12.seq
12913 TTCGTGTAAAAGACCAATAAAATTAA--CCTGATAAGTCTTATATCCCAT cjb111_a12.seq
CTCTGACAGACGATAATTCATTTCAGAGTCAACAAAATCAATAAACATCT Majority
13160 13170 13180 13190 13200
13023 CTCTGACAGACGATAATTCATTTCAGAGTCAACAAAATCAATAAACATCT
13024 CTCTGACAGACGATAATTCATTTCAGAGTCAACAAAATCAATAAACATCT 2603_a12.seq
12927 AAATAGCGCAGTATICGTCITGTIA-AITCIITAAAA---GTAAATACAT 18rs21_a12.seq
12921 CTCTGACAGACGATAATTCATTTCAGAGTCAACAAAATCAATAAACATCT 515_a12.seq
12961 CTCTGACAGACGATAATTCATTTCAGAGTCAACAAAATCAATAAACATCT cjb111_a12.seq
CTTCTGCAAAGCCAGATGTTTCTTCGAAAACGCTCGTTTTCATTAAAGCA Majority
13210 13220 13230 13240 13250
13073 CTTCTGCAAAGCCAGATGTTTCTTCGAAAACGCTCGTTTTCATTAAAGCA
13074 CTTCTGCAAAGCCAGATGTTTCTTCGAAAACGCTCGTTTTCATTAAAGCA 2603_a12.seq
12973 CTCTTAAAGAGATAGCTIGATATAGGGAGACAATAAATTCAGTAATAACT 18rs21_a12.seq
12971 CTTCTGCAAAGCCAGATGTTTCTTCGAAAACGCTCGTTTTCATTAAAGCA 515_a12.seq
13011 CTTCTGCAAAGCCAGATGTTTCTTCAAAAACGCTCGTTTTCATTAAAGCA cjb111_a12.seq
GCCGAAGTAATACACTCTTCAATTTCT-----TTATAGTCAAATTCT-TG Majority
13260 13270 13280 13290 13300
13123 GCCGAAGTAATACACTCTTCAATTTCT-----TTATAGTCAAATTCT-TG
13124 GCCGAAGTAATACACTCTTCAATTTCT-----TTATAGTCAAATTCT-TG 2603_a12.seq
13023 GIAGAGATAATAGCICCATAGIACCTAAAATTGGTAITAAAGTATATI 515_a12.seq
13021 GCCGAAGTAATACACTCTTCAATTTCT-----TTATAGTCAAATTCT-TG
13061 GCCGAAGTAATACACTCTTCAATTTCT-----TTATAGTCAAATTCT-TG cjb111_a12.seq
CATCACTAAATTT-TCACGGTTCATATCTTGATACAAACAAGATAACATA Majority
13310 13320 13330 13340 13350
13167 CATCACTAAATTT-TCACGGTTCATATCTTGATACAAACAAGATAACATA
13168 CATCACTAAATTT-TCACGGTTCATATCTTGATACAAACAAGATAACATA 2603_a12.seq
13073 AAGCACAACATTTGCCACAAGTCCAATAACTGCAGACAIIGIGTAAGCTI 18rs21_a12.seq
13065 CATCACTAAATTT-TCACGGTTCATATCTTGATACAAACAAGATAACATA 515_a12.seq
13105 CATAACTAAATGT-TCACGGTTCATATCTTGATACAAACAAGATAACATA cjb111_a12.seq
C----CGACCTTAGGTAAATGAAGGTAATTTTCATAA----TTATCTATC Majority
13360 13370 13380 13390 13400
13216 C----CGACCTTAGGTAAATGAAGGTAATTTTCATAA----TTATCTATC
13217 C----CGACCTTAGGTAAATGAAGGTAATTTTCATAA----TTATCTATC 2603_a12.seq
13123 IIGIAC,GICTTGAAGCCAGTAGAIACIGTGTCCTAAAGCGTTACCATAA 18rs21_a12.seq
13114 C----CGACCTTAGGTAAATGAAGGTAATTTTCATAA----TTATCTATC 515_a12.seq
13154 C----CGACCTTAGGTAAAAGAAGGTAATTTTCATAA----TTATCTATC cjb111_a12.seq
AAATCACCTAGGACAACCGAATCTTGATCTAAAGTCAAGAACCAATCAAA Majority
13410 13420 13430 13440 13450
13258 AAATCACCTAGGACAACCGAATCTTGATCTAAAGTCAAGAACCAATCAAA
13259 AAATCACCTAGGACAACCGAATCTTGATCTAAAGTCAAGAACCAATCAAA 2603_a12.seq
13173 GAAATGCAAATGATCATCAAAGACTCAACAGCCATIAGCGGACTGACAGC 18rs21_a12.seq
13156 AAATCACCTAGGACAACCGAATCTTGATCTAAAGTCAAGAACCAATCAAA 515_a12.seq
13196 AAATCACCTAGGACAACCGAATCTTGATCTAAAGTCAAGAACCAATCAAA cjb111_a12.seq
TTCTTGTGCTACTGCAAATTGACCGATACAGTTCAAAGCATATGCAATTC Majority
13460 13470 13480 13490 13500
13308 TTCTTGTGCTACTGCAAATTGACCGATACAGTTCAAAGCATATGCAATTC
13309 TTCTTGTGCTACTGCAAATTGACCGATACAGTTCAAAGCATATGCAATTC 2603_a12.seq
13223 TTCATACCCTTTCCCAAA--AAAGAAGACAGCAAATGICGAACTGAAATTC 18rs21_a12.seq
13206 TTCTTGTGCTACTGCAAATTGACCAATACAGTTCAAAGCATATGCAATTC 515_a12.seq
13246 TTCTTGTGCTACTGCAAATTGACCGATACAGTTCAAAGCATATGCAATTC cjb111_a12.seq
h36b_a12.seq

FIGURE 19AC

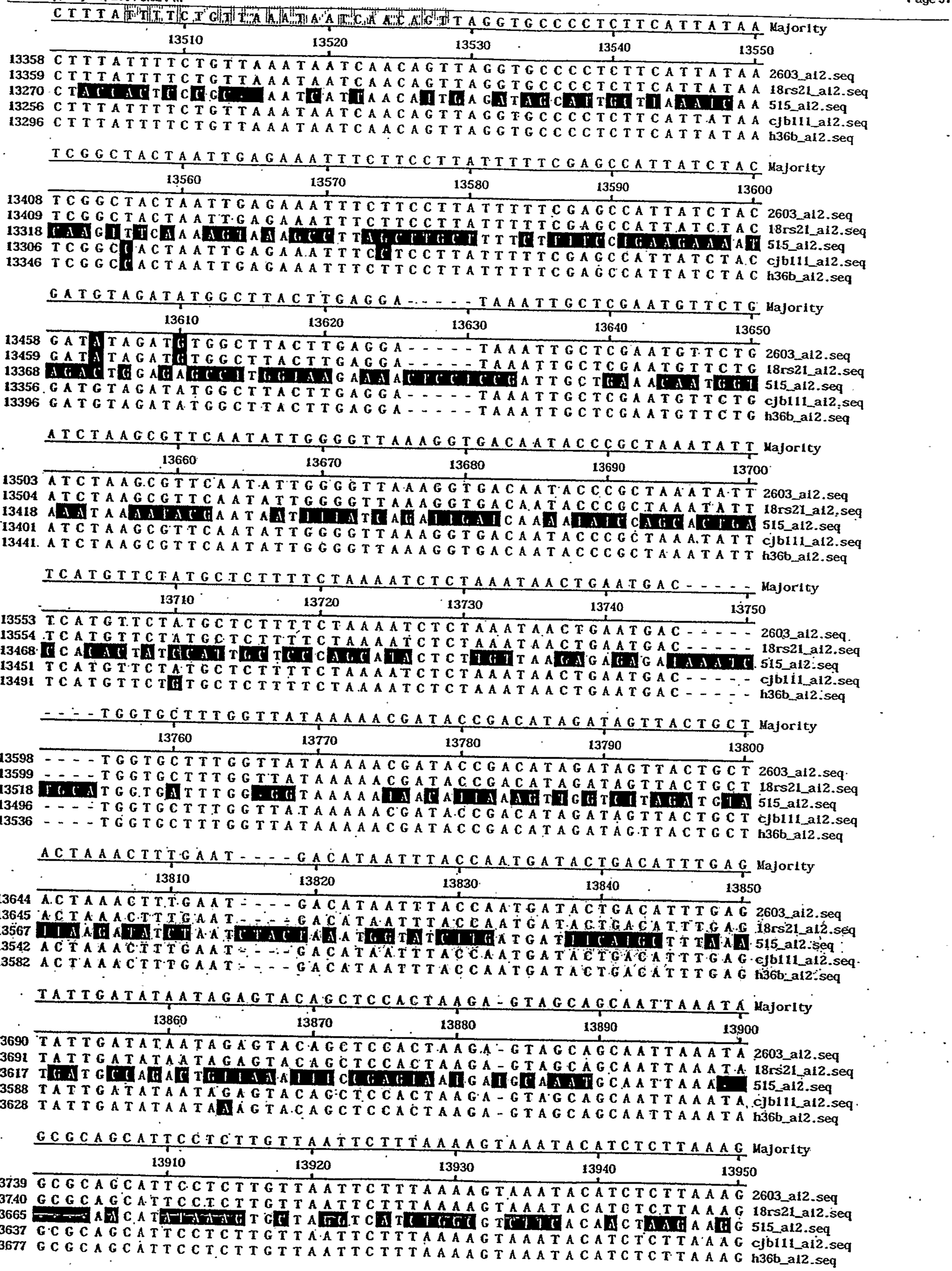
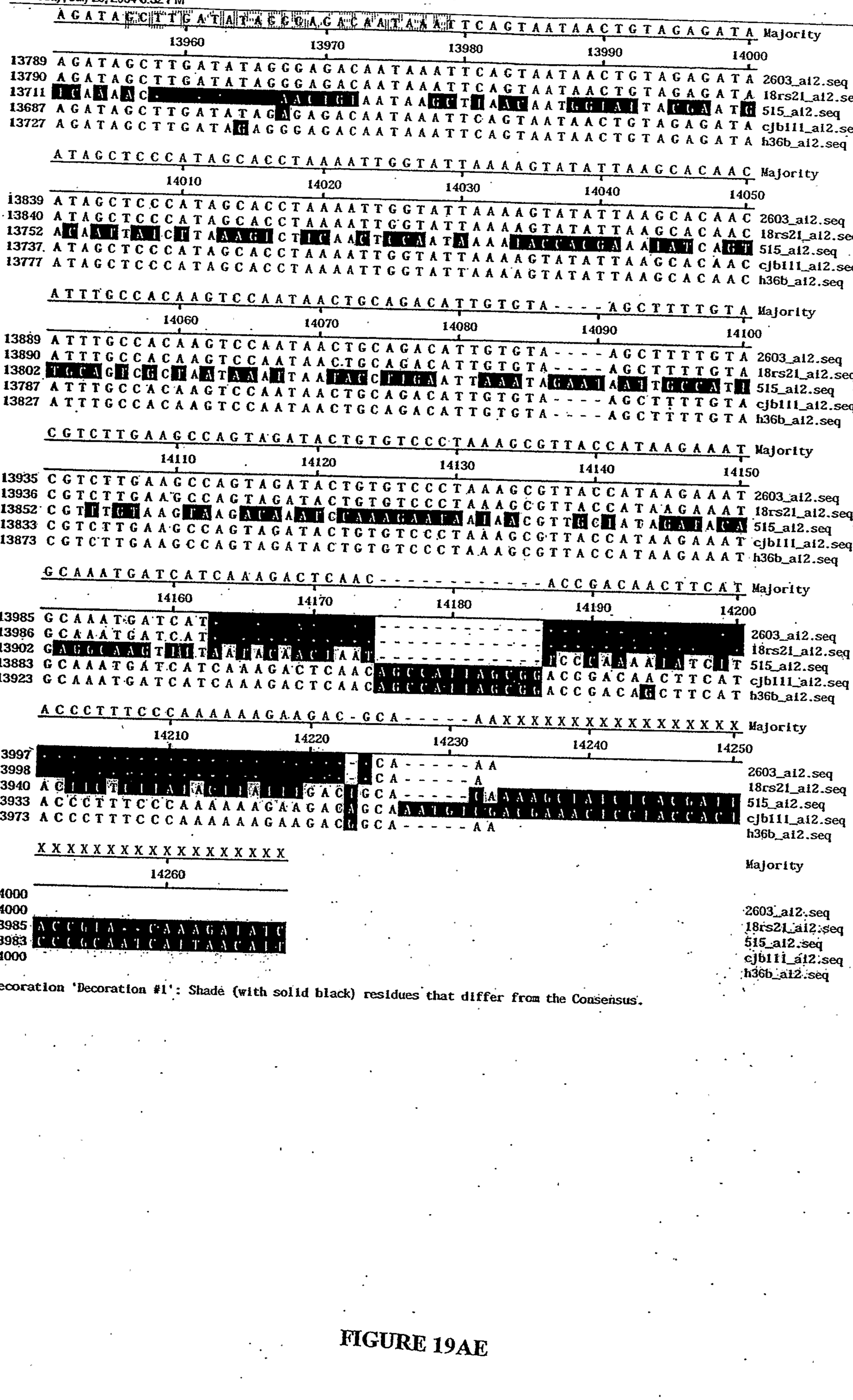

CTTTATTTTCTGTTAAATAATCAACAGTTAGGTGCCCCTCTTCATTATAA Majority
13510 13520 13530 13540 13550
13358 CTTTATTTTCTGTTAAATAATCAACAGTTAGGTGCCCCTCTTCATTATAA
13359 CTTTATTTTCTGTTAAATAATCAACAGTTAGGTGCCCCTCTTCATTATAA 2603_a12.seq
13270 CTACTACTCCGCG.AATCATTAACATTGAGATAGCATTGTTAAATCAA 18rs21_a12.seq
13256 CTTTATTTTCTGTTAAATAATCAACAGTTAGGTGCCCCTCTTCATTATAA 515_a12.seq
13296 CTTTATTTTCTGTTAAATAATCAACAGTTAGGTGCCCCTCTTCATTATAA cjb111_a12.seq
TCGGCTACTAATTGAGAAATTTCTTCCTTATTTTTCGAGCCATTATCTAC Majority
13560 13570 13580 13590 13600
13408 TCGGCTACTAATTGAGAAATTTCTTCCTTATTTTTCGAGCCATTATCTAC
13409 TCGGCTACTAATTGAGAAATTTCTTCCTTATTTTTCGAGCCATTATCTAC 2603_a12.seq
13318 CAAGTTCAAAAGTAAAGCCTTAGCTTGCTTTTCTTTCGAGCCATTATCTAC 18rs21_a12.seq
13306 TCGGCCACTAATTGAGAAATTTCGTCCTTATTTTTCGAGCCATTATCTAC 515_a12.seq
13346 TCGGCCACTAATTGAGAAATTTCTTCCTTATTTTTCGAGCCATTATCTAC cjb111_a12.seq
13408 TCGGCTACTAATTGAGAAATTTCTTCCTTATTTTTCGAGCCATTATCTAC h36b_a12.seq
GATGTAGATATGGCTTACTTGAGGA------TAAATTGCTCGAATGTTCTG Majority
13610 13620 13630 13640 13650
13458 GATATAGATGTGGCTTACTTGAGGA------TAAATTGCTCGAATGTTCTG
13459 GATATAGATGTGGCTTACTTGAGGA------TAAATTGCTCGAATGTTCTG 2603_a12.seq
13368 AGACTGGAGAGCCTTGGTAAGAAAACTCCTCCGATTGCTGAAATAATGGT 18rs21_a12.seq
13356 GATGTAGATATGGCTTACTTGAGGA------TAAATTGCTCGAATGTTCTG 515_a12.seq
13396 GATGTAGATATGGCTTACTTGAGGA------TAAATTGCTCGAATGTTCTG cjb111_a12.seq
13458 GATATAGATGTGGCTTACTTGAGGA------TAAATTGCTCGAATGTTCTG h36b_a12.seq
ATCTAAGCGTTCAATATTGGGGTTAAAGGTGACAATACCCGCTAAATATT Majority
13660 13670 13680 13690 13700
13503 ATCTAAGCGTTCAATATTGGGGTTAAAGGTGACAATACCCGCTAAATATT
13504 ATCTAAGCGTTCAATATTGGGGTTAAAGGTGACAATACCCGCTAAATATT 2603_a12.seq
13418 AAATAAAATACGAATAATTTATCAGATTGAICAAAATATCCAGCACTGA 18rs21_a12.seq
13401 ATCTAAGCGTTCAATATTGGGGTTAAAGGTGACAATACCCGCTAAATATT 515_a12.seq
13441 ATCTAAGCGTTCAATATTGGGGTTAAAGGTGACAATACCCGCTAAATATT cjb111_a12.seq
13503 ATCTAAGCGTTCAATATTGGGGTTAAAGGTGACAATACCCGCTAAATATT h36b_a12.seq
TCATGTTCTATGCTCTTTTCTAAAATCTCTAAATAACTGAATGAC------ Majority
13710 13720 13730 13740 13750
13553 TCATGTTCTATGCTCTTTTCTAAAATCTCTAAATAACTGAATGAC------
13554 TCATGTTCTATGCTCTTTTCTAAAATCTCTAAATAACTGAATGAC------ 2603_a12.seq
13468 CCATACTATGCATTGCTCCCAGCATACTCTTGTTAAGAGAGAGAAAATC 18rs21_a12.seq
13451 TCATGTTCTATGCTCTTTTCTAAAATCTCTAAATAACTGAATGAC------ 515_a12.seq
13491 TCATGTTCTGTGCTCTTTTCTAAAATCTCTAAATAACTGAATGAC------ cjb111_a12.seq
13553 TCATGTTCTATGCTCTTTTCTAAAATCTCTAAATAACTGAATGAC------ h36b_a12.seq
----TGGTGCTTTGGTTATAAAAACGATACCGACATAGATAGTTACTGCT Majority
13760 13770 13780 13790 13800
13598 ----TGGTGCTTTGGTTATAAAAACGATACCGACATAGATAGTTACTGCT
13599 ----TGGTGCTTTGGTTATAAAAACGATACCGACATAGATAGTTACTGCT 2603_a12.seq
13518 TGCATGGTGATTTGGAGGTAAAAAIAACAIIAAGTIGGTCITAGATGCTA 18rs21_a12.seq
13496 ----TGGTGCTTTGGTTATAAAAACGATACCGACATAGATAGTTACTGCT 515_a12.seq
13536 ----TGGTGCTTTGGTTATAAAAACGATACCGACATAGATAGTTACTGCT cjb111_a12.seq
13598 ----TGGTGCTTTGGTTATAAAAACGATACCGACATAGATAGTTACTGCT h36b_a12.seq
ACTAAACTTTGAAT----GACATAATTTACCAATGATACTGACATTTGAG Majority
13810 13820 13830 13840 13850
13644 ACTAAACTTTGAAT----GACATAATTTACCAATGATACTGACATTTGAG
13645 ACTAAACTTTGAAT----GACATAATTTACCAATGATACTGACATTTGAG 2603_a12.seq
13567 IIAAGATATCTAATCTACTAAATGGTATCITGATGATIICAIGCTTTAAA 18rs21_a12.seq
13542 ACTAAACTTTGAAT----GACATAATTTACCAATGATACTGACATTTGAG 515_a12.seq
13582 ACTAAACTTTGAAT----GACATAATTTACCAATGATACTGACATTTGAG cjb111_a12.seq
13644 ACTAAACTTTGAAT----GACATAATTTACCAATGATACTGACATTTGAG h36b_a12.seq
TATTGATATAATAGAGTACAGCTCCACTAAGA-GTAGCAGCAATTAAATA Majority
13860 13870 13880 13890 13900
13690 TATTGATATAATAGAGTACAGCTCCACTAAGA-GTAGCAGCAATTAAATA
13691 TATTGATATAATAGAGTACAGCTCCACTAAGA-GTAGCAGCAATTAAATA 2603_a12.seq
13617 TGATGCCAGACTGIIAAAIIICCGAGIAIGAIGCAAATGCAATTAAA. 18rs21_a12.seq
13588 TATTGATATAATAGAGTACAGCTCCACTAAGA-GTAGCAGCAATTAAATA 515_a12.seq
13628 TATTGATATAATAAAGTACAGCTCCACTAAGA-GTAGCAGCAATTAAATA cjb111_a12.seq
13690 TATTGATATAATAGAGTACAGCTCCACTAAGA-GTAGCAGCAATTAAATA h36b_a12.seq
GCGCAGCATTCCTCTTGTTAATTCTTTAAAAGTAAATACATCTCTTAAAG Majority
13910 13920 13930 13940 13950
13739 GCGCAGCATTCCTCTTGTTAATTCTTTAAAAGTAAATACATCTCTTAAAG
13740 GCGCAGCATTCCTCTTGTTAATTCTTTAAAAGTAAATACATCTCTTAAAG 2603_a12.seq
13665 ------AACATAIAAAGTGCTAGGTCATCTTGGCGTCTTCACACTAAGAAGG 18rs21_a12.seq
13637 GCGCAGCATTCCTCTTGTTAATTCTTTAAAAGTAAATACATCTCTTAAAG 515_a12.seq
13677 GCGCAGCATTCCTCTTGTTAATTCTTTAAAAGTAAATACATCTCTTAAAG cjb111_a12.seq
13739 GCGCAGCATTCCTCTTGTTAATTCTTTAAAAGTAAATACATCTCTTAAAG h36b_a12.seq

FIGURE 19AD



Decoration 'Decoration #1': Shade (with solid black) residues that differ from the Consensus.

FIGURE 19AE

```
GTTGCTTCATAAGTTGTCTGAAGGCCTAATTCTAAAGTCACATGCATTCT Majority
          10        20        30        40        50
1         G
1         GTTGCTTCATAAGTTGTCTGAAGGCCTAATTCTAAAGTCACATGCATTCT 2603_a12.seq
1         GTTGCTTCATAAGTTGTCTGAAGGCCTAATTCTAAAGTCACATGCATTCT nem316_a12.seq

TTCAGAAAGTTCAGCGAGATAGTATATAGTTTCATCAGGTAAGCAATCCG Majority
          60        70        80        90       100
2
51        TTCAGAAAGTTCAGCGAGATAGTATATAGTTTCATCAGGTAAGCAATCCG 2603_a12.seq
51        TTCAGAAAGTTCAGCGAGATAGTATATAGTTTCATCAGGTAAGCAATCCG nem316_a12.seq

GCCTTGTTCCGATGTTGATTCCGATAACTCCTGGCTCATTAATAGCCTGT Majority
         110       120       130       140       150
2         GCCTTGTTCCGATGTTGATTCCGATAACTCCTGGCTCATTAATAGCCTGT 2603_a12.seq
101       GCCTTGTTCCGATGTTGATCCCGATAACTCCTGGCTCATTAATAGCCTGT nem316_a12.seq

TCGTAACGCTCTTTAATTATCTCTAACTTAGCATGGGTATTGGTAAAATT Majority
         160       170       180       190       200
52        TCGTAACGCTCTTTAATTATCTCTAACTTAGCATGGGTATTGGTAAAATT 2603_a12.seq
151       TCGTAACGCTCTTTAATTATCTCTAACTTAGCATGGGTATTGGTAAAATT nem316_a12.seq

TTGAAAATAGACTAAGTATTTATTAACCTCGGGCCACTTTCTATGCATGA Majority
         210       220       230       240       250
102       TTGAAAATAGACTAAGTATTTATTAACCTCAGGCCACTTTCTATGCATGA 2603_a12.seq
201       TTGAAAATAGACTAAGTATTTATTAACCTCGGGCCACTTTCTATGCATGA nem316_a12.seq

AATCAATTTCTTTATAGAATTGTTCACGAATAGGAGCTTCTGGAGCAACT Majority
         260       270       280       290       300
152       AATCAATTTCTTTATAGAATTGTTCACGAATAGGAGCTTCTGGAGCAACT 2603_a12.seq
251       AATCAATTTCTTTATAGAATTGTTCACGAATAGGAGCTTCTGGAGCAACT nem316_a12.seq

ATAGCATCCCCTGAACCAGAAACTGTGCAAAAAGTGCATCCTCCTCTAGC Majority
         310       320       330       340       350
202       ATAGCATCCCCTGAACCAGAAACTGTGCAAAAAGTGCACCTCCTCTAGC 2603_a12.seq
301       ATAGCATCCCCTGAACCAGAAACTGTGCAAAAAGTGCATCCTCCTCTAGC nem316_a12.seq

AACTGTTCCGTCTCTGTTAGGACAGTCAAAACCAGCATCTATAGGTAATT Majority
         360       370       380       390       400
252       AACTGTTCCGTCTCTGTTAGGACAGTCAAAACCAGCATCTATAGGTAATT 2603_a12.seq
351       AACTGTTCCGTCTCTGTTAGGACAGTCAAAACCAGCATCTATAGGTAATT nem316_a12.seq

TAAATATTTTTTCTCCAAAGAGTTCTCGATAATAATCATTAATCGCACGA Majority
         410       420       430       440       450
302       TAAATATTTTTTCTCCAAAGAGTTCTCGATAATAATCATTAATCGCACGA 2603_a12.seq
401       TAAATATTTTTTCTCCAAAGAGTTCTCGATAATAATCATTAATCGCACGA nem316_a12.seq

TAACGTTTTTTCATAGGATAATTGTATCACAATTTTAACTAAAATAACCT Majority
         460       470       480       490       500
352       TAACGTTTTTTCATAGGATAATTGTATCACAATTTTAACTAAAATAACCT 2603_a12.seq
451       TAACGTTTTTTCATAGGATAATTGTATCACAATTTTAACTAAAATAACCT nem316_a12.seq

CACTACTACAATAAAACTTAAAAAGATTGGAACGTCAGTTAGTCCCAATC Majority
         510       520       530       540       550
402       CACTACTACAATAAAACTAAAAAAGATTGGAACGTCAGTTAGTCCCAATC 2603_a12.seq
501       CACTACTACAATAAAACTTAAAAAGATTGGAACGTCAGTTAGTCCCAATC nem316_a12.seq

TTTTATTTACTTCACTTTCTTTAACCAATCCTTGGCTAAAAAGATATACG Majority
         560       570       580       590       600
452       TTTTATTTACTTCACTTTCTTTAACCAATCCTTGGCTAAAAAGATATACG 2603_a12.seq
551       TTTTATTTACTTCACTTTCTTTAACCAATCCTTGGCTAAAAAGATATACG nem316_a12:seq

CAGTTAGATTCAAAATACCATAAGCAAGTATAAAACCAGCTAAAACATCT Majority
         610       620       630       640       650
502       CAGTTAGATTCAAAATACCATAAGCAAGTATAAAACCAGCTAAAACATCT 2603_a12.seq
601       CAGTTAGATTCAAAATACCATAAGCAAGTATAAAACCAGCTAAAACATCT nem316_a12.seq
```

Figure 20

```
GTCGGAAAATGAACCCCTAGGTAAATACGAGATAACCCAATTAAAAAAAT Majority
        660       670       680       690       700
552 GTCGGAAAATGAACCCCTAGGTAAATACGAGATAACCCAATTAAAAAAAT 2603_a12.seq
651 GTCGGAAAATGAACCCCTAGGTAAATACGAGATAACCCAATTAAAAAAAT nem316_a12.seq

GAGCAAACCCAAAGTACCTTGGCACAACAGTTTCCATATACTCTTAGGCA Majority
        710       720       730       740       750
602 GAGCAAACCCAAAGTACCTTGGCACAACAGTTTCCATATACTCTTAGGCA 2603_a12.seq
701 GAGCAAACCCAAAATACCTTGGCACAACAGTTTCCATATACTCTTAGGCA nem316_a12.seq

TATAGTACTGCAATAAAATAATAATACTCCCAAATATCATAAATGTTCCC Majority
        760       770       780       790       800
652 TATAGTACTGCAATAAAATAATAATACTCCCAAATATCATAAATGTTCCC 2603_a12.seq
751 TATAGTACTGCAATAAAATAATAATACTCCCAAATATCATAAATGTTCCC nem316_a12.seq

ATCGAGTGCCCACTGGGAAACGAATAGCCACCTGCAAATACTAAATGGGT Majority
        810       820       830       840       850
702 ATCGAGTGCCCACTGGGAAACGAATAGCCACCTGCAAATACTAAATGGGT 2603_a12.seq
801 ATCGAGTGCCCACTGGGAAACGAATAGCCACCTGCAAATACTAAATGGGT nem316_a12.seq

TAAAGTTGGTCTTACTCTTTGAAAAATAAGTTTTAAAGAAGTATACATA Majority
        860       870       880       890       900
752 TAAAGTTGGTCTTACTCTTTGAAAAATAAGTTTTAAAGAAAGTATACATA 2603_a12.seq
851 TAAAGTTGGTCTCACTCTTTGAAAAATAAGTTTTAAAGAAAGTATACATA nem316_a12.seq

TACCAGAGATAATAGCATTTACTGCGATAAATCTAGCTTGAGGATACCAC Majority
        910       920       930       940       950
802 TACCAGAGATAATAGCATTTACTGCGATAAATCTAGCTTGAGGATACCAC 2603_a12.seq
901 TACCAGAGATAATAGCATTTACTGCGATAAATCTAGCTTGAGGATACCAC nem316_a12.seq

TTCTTAAGGTAACAGAAAGTGACGCTCATAATCGCAATAGCTATCTGGCT Majority
        960       970       980       990      1000
852 TTCTTAAGGTAACAGAAAGTGACGCTCATAATCGCAATAGCTATCTGGCT 2603_a12.seq
951 TTCTTAAGGTAACAGAAAGTGACGCTCATAATCGCAATAGCTATCTGGCT nem316_a12.seq

TACAGTATTACCAATCACAGTGATTAACTTGAAAAATCTTGTAGAAAGAT Majority
       1010      1020      1030      1040      1050
902 TACAGTATTACCAATCACAGTGATTAACTTGAAAAATCTTGTAGAAAGAT 2603_a12.seq
1001 TACAGTATTACCAATCACAGTGATTAACTTGAAAAATCTTGTAGAAAGAT nem316_a12.seq

TTGGCAACTGTCCTCTAACACTTTCTTGAATGTTTTGGTCAAATGCAATT Majority
       1060      1070      1080      1090      1100
952 TTGGCAACTGTCCTCTAACACTTTCTTGAATAGTTTGGTCAAATGAAATT 2603_a12.seq
1051 TTGGCAACTGTCCTCTAACACTTTCTTGAATGTTTTGGTCAAATGCAATT nem316_a12.seq

ACAGTGTCGGGCCAATATTTGATGACCAATCCTAAACTGAAAAATAAGAT Majority
       1110      1120      1130      1140      1150
1002 ACAGTGTCGGGCCAATATTTGATGACCAATCCTAAACTGAAAAATAAGAT 2603_a12.seq
1101 ACAGTGTCGGGCCAATATTTGATGACCAATCCTAAACTGAAAAATAAGAT nem316_a12.seq

AATAGCAATAAATGCTTGAATAAGTTTACTATTTTGACGAGATAACATTA Majority
       1160      1170      1180      1190      1200
1052 AATAGCAATAAATGCTTGAATAAGTTTACTATTTTGACGAGATAACATTA 2603_a12.seq
1151 AATAGCAATAAATGCTTGAATAAGTTTACTATTTTGACGAGATAACATTA nem316_a12.seq

GTCTTTTTATATCTTTCTAATATTGGCAAACAAGCCACGTAAGTTAGATA Majority
       1210      1220      1230      1240      1250
1102 GTCTTTTTATATCTTTCTAATATTGGCAAACAAGCCACGTAAGTTAGATA 2603_a12.seq
1201 GTCTTTTTATATCTTTCTAATATTGGCAAACAAGCCACGTAAGTTAGATA nem316_a12.seq

GAAAACAATCGAAATTAAAATTCCCTCAACGATATTAAATGGAATAACCA Majority
       1260      1270      1280      1290      1300
1152 GAAAACAATCGAAATTAAAATTCCCTCAACGATATTAAATGGAATAACCA 2603_a12.seq
1251 GAAAACAATCGAAATTAAAATTCCCTCAACGATATTAAATGGAATAACCA nem316_a12.seq
```

FIGURE 20A

```
                TTGTTAAAAGGTAATTGCCTACACCAATAAATGTTCTGATATCAAAGTTA Majority
                        1310        1320        1330        1340        1350
1202 TTGTTAAAAGGTAATTGCCTACACCAATAAATGTTCTGATATCAAAGTTA 2603_a12.seq
1301 TTGTTAAAAGGTAATTGCCTACACCAATAAATGTTCTGATATCAAAGTTA nem316_a12.seq

                GCAAATATAGCATACAAAGGAATCGCAAAGACATAGTTGAGAGCTACCAT Majority
                        1360        1370        1380        1390        1400
1252 GCAAATATAGCATACAAAGGAATCGCAAAGACATAGTTGAGAGCTACCAT 2603_a12.seq
1351 GCAAATATAGCATACAAAGGAATCGCAAAGACATAGTTGAGAGCTACCAT nem316_a12.seq

                AGATACGGTCAAGCTAACTGTACCAAATAGACTAGCTTTAATAAAATCTT Majority
                        1410        1420        1430        1440        1450
1302 AGATACGGTCAAGCTAACTGTACCAAATAGACTAGCTTTAATAAAATCTT 2603_a12.seq
1401 AGATACNGTCAAGCTAACTGTACCAAATAGACTAGCTTTAATAAAATCTT nem316_a12.seq

                TTGCACTCTCTCTATTTTTCCAGAAAATAGCGAAACTTGCTAAAAATAAA Majority
                        1460        1470        1480        1490        1500
1352 TTGCACTCTCTCTATTTTTCCAGAAAATAGCGAAACTTGCTAAAAATAAA 2603_a12.seq
1451 TTGCACTCTCTCTATTTTTCCAGAAAATAGCGAAACTTGCTAAAAATAAA nem316_a12.seq

                GCTAGAGCAACCATATTCATCGGTAAACCGATAAAGGTTTCTGGACCACG Majority
                        1510        1520        1530        1540        1550
1402 GCTAGAGCAACCATATTCATCGGTAAACCNATAAANGTTTCTGGACCACG 2603_a12.seq
1501 GCTAGAGCAACCATATTCATCGGTAAACCGATAAAGGTTTCTGGACCACG nem316_a12.seq

                ATTAGCAAGTATAACTTTTAAAAGTGATCTTAATAAGAGTACACCATAAC Majority
                        1560        1570        1580        1590        1600
1452 ATTAGCAAGTATAACTTTTAAAAGTGATCTTAATAAGAGTACACCATAAC 2603_a12.seq
1551 ATTAGCAAGTATAACTTTTAAAAGTGATCTTAATAAGAGTACACCATAAC nem316_a12.seq

                TTGATTTCAAATCAAATAAAATAAAAGCAACTAACATCGGAAGGATTGAA Majority
                        1610        1620        1630        1640        1650
1502 TTGATTTCAAATCAAATAAAATAAAAGCAACTAACATCGGAAGGATTGAA 2603_a12.seq
1601 TTGATTTCAAATCAAATAAAATAAAAGCAACTAACATCGGAAGGATTGAA nem316_a12.seq

                AAATCAACCTTTAAAAATTCTGCTCCTGGTATTAATGGAAATGAAACCAT Majority
                        1660        1670        1680        1690        1700
1552 AAATCAACCTTTAAAAATTCTGCTCCTGGTATTAATGGAAATGAAACCAT 2603_a12.seq
1651 AAATCAACCTTTAAAAATTCTGCTCCTGGTATTAATGGAAATGAAACCAT nem316_a12.seq

                CATCAATACAAAAGATAAGGCAGAAAGAATGGCGATTGTCACCATTTTAC Majority
                        1710        1720        1730        1740        1750
1602 CATCAATACAAAAGATAAGGCAGAAAGAATGGCGATTGTCACCATTTTAC 2603_a12.seq
1701 CATCAATACAAAAGATAAGGCAGAAAGAATGGCGATTGTCACCATTTTAC nem316_a12.seq

                GTGTATTTGTCATAAAAAAATTCCTCCAATTTAAATAAATTGAAAGAAGC Majority
                        1760        1770        1780        1790        1800
1652 GTGTATTTGTCATAAAAAAATTCCTCCAATTTAAATAAATTGAAAGAAGC 2603_a12.seq
1751 GTGTATTTGTCATAAAAAAATTCCTCCAATTTAAATAAATTGAAAGAAGC nem316_a12.seq

                TCCAAAGGTAAGCGTATGTACGCGAAAAAAACCTTTGTCTTCTCCCATCC Majority
                        1810        1820        1830        1840        1850
1702 TCCAAAGGTAAGCGTATGTACGCGAAAAAAN.CCTTTGTCTTCTCCCATCC 2603_a12.seq
1801 TCCAAAGGTAAGCGTANGTACGCGAAAAAAACCTTTGTCTTCTCCCATCC nem316_a12.seq

                AGACTTTACTGTCGGTTGTGGAATCTCACCACATCAGCTTTCGCTCGCGG Majority
                        1860        1870        1880        1890        1900
1751 AGACTTTACTGTCGGTTGTGGAATCTCACCACATCAGCTTTCGCTCGCGG 2603_a12.seq
1851 AGACTTTACTGTCGGTTGTGGAATCTCACCACATCAGCTTTCGCTCGCGG nem316_a12.seq

                ACTGATGCTTCACAACTGACAAATAAGTTGGAAGCGATTACCGCCGGTCG Majority
                        1910        1920        1930        1940        1950
1801 ACTGATGCTTCACAACTGACAAATAAGTTGGAAGCGATTACCGCCGGTCG 2603_a12.seq
1901 ACTGATGCTTCACAACTGACAAATAAGTTGGAAGCGATTACCGCCGGTCG nem316_a12.seq
```

FIGURE 20B

```
      GGAATTACACCCTGCCCTGAAGACACCTATAGCATAACAAAAAAAACTTG Majority
                1960      1970      1980      1990      2000
1851  GGAATTACACCCTGCCCTGAAGACACCTATAGCATAACAAAAAAAACTTG 2603_a12.seq
1951  GGAATTACACCCTGCCCTGAAGACACCTATAGCATAACAAAAAAAACTTG nem316_a12.seq

      CAATTGCAAGTTTTTTAATTACTAATTAGTAGTAGTGATTAAAAATCATA Majority
                2010      2020      2030      2040      2050
1901  CAATTGCAAGTTTTTTAATTACTAATTAGTAGTAGTGATTAAAAATCATA 2603_a12.seq
2001  CAATTGCAAGTTTTTTAATTACTAATTAGTAGTAGTGATTAAAAATCATA nem316_a12.seq

      TTAATACCAAATTACTATGCTGTATCGTTTCTTTCAGATTTGCTATTTTT Majority
                2060      2070      2080      2090      2100
1951  TTAATACCAAATTACTATACTGTATCGTTTCTTTCAGATTTGCTATTTTT 2603_a12.seq
2051  TTAATACCAAATTACTATGCTGTATCGTTTCTTTCAGATTTGCTATTTTT nem316_a12.seq

      AGTTTTTCTTAAAAAGATAAACAAAATTCCCAAAATAATACAACCAAGAA Majority
                2110      2120      2130      2140      2150
2001  AGTTTTTCTTAAAAAGATAAACAAAATTCCCAAAATAATACAACCAAGAA 2603_a12.seq
2101  AGTTTTTCTTAAAAAGATAAACAAAATTCCCAAAATAATACAACCAAGAA nem316_a12.seq

      TTGTCAGTCCTCCACCAATAATCATTCCTGTTTTAGGAAGAAATGATTGT Majority
                2160      2170      2180      2190      2200
2051  TTGTCAGTCCTCCACCAATAATCATTCCTGTTTTAGGAAGAAATGATTGT 2603_a12.seq
2151  TTGTCAGTCCTCCACCAATAATCATTCCTGTTTTAGGAAGAAATGATTGT nem316_a12.seq

      GGAAAAAGCGGTTGTGATGGTTTAGGATTTGTTGGTGGAGGAGTTTCTTT Majority
                2210      2220      2230      2240      2250
2101  GGAAAAAGCGGTTGTGATGGTTTAGGATTTGTTGGTGGAGGAGTTTCTTT 2603_a12.seq
2201  GGAAAAAGCGGTTGTGATGGTTTAGGATTTGTTGGTGGAGGAGTTTCTTT nem316_a12.seq

      TTCGTTTTCTACCTCTACTTCCTGTGTTTTATTAGCAACTACAGCAACTA Majority
                2260      2270      2280      2290      2300
2151  TTCGTTTTCTACCTCTACTTCCTGTGTTTTATTAGCAACTACAGCAACTA 2603_a12.seq
2251  TTCGTTTTCTACCTCTACTTCCTGTGTTTTATTAGCAACTACAGCAACTA nem316_a12.seq

      CAGCATCCTTCATAGATATACGGTAACCAGTTAGTGCTTTTGCTTCTCGA Majority
                2310      2320      2330      2340      2350
2201  CAGCATCCTTCATAGATATACGGTAACCAGTTAGTGCTTTTGCTTCTCGA 2603_a12.seq
2301  CAGCATCCTTCATAGATATACGGTAACCAGTTAGTGCTTTTGCTTCTCGA nem316_a12.seq

      AAAATATACTTACCAGGTAATAAACCTTCAACCTCAATTTCTCCCTTATC Majority
                2360      2370      2380      2390      2400
2251  AAAATATACTTACCAGGTAATAAACCTTCAACCTCAATTTCTCCCTTATC 2603_a12.seq
2351  AAAATATACTTACCAGGTAATAAACCTTCAACCTCAATTTCTCCCTTATC nem316_a12.seq

      ATCAGTTACTAATGAAGTAATCCCATCTTGATCGGTCGTAAATCGTCCAT Majority
                2410      2420      2430      2440      2450
2301  ATCAGTTACTAATGAAGTAATCCCATCTTGATCGGTCGTAAATCGTCCAT 2603_a12.seq
2401  ATCAGTTACTAATGAAGTAATCCCATCTTGATCGGTCGTAAATCGTCCAT nem316_a12.seq

      TTTTAAAGCGAACTGGCTGATTCTGGTTATCGTATAATACAAATATTACT Majority
                2460      2470      2480      2490      2500
2351  TTTTAAAGCGAACTGGCTGATTCTGGTTATCGTATAATACAAATATTACT 2603_a12.seq
2451  TTTTAAAGCGAACTGGCTGATTCTGGTTATCGTATAATACAAATATTACT nem316_a12.seq

      CCTGATAGCCTTTTCTTTATCTTTCCTTCTTTTGTATATTTAATAAGTTT Majority
                2510      2520      2530      2540      2550
2401  CCTGATAGCCTTTTCTTTATCTTTCCTTCTTTTGTATATTTAATAAGTTT 2603_a12.seq
2501  CCTGATAGCCTTTTCTTTATCTTTCCTTCTTTTGTATATTTAATAAGTTT nem316_a12.seq

      TAATCGGCCTGTTTCAACTTTTCGCTTAGGATTTATCTGTAATTGATTTG Majority
                2560      2570      2580      2590      2600
2451  TAATCGGCCTGTTTCAACTTTTCGCTTAGGATTTATCTGTAATTGATTTG 2603_a12.seq
2551  TAATCGGCCTGTTTCAACTTTTCGCTTAGGATTTATCTGTAATTGATTTG nem316_a12.seq
```

FIGURE 20C

```
ATAACTTATCATCTGGTATTTCAATATAAAAAGGTACTATTGTTGAAACG Majority
          2610      2620      2630      2640      2650
2501 ATAACTTATCATCTGGTAATTCAATATAAAAAGGTACTATTGTTGAAACG 2603_ai2.seq
2601 ATAACTTATCATCTGGTATTTCAATATAAAAAGGTACTATTGTTGAAACG nem316_ai2.seq

CTTTGATCAGCTTTATAAGCACGACCAAAGTACGAACCATTTGGGAGTGC Majority
          2660      2670      2680      2690      2700
2551 CTTTGATCAGCTTTATAAGCACGACCAAAGTACGAACCATTTGGGAGTGC 2603_ai2.seq
2651 CTTTGATCAGCTTTATAAGCACGACCAAAGTACGAACCATTTGGGAGTGC nem316_ai2.seq

TATCTTTGTCTGACCATTAGTATCAGTAGGAGAAGTCAAGATACTCTTAT Majority
          2710      2720      2730      2740      2750
2601 TATCTTTGTCTGACCATTAGTATCAGTAGGAGAAGTCAAGATACTCTTAT 2603_ai2.seq
2701 TATCTTTGTCTGACCATTAGTATCAGTAGGAGAAGTCAAGATACTCTTAT nem316_ai2.seq

ACTTCTGGTTCAATTCGCTATCTGTCATTTGGCTCAATAAATCAACTTTT Majority
          2760      2770      2780      2790      2800
2651 ACTTCTGGTTCAATTCGCTATCTGTCATTTGGCTCAATAAATCAACTTTT 2603_ai2.seq
2751 ACTTCTGGTTCAATTCGCTATCTGTCATTTGGCTCAATAAATCAACTTTT nem316_ai2.seq

AAGTTGTCAGTCACAGTCCATAAACGATAAGAAATCCCCTCCTCTGTAGT Majority
          2810      2820      2830      2840      2850
2701 AAGTTGTCAGTCACAGTCCATAAACGATAAGAAATCCCCTCCTCTGTAGT 2603_ai2.seq
2801 AAGTTGTCAGTCACAGTCCATAAACGATAAGAAATCCCCTCCTCTGTAGT nem316_ai2.seq

ATTTGGCTGAAGTCCTATCTGTGTGATTGTTAGTTGATTAGGGGTATCAG Majority
          2860      2870      2880      2890      2900
2751 ATTTGGCTGAAGTCCTATCTGTGTGATTGTTAGTTGATTAGGGGTATCAG 2603_ai2.seq
2851 ATTTGGCTGAAGTCCTATCTGTGTGATTGTTAGTTGATTAGGGGTATCAG nem316_ai2.seq

CATTTACACTGGCTACCGAAAAAAACGCTAATTGTACCAATCCTAAAAAG Majority
          2910      2920      2930      2940      2950
2801 CATTTACACTGGCTACCGAAAAAAACGCTAATTGTACCAATCCTAAAAAG 2603_ai2.seq
2901 CATTTACACTGGCTACCGAAAAAAACGCTAATTGTACCAATCCTAAAAAG nem316_ai2.seq

CAACATAGTAGAAGTCCTAAACTTTTTTCTAATCTTTTTTCATTTTTGATTT Majority
          2960      2970      2980      2990      3000
2851 CAACATAGTAGAAGTCCTAAACTTTTTTCTAATCTTTTTTCATTTTTGATTT 2603_ai2.seq
2951 CAACATAGTAGAAGTCCTAAACTTTTTTCTAATCTTTTTTCATTTTTGATTT nem316_ai2.seq

CCCTTTCTTTTTCTCTCTTTAAATTTTCGTTTTAAATATAATAGTAAAGC Majority
          3010      3020      3030      3040      3050
2901 CCCTTTCTTTTTCTCTCTTTAAATTTTCGTTTTAAATATAATAGTAAAGC 2603_ai2.seq
3001 CCCTTTCTTTTTCTCTCTTTAAATTTTCGTTTTAAATATAATAGTAAAGC nem316_ai2.seq

GACTAATATAAGAATAACTAGGATTGATAAGAGGAAATAAAGTTTATAGT Majority
          3060      3070      3080      3090      3100
2951 GACTAATATAAGAATAACTAGGATTGATAAGAGGAAATAAAGTTTATAGT 2603_ai2.seq
3051 GACTAATATAAGAATAACTAGGATTGATAAGAGGAAATAAAGTTTATAGT nem316_ai2.seq

GTGTTTGCAATTCTTTCATTAAATAGTTCTTTTCTTTAACAGGAGGTACA Majority
          3110      3120      3130      3140      3150
3001 GTGTTTGCAATTCTTTCATTAAATAGTTCTTTTCTTTAACAGGAGGTACA 2603_ai2.seq
3101 GTGTTTGCAATTCTTTCATTAAATAGTTCTTTTCTTTAACAGGAGGTACA nem316_ai2.seq

TACTTGATTCGATGCCCTCTAACTAGTAAACGATGTGAATTAATCGAATA Majority
          3160      3170      3180      3190      3200
3051 TACTTGATTCGATGCCCTCTAACTAGTAAACGATGTGAATTAATCGAATA 2603_ai2.seq
3151 TACTTGATTCGATGCCCTCTAACTAGTAAACGATGTGAATTAATCGAATA nem316_ai2.seq

AGGTGTACATGTTAGCAAAGTCGCATAATCCTTACCTTTAACAACCAATA Majority
          3210      3220      3230      3240      3250
3101 AGGTGTACATGTTAGCAAAGTCGCATAATCCTTACCTTTAACAACCAATA 2603_ai2.seq
3201 AGGTGTACATGTTAGCAAAGTCGCATAATCCTTACCTTTAACAACCAATA nem316_ai2.seq
```

FIGURE 20D

```
                ATTTAGAAAAATTATCTGGCTTTACAACACTTATTTGATCAACCTTATAG Majority
                    3260      3270      3280      3290      3300
3151 ATTTAGAAAAATTATCTGGCTTTACAACACTTATTTGATCAACCTTATAG 2603_a12.seq
3251 ATTTAGAAAAATTATCTGGCTTTACAACACTTATTTGATCAACCTTATAG nem316_a12.seq

                GCTAAAACTTCTTTGATATTATGAATATAAAAAATTTTTCCTTTTTTTAAG Majority
                    3310      3320      3330      3340      3350
3201 GCTAAAACTTCTTTGATATTATGAATATAAAAAATTTTTCCTTTTTTTAAG 2603_a12.seq
3301 GCTAAAACTTCTTTGATATTATGAATATAAAAAATTTTTCCTTTTTTTAAG nem316_a12.seq

                TTTATCTAAATCTGTAAATAACTTAGCTTTAGGTAAGCCGCGATGAGCTG Majority
                    3360      3370      3380      3390      3400
3251 TTTATCTAAATCTGTAAATAACTTAGCTTTAGGTAAGCCGCGATGAGCTG 2603_a12.seq
3351 TTTATCTAAATCTGTAAATAACTTAGCTTTAGGTAAGCCGCGATGAGCTG nem316_a12.seq

                TGATAACAGTATGTGAACTTTTTCCACCAATTGGCAAGGAGGTTCCTTCA Majority
                    3410      3420      3430      3440      3450
3301 TGATAACAGTATGTGAACTTTTTCCACCAATTGGCAAGGAGGTTCCTTCA 2603_a12.seq
3401 TGATAACAGTATGTGAACTTTTTCCACCAATTGGCAAAGAGTTCCTTCA nem316_a12.seq

                AGGTGTCCTGCTCCTTTTTCAAGAACACTACTGGTAGTCCCCGCATAGAT Majority
                    3460      3470      3480      3490      3500
3351 AGGTGTCCTGCTCCTTTTTCAAGAACACTACTGGTAGTCCCCGCATAGAT 2603_a12.seq
3451 AGGTGTCCTGCTCCTTTTTCAAGAACACTACTGGTAGTCCCCGCATAGAT nem316_a12.seq

                AGGTAATTTTTGCTTGATAGACGGTATATCAATATATCCAATCATTTCAG Majority
                    3510      3520      3530      3540      3550
3401 AGGTAATTTTTGCTTGATAGACGGTATATCAATATATCCAATCATTTCAG 2603_a12.seq
3501 AGGTAATTTTTGCTTGATAGACGGTATATCAATATATCCAATCATTTCAG nem316_a12.seq

                TAATCTCAAGCATGTGGGCGTATTCAGCAATACCTTTTTTTTCTTTTTCA Majority
                    3560      3570      3580      3590      3600
3451 CAATCTCAAGCATGTGGGCGTATTCAGCAATACCTTTTTTTTCTTTTTCA 2603_a12.seq
3551 TAATCTCAAGCATATGGGCGTATTCAGCAATACCTTTTTTTTCTTTTTCA nem316_a12.seq

                GTATAGGGATCTGATAGGCGGCTTGGGTCCAGTGTTCTATTATAAGCTTT Majority
                    3610      3620      3630      3640      3650
3501 GTATAGGGATCTGATAGGCGGCTTGGGTCCAGTGTTCTATTATAAGCTTT 2603_a12.seq
3601 GTATAGGGATCTGATAGGCGGCTTGGGTCCAGTGTTCTATTATAAGCTTT nem316_a12.seq

                TGCTAACTCAAATCGTCTATTAATCTCTTTAGTATTTAATTTTTGGGTTT Majority
                    3660      3670      3680      3690      3700
3551 TGCTAACTCAAATCGTCTATTAATCTCTTTAGTATTTAATTTTTGGGTTT 2603_a12.seq
3651 TGCTAACTCAAATCGTCTATTAATCTCTTTAGCATTTAATTTTTGGGTTT nem316_a12.seq

                GATTATCAAAGTTAGTTACTTGATTATTAGCTTTAATATTATAGTACCAA Majority
                    3710      3720      3730      3740      3750
3601 GATTATCAAAGTTAGTTACTTGATTATTAGCTTTAATATTATAGTACCAA 2603_a12.seq
3701 GATTATCAAAGTTAGTTACTTGATTATTAGCTTTAATATTATAGTACCAA nem316_a12.seq

                TTTGAAATAAAAGGATATGAGGTTATCAAAAGACCAACTAAGAACAATAG Majority
                    3760      3770      3780      3790      3800
3651 TTTGAAATAAAAGGATATGAGGTTATCAAAAGACCAACTAAGAACAATAG 2603_a12.seq
3751 TTTGAAATAAAAGGATATGAGGTTATCAAAAGACCAACTAAGAACAATAG nem316_a12.seq

                TATCAGGCCTACATTCATCCATCGATTTAAAACGACCGATTTCTTAAGGT Majority
                    3810      3820      3830      3840      3850
3701 TATCAAGCCTACATTCATCCATCGATTTAAAACGACCGATTTCTTAAGGT 2603_a12.seq
3801 TATCAGGCCTACATTCATCCATCGATTTAAAACGACCGATTTCTTAAGGT nem316_a12.seq

                TTTTCTGAAATTTTCCTCCCATTATGATTCAATTCCTTTTCTAACACTTG Majority
                    3860      3870      3880      3890      3900
3751 TTTTCTGAAATTTTCCTCCCATTATGATTCAATTCCTTTTCTAACACTTG 2603_a12.seq
3851 TTTTCTGAAATTTTCCTCCCATTATGATTCAATTCCTTTTCTAACACTTG nem316_a12.seq
```

FIGURE 20E

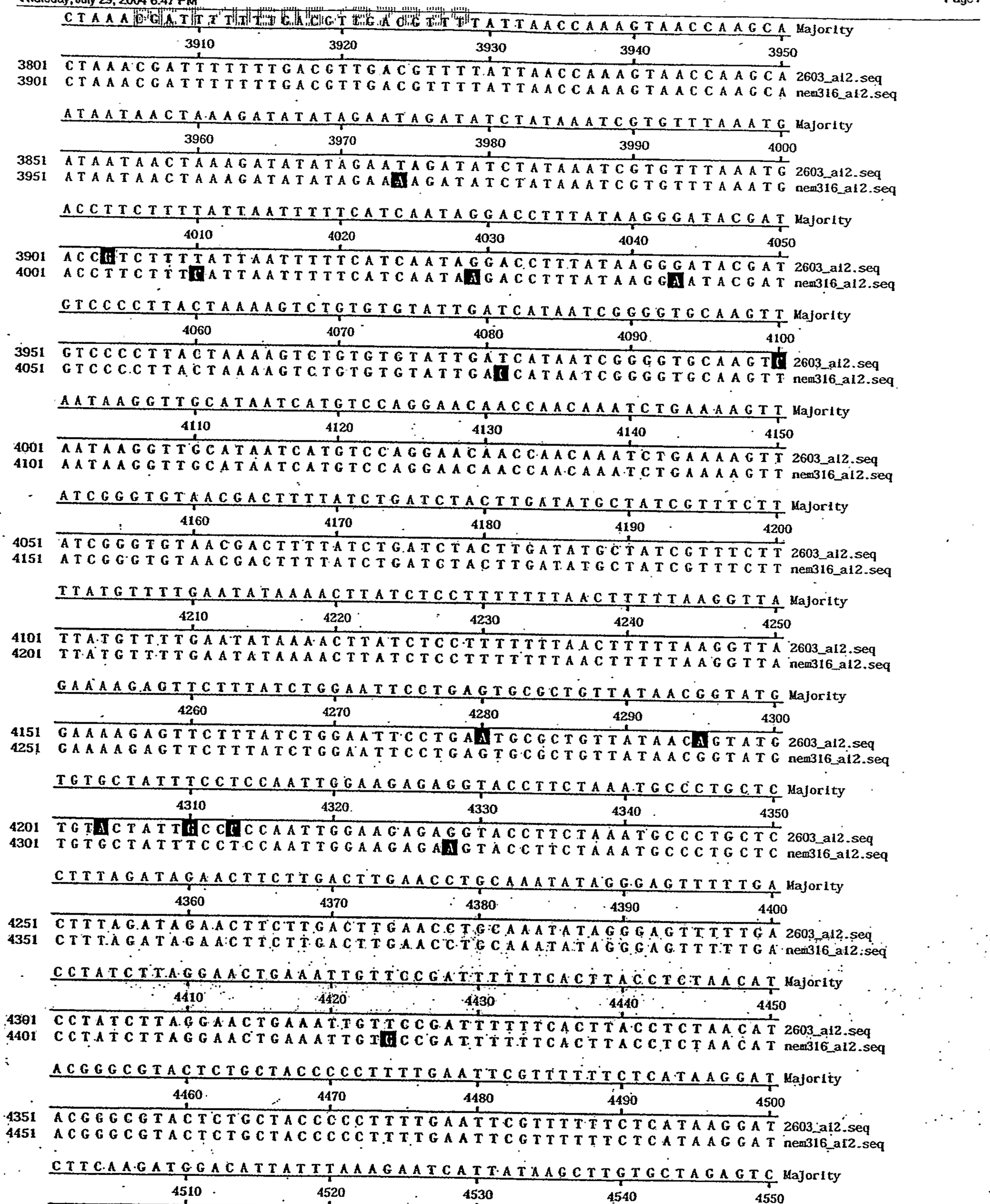

CTAAACGATTTTTTTGACGTTGACGTTTTTATTAACCAAAGTAACCAAGCA Majority
3910 3920 3930 3940 3950
3801 CTAAACGATTTTTTTGACGTTGACGTTTTTATTAACCAAAGTAACCAAGCA 2603_a12.seq
3901 CTAAACGATTTTTTTGACGTTGACGTTTTTATTAACCAAAGTAACCAAGCA nem316_a12.seq
ATAATAACTAAAGATATATAGAATAGATATCTATAAATCGTGTTTAAATG Majority
3960 3970 3980 3990 4000
3851 ATAATAACTAAAGATATATAGAATAGATATCTATAAATCGTGTTTAAATG 2603_a12.seq
3951 ATAATAACTAAAGATATATAGAAAAGATATCTATAAATCGTGTTTAAATG nem316_a12.seq
ACCTTCTTTTATTAATTTTTCATCAATAGGACCTTTATAAGGGATACGAT Majority
4010 4020 4030 4040 4050
3901 ACCGTCTTTTATTAATTTTTCATCAATAGGACCTTTATAAGGGATACGAT 2603_a12.seq
4001 ACCTTCTTTCATTAATTTTTCATCAATAAGACCTTTATAAGGAATACGAT nem316_a12.seq
GTCCCCTTACTAAAAGTCTGTGTGTATTGATCATAATCGGGGTGCAAGTT Majority
4060 4070 4080 4090 4100
3951 GTCCCCTTACTAAAAGTCTGTGTGTATTGATCATAATCGGGGTGCAAGTC 2603_a12.seq
4051 GTCCCCTTACTAAAAGTCTGTGTGTATTGARCATAATCGGGGTGCAAGTT nem316_a12.seq
AATAAGGTTGCATAATCATGTCCAGGAACAACCAACAAATCTGAAAAGTT Majority
4110 4120 4130 4140 4150
4001 AATAAGGTTGCATAATCATGTCCAGGAACAACCAACAAATCTGAAAAGTT 2603_a12.seq
4101 AATAAGGTTGCATAATCATGTCCAGGAACAACCAACAAATCTGAAAAGTT nem316_a12.seq
ATCGGGTGTAACGACTTTTATCTGATCTACTTGATATGCTATCGTTTCTT Majority
4160 4170 4180 4190 4200
4051 ATCGGGTGTAACGACTTTTATCTGATCTACTTGATATGCTATCGTTTCTT 2603_a12.seq
4151 ATCGGGTGTAACGACTTTTATCTGATCTACTTGATATGCTATCGTTTCTT nem316_a12.seq
TTATGTTTTGAATATAAAACTTATCTCCTTTTTTTAACTTTTTAAGGTTA Majority
4210 4220 4230 4240 4250
4101 TTATGTTTTGAATATAAAACTTATCTCCTTTTTTTAACTTTTTAAGGTTA 2603_a12.seq
4201 TTATGTTTTGAATATAAAACTTATCTCCTTTTTTTAACTTTTTAAGGTTA nem316_a12.seq
GAAAAGAGTTCTTTATCTGGAATTCCTGAGTGCGCTGTTATAACGGTATG Majority
4260 4270 4280 4290 4300
4151 GAAAAGAGTTCTTTATCTGGAATTCCTGAATGCGCTGTTATAACAGTATG 2603_a12.seq
4251 GAAAAGAGTTCTTTATCTGGAATTCCTGAGTGCGCTGTTATAACGGTATG nem316_a12.seq
TGTGCTATTTCCTCCAATTGGAAGAGAGGTACCTTCTAAATGCCCTGCTC Majority
4310 4320 4330 4340 4350
4201 TGTACTATTGCCCCAATTGGAAGAGAGGTACCTTCTAAATGCCCTGCTC 2603_a12.seq
4301 TGTGCTATTTCCTCCAATTGGAAGAGAAGTACCTTCTAAATGCCCTGCTC nem316_a12.seq
CTTTAGATAGAACTTCTTGACTTGAACCTGCAAATATAGGGAGTTTTTGA Majority
4360 4370 4380 4390 4400
4251 CTTTAGATAGAACTTCTTGACTTGAACCTGCAAATATAGGGAGTTTTTGA 2603_a12.seq
4351 CTTTAGATAGAACTTCTTGACTTGAACCTGCAAATATAGGGAGTTTTTGA nem316_a12.seq
CCTATCTTAGGAACTGAAATTGTTCCGATTTTTTCACTTACCTCTAACAT Majority
4410 4420 4430 4440 4450
4301 CCTATCTTAGGAACTGAAATTGTTCCGATTTTTTCACTTACCTCTAACAT 2603_a12.seq
4401 CCTATCTTAGGAACTGAAATTGTGCCGATTTTTTCACTTACCTCTAACAT nem316_a12.seq
ACGGGCGTACTCTGCTACCCCCTTTTGAATTCGTTTTTTCTCATAAGGAT Majority
4460 4470 4480 4490 4500
4351 ACGGGCGTACTCTGCTACCCCCTTTTGAATTCGTTTTTTCTCATAAGGAT 2603_a12.seq
4451 ACGGGCGTACTCTGCTACCCCCTTTTGAATTCGTTTTTTCTCATAAGGAT nem316_a12.seq
CTTCAAGATGGACATTATTTAAAGAATCATTATAAGCTTGTGCTAGAGTC Majority
4510 4520 4530 4540 4550
4401 CTTCAAGATGGACATTATTTAAAGAATCATTATAAGCTTGTGCTAGAGCC 2603_a12.seq
4501 CTTCAAGATGGACATTATTTAAAGAATCATTATAAGCTTGTGCTAGAGTC nem316_a12.seq

FIGURE 20F

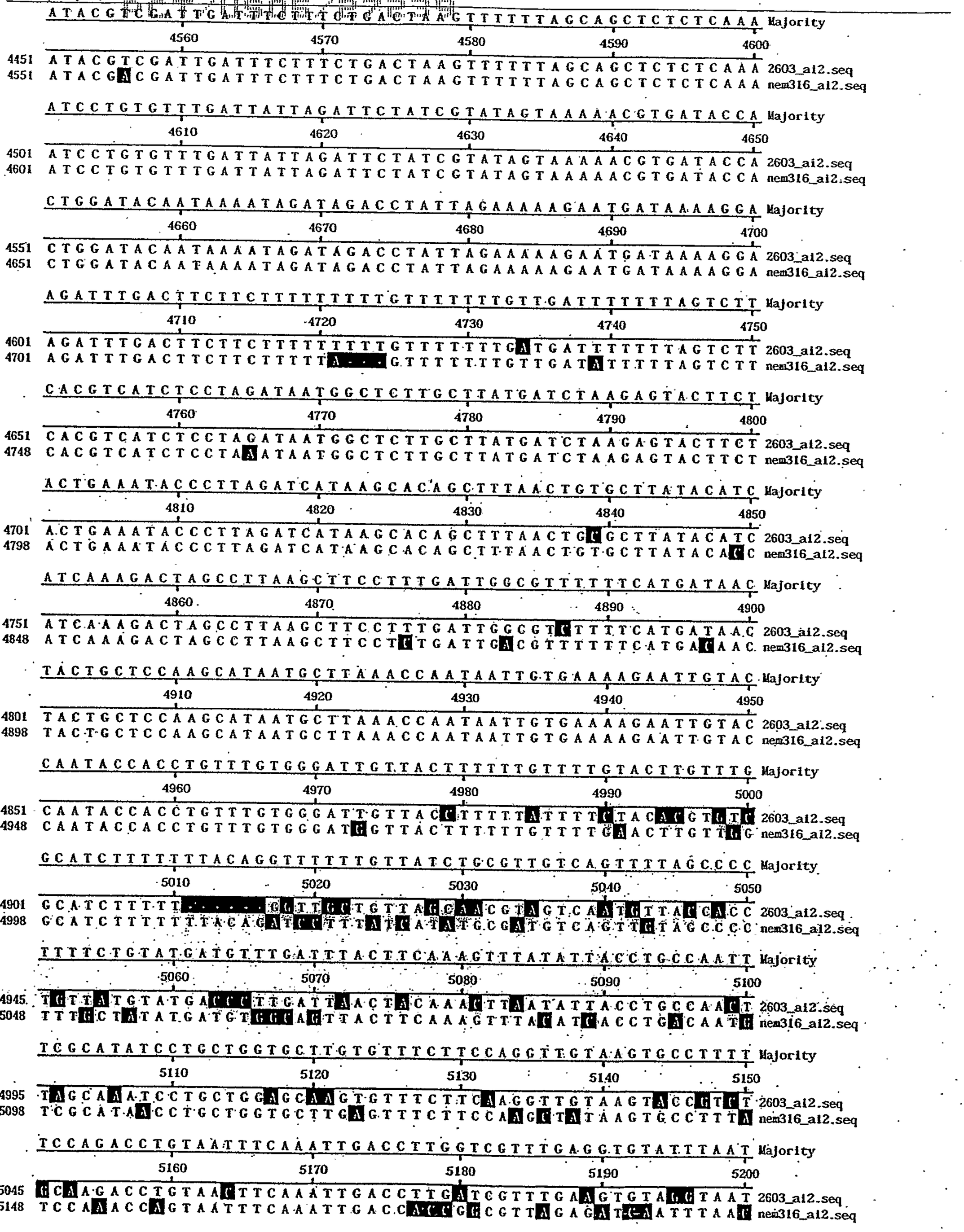

ATACGTCGATTGATTTCTTTCTGACTAAGTTTTTTAGCAGCTCTCTCAAA Majority
4560 4570 4580 4590 4600
4451 ATACGTCGATTGATTTCTTTCTGACTAAGTTTTTTAGCAGCTCTCTCAAA 2603_a12.seq
4551 ATACGACGATTGATTTCTTTCTGACTAAGTTTTTTAGCAGCTCTCTCAAA nem316_a12.seq
ATCCTGTGTTTGATTATTAGATTCTATCGTATAGTAAAAACGTGATACCA Majority
4610 4620 4630 4640 4650
4501 ATCCTGTGTTTGATTATTAGATTCTATCGTATAGTAAAAACGTGATACCA 2603_a12.seq
4601 ATCCTGTGTTTGATTATTAGATTCTATCGTATAGTAAAAACGTGATACCA nem316_a12.seq
CTGGATACAATAAAATAGATAGACCTATTAGAAAAAGAATGATAAAAGGA Majority
4660 4670 4680 4690 4700
4551 CTGGATACAATAAAATAGATAGACCTATTAGAAAAAGAATGATAAAAGGA 2603_a12.seq
4651 CTGGATACAATAAAATAGATAGACCTATTAGAAAAAGAATGATAAAAGGA nem316_a12.seq
AGATTTGACTTCTTCTTTTTTTTTGTTTTTTTGTTGATTTTTTTAGTCTT Majority
4710 4720 4730 4740 4750
4601 AGATTTGACTTCTTCTTTTTTTTTTGTTTTTTTGATGATTTTTTTAGTCTT 2603_a12.seq
4701 AGATTTGACTTCTTCTTTTTA---GTTTTTTTGTTGATTTTTTTAGTCTT nem316_a12.seq
CACGTCATCTCCTAGATAATGGCTCTTGCTTATGATCTAAGAGTACTTCT Majority
4760 4770 4780 4790 4800
4651 CACGTCATCTCCTAGATAATGGCTCTTGCTTATGATCTAAGAGTACTTCT 2603_a12.seq
4748 CACGTCATCTCCTAAATAATGGCTCTTGCTTATGATCTAAGAGTACTTCT nem316_a12.seq
ACTGAAATACCCTTAGATCATAAGCACAGCTTTAACTGTGCTTATACATC Majority
4810 4820 4830 4840 4850
4701 ACTGAAATACCCTTAGATCATAAGCACAGCTTTAACTGCGCTTATACATC 2603_a12.seq
4798 ACTGAAATACCCTTAGATCATAAGCACAGCTTTAACTGTGCTTATACAGC nem316_a12.seq
ATCAAAGACTAGCCTTAAGCTTCCTTTGATTGGCGTTTTTTCATGATAAC Majority
4860 4870 4880 4890 4900
4751 ATCAAAGACTAGCCTTAAGCTTCCTTTGATTGGCGTCTTTTCATGATAAC 2603_a12.seq
4848 ATCAAAGACTAGCCTTAAGCTTCCTCTGATTGACGTTTTTTCATGACAAC nem316_a12.seq
TACTGCTCCAAGCATAATGCTTAAACCAATAATTGTGAAAAGAATTGTAC Majority
4910 4920 4930 4940 4950
4801 TACTGCTCCAAGCATAATGCTTAAACCAATAATTGTGAAAAGAATTGTAC 2603_a12.seq
4898 TACTGCTCCAAGCATAATGCTTAAACCAATAATTGTGAAAAGAATTGTAC nem316_a12.seq
CAATACCACCTGTTTGTGGGATTGTTACTTTTTTGTTTTGTACTTGTTTG Majority
4960 4970 4980 4990 5000
4851 CAATACCACCTGTTTGTGGGATTGTTACCTTTTTTATTTTGTACAGTGTC 2603_a12.seq
4948 CAATACCACCTGTTTGTGGGATGGTTACTTTTTTGTTTTGAACTTGTTTG nem316_a12.seq
GCATCTTTTTTTACAGGTTTTTTGTTATCTGCGTTGTCAGTTTTAGCCCC Majority
5010 5020 5030 5040 5050
4901 GCATCTTTTT-----GGTTGCTGTTAGCAACGTAGTCAATGTTACCACC 2603_a12.seq
4998 GCATCTTTTTTTACAGATCCTTTATCATATGCGATGTCAGTTGTAGCCCC nem316_a12.seq
TTTTCTGTATGATGTTTGATTTACTTCAAAGTTTATATTACCTGCCAATT Majority
5060 5070 5080 5090 5100
4945 TGTTATGTATGACCCTTGATTAACTACAAACTTAATATTACCTGCCAACT 2603_a12.seq
5048 TTTGCTATATGATGTGGCAGTTACTTCAAAGTTTACATCACCTGACAATG nem316_a12.seq
TCGCATATCCTGCTGGTGCTTGTGTTTCTTCCAGGTTGTAAGTGCCTTTT Majority
5110 5120 5130 5140 5150
4995 TAGCAAATCCTGCTGGAGCAAGTGTTTCTTCAAGGTTGTAAGTACCGTCT 2603_a12.seq
5098 TCGCATAACCTGCTGGTGCTTGAGTTTCTTCCAAGCTATAAGTGCCTTTA nem316_a12.seq
TCCAGACCTGTAATTTCAAATTGACCTTGGTCGTTTGAGGTGTATTTAAT Majority
5160 5170 5180 5190 5200
5045 GCAAGACCTGTAAGTTCAAATTGACCTTGATCGTTTGAAGTGTAGGTAAT 2603_a12.seq
5148 TCCAAACCAGTAATTTCAAATTGACCACGGCGTTAGAGATCAATTTAAT nem316_a12.seq

FIGURE 20G

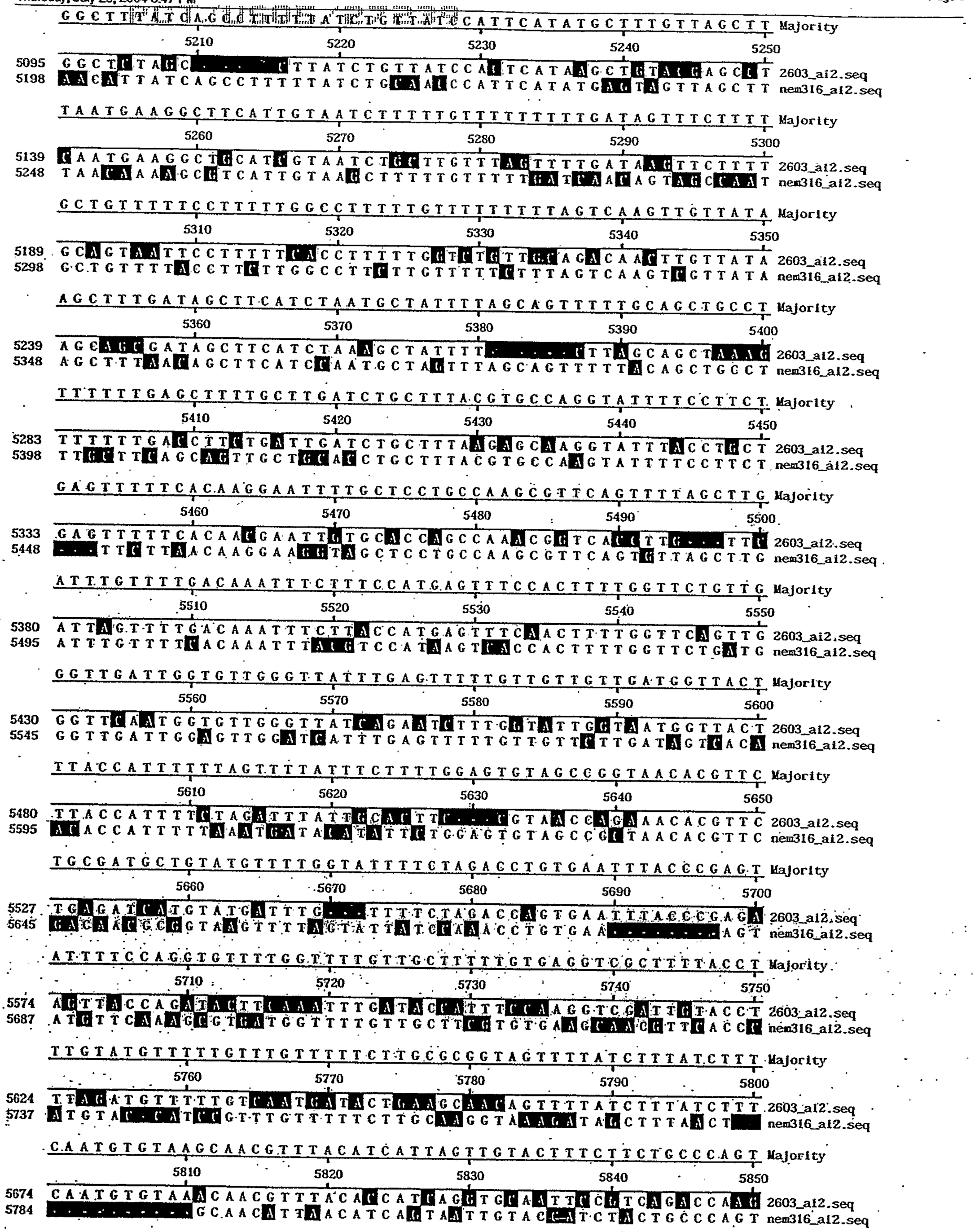

GGCTTTATCAGCCTTTTATCTGTTATCCATTCATATGCTTTGTTAGCTT Majority
5210 5220 5230 5240 5250
5095 GGCTCTAGC...GTTATCTGTTATCCATTCATAAGCTGTACGAGCCT 2603_a12.seq
5198 AACATTATCAGCCTTTTTTATCTGCAATCCATTCATATGAGTAGTTAGCTT nem316_a12.seq
TAATGAAGGCTTCATTGTAATCTTTTTGTTTTTTTTTTGATAGTTTCTTTT Majority
5260 5270 5280 5290 5300
5139 CAATGAAGGCTGCATCGTAATCTGCTTGTTTAGTTTTGATAAGTTCTTTT 2603_a12.seq
5248 TAACAAAAGCGTCATTGTAAGCTTTTTGTTTTTGATCAACAGTAGCCAAT nem316_a12.seq
GCTGTTTTTCCTTTTTGGCCTTTTTGTTTTTTTTTAGTCAAGTTGTTATA Majority
5310 5320 5330 5340 5350
5189 GCAGTAATTCCTTTTTCACCTTTTTGGTCTGTTGCAGACAAGTTGTTATA 2603_a12.seq
5298 GCTGTTTTACCTTCTTGGCCTTCTTGTTTTTCTTTAGTCAAGTCGTTATA nem316_a12.seq
AGCTTTGATAGCTTCATCTAATGCTATTTTAGCAGTTTTTGCAGCTGCCT Majority
5360 5370 5380 5390 5400
5239 AGCAGCGATAGCTTCATCTAAAGCTATTTT.....CTTAGCAGCTAAAG 2603_a12.seq
5348 AGCTTTAACAGCTTCATCTAATGCTAGTTTAGCAGTTTTTACAGCTGCCT nem316_a12.seq
TTTTTTGAGCTTTTGCTTGATCTGCTTTACGTGCCAGGTATTTTCCTTCT Majority
5410 5420 5430 5440 5450
5283 TTTTTTGACCTTCTGATTGATCTGCTTTAAGAGCAAGGTATTTACCTGCT 2603_a12.seq
5398 TTGCTTCAGCAGTTGCTGCACCTGCTTTACGTGCCAAGTATTTTCCTTCT nem316_a12.seq
GAGTTTTTCACAAGGAATTTTGCTCCTGCCAAGCGTTCAGTTTTAGCTTG Majority
5460 5470 5480 5490 5500
5333 GAGTTTTTCACAACGAATTGTGCACCAGCCAAACGGTCACTTG....TTTT 2603_a12.seq
5448 ....TTGTTTAACAAGGAAGGTAGCTCCTGCCAAGCGTTCAGTGTTAGCTTG nem316_a12.seq
ATTTGTTTTGACAAATTTCTTTCCATGAGTTTCCACTTTTGGTTCTGTTG Majority
5510 5520 5530 5540 5550
5380 ATTAGTTTTGACAAATTTCTTACCATGAGTTTCAACTTTTGGTTCAGTTG 2603_a12.seq
5495 ATTTGTTTTTCACAAATTTACGTCCATAAGTCACCACTTTTGGTTCTGATG nem316_a12.seq
GGTTGATTGGTGTTGGGTTATTTGAGTTTTTGTTGTTGTTGATGGTTACT Majority
5560 5570 5580 5590 5600
5430 GGTTGCAATGGTGTTGGGTTATCAGAATCTTTGGTATTGGTAATGGTTACT 2603_a12.seq
5545 GGTTGATTGGAGTTGGATCATTTGAGTTTTTGTTGTTCTTGATAGTCACA nem316_a12.seq
TTACCATTTTTTAGTTTTATTTCTTTTGGAGTGTAGCCGGTAACACGTTC Majority
5610 5620 5630 5640 5650
5480 TTACCATTTTCTAGATTTATTGCACTTC....TGTAACCAGAACACGTTC 2603_a12.seq
5595 ACCATTTTTAAATGATACATTTCTGGAGTGTAGCCGCTAACACGTTC nem316_a12.seq
TGCGATGCTGTATGTTTTGGTATTTTCTAGACCTGTGAATTTACCCGAGT Majority
5660 5670 5680 5690 5700
5527 TGAGATCATGTATGATTTG....TTTTCTAGACCAGTGAATTTACCCGAGA 2603_a12.seq
5645 GACAACGCGGTAAGTTTTAGTATTATCGAAACCTGTGAA.........AGT nem316_a12.seq
ATTTCCAGGTGTTTTGGTTTTGTTGCTTTTTGTGAGGTCGCTTTTACCT Majority
5710 5720 5730 5740 5750
5574 AGTTACCAGATACTTCAAATTTGATACCATTTCCAAGGTCGATTGTACCT 2603_a12.seq
5687 ATGTTCAAAGCGTGATGGTTTTGTTGCTTCGTGTGAAGCAACGTTCACCT nem316_a12.seq
TTGTATGTTTTTGTTTGTTTTTCTTGCGCGGTAGTTTTATCTTTATCTTT Majority
5760 5770 5780 5790 5800
5624 TTAGATGTTTTTGTCAATGATACTGAAGCAAGCAGTTTTATCTTTATCTTT 2603_a12.seq
5737 ATGTAC.CATCCGTTGTTTTTCTTGCAAGGTAAAGATAGCTTTAACT... nem316_a12.seq
CAATGTGTAAGCAACGTTTACATCATTAGTTGTACTTTCTTCTGCCCAGT Majority
5810 5820 5830 5840 5850
5674 CAATGTGTAAACAACGTTTACAGCATCAGGTGTAATTCCGTCAGACCAAG 2603_a12.seq
5784 .....GCAACATTACATCATTACTTGTACCATCTACTGCCCAGT nem316_a12.seq

FIGURE 20H

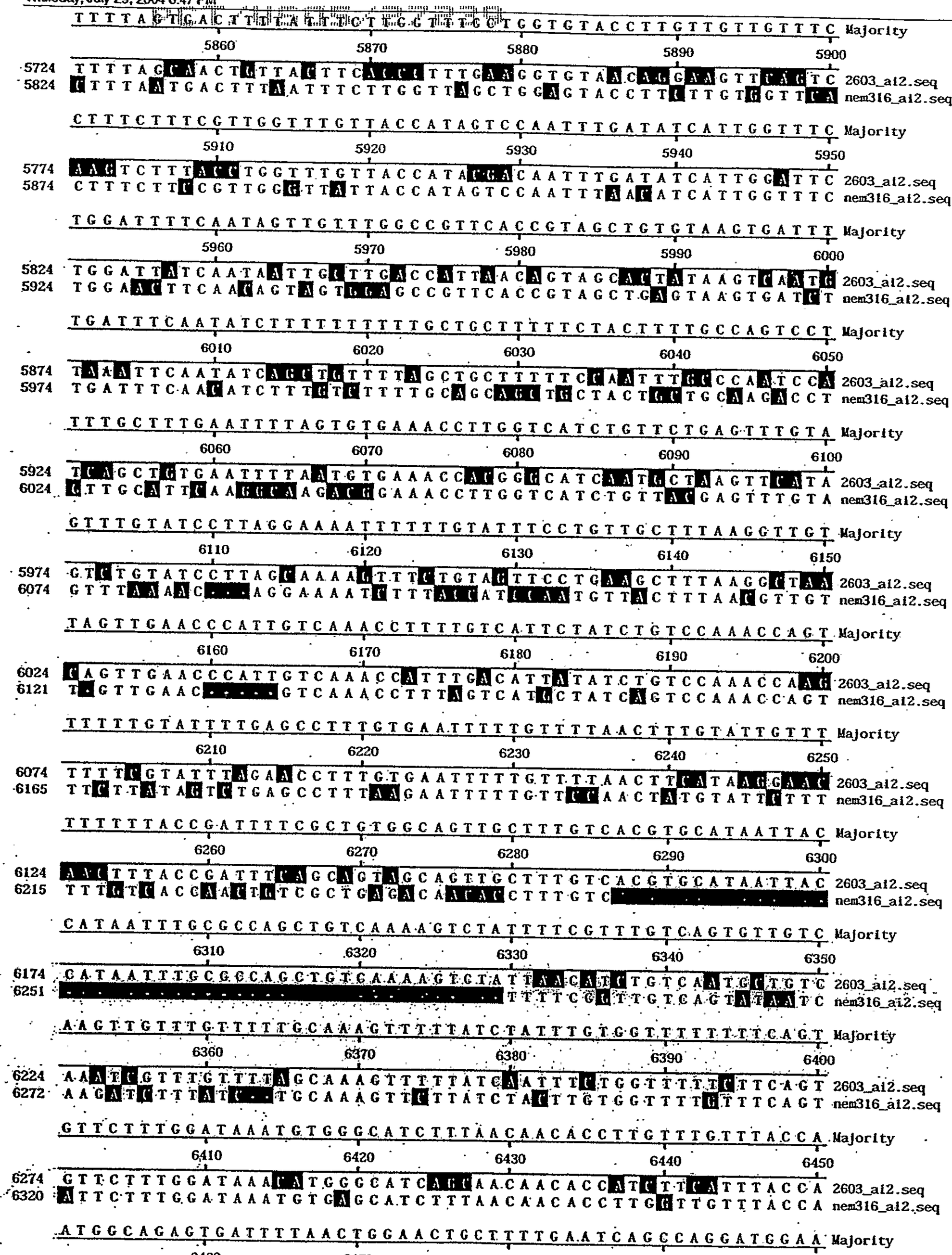

TTTTAGTGACTTTATTCTCGTTCGTGGTGTACCTTGTTGTTGTTTC Majority
5860 5870 5880 5890 5900
5724 TTTTAGCAACTGTTAGTTCACCCTTTGAAGGTGTAACAGGAAGTTCAGTC 2603_a12.seq
5824 CTTTAATGACTTTAATTTCTTGGTTAGCTGGAGTACCTTCTTGTGGTTCA nem316_a12.seq
CTTTCTTTCGTTGGTTTGTTACCATAGTCCAATTTGATATCATTGGTTTC Majority
5910 5920 5930 5940 5950
5774 AAGTCTTTACCTGGTTTGTTACCATACTGACAATTTGATATCATTGGATTC 2603_a12.seq
5874 CTTTCTTCGTTGGGTTATTACCATAGTCCAATTTAACATCATTGGTTTC nem316_a12.seq
TGGATTTTCAATAGTTGTTTGGCCGTTCACCGTAGCTGTGTAAGTGATTT Majority
5960 5970 5980 5990 6000
5824 TGGATTATCAATAATTGCTTGACCATTAACAGTAGCACTATAAGTCAATG 2603_a12.seq
5924 TGGAACTTCAACAGTAGTGGAGCCGTTCACCGTAGCTGAGTAAGTGATCT nem316_a12.seq
TGATTTCAATATCTTTTTTTTTTGCTGCTTTTTCTACTTTTGCCAGTCCT Majority
6010 6020 6030 6040 6050
5874 TAAATTCAATATCAGCTGTTTTAGCTGCTTTTTCCAATTTGCCCAATCCA 2603_a12.seq
5974 TGATTTCAACATCTTTGTCTTTTGCAGCAGCTGCTACTGCTGCAAGACCT nem316_a12.seq
TTTGCTTTGAATTTTAGTGTGAAACCTTGGTCATCTGTTCTGAGTTTGTA Majority
6060 6070 6080 6090 6100
5924 TCAGCTGTGAATTTTAATGTGAAACCACGGGCATCAATGCTAAGTTCATA 2603_a12.seq
6024 GTTGCATTCAAGGCAAGACGGAAACCTTGGTCATCTGTTACGAGTTTGTA nem316_a12.seq
GTTTGTATCCTTAGGAAAATTTTTTGTATTTCCTGTTGCTTTAAGGTTGT Majority
6110 6120 6130 6140 6150
5974 GTCTGTATCCTTAGCAAAAGTTTCTGTAGTTCCTGAAGCTTTAAGGCTAA 2603_a12.seq
6074 GTTTAAAAC.....AGGAAAATTTTTACATCAATGTTACTTTAACGTTGT nem316_a12.seq
TAGTTGAACCCATTGTCAAACCTTTTGTCATTCTATCTGTCCAAACCAGT Majority
6160 6170 6180 6190 6200
6024 CAGTTGAACCCATTGTCAAACCATTTGACATTATATCTGTCCAAACCAAG 2603_a12.seq
6121 T.GTTGAAC.....GTCAAACCTTTAGTCATGCTATCAGTCCAAACCAGT nem316_a12.seq
TTTTTGTATTTTGAGCCTTTGTGAATTTTTGTTTTAACTTTGTATTGTTT Majority
6210 6220 6230 6240 6250
6074 TTTTTGTATTTAGAACCTTTGTGAATTTTTGTTTTAACTTCATAAGGAAC 2603_a12.seq
6165 TTCTTATAGTCTGAGCCTTTTAAGAATTTTTGTTCCAACTATGTATTCTTT nem316_a12.seq
TTTTTTACCGATTTTCGCTGTGGCAGTTGCTTTGTCACGTGCATAATTAC Majority
6260 6270 6280 6290 6300
6124 AACTTTACCGATTTCAGCAGTAGCAGTTGCTTTGTCACGTGCATAATTAC 2603_a12.seq
6215 TTTGTCACGAACTGTCGCTGAGACAAACTTTGTC..... nem316_a12.seq
CATAATTTGCGCCAGCTGTCAAAAGTCTATTTTCGTTTGTCAGTGTTGTC Majority
6310 6320 6330 6340 6350
6174 CATAATTTGCGCCAGCTGTCAAAAGTCTATTAACATCTGTCAATGCTGTC 2603_a12.seq
6251 .....TTTTCGGTTGTCAGTAATAATC nem316_a12.seq
AAGTTGTTTGTTTTTGCAAAGTTTTTATCTATTTGTGGTTTTTTTTCAGT Majority
6360 6370 6380 6390 6400
6224 AAATCGTTTGTTTTAGCAAAGTTTTTATCAATTTCTGGTTTTTCTTCAGT 2603_a12.seq
6272 AAGATCTTTATC..TGCAAAGTTCTTATCTACTTGTGGTTTTGTTTCAGT nem316_a12.seq
GTTCTTTGGATAAATGTGGGCATCTTTAACAACACCTTGTTTGTTTACCA Majority
6410 6420 6430 6440 6450
6274 GTTCTTTGGATAAACATGGGCATCAGCAACAACACCCATCTTCATTTACCA 2603_a12.seq
6320 ATTCTTTGGATAAATGTGAGCATCTTTAACAACACCTTGGTTGTTTACCA nem316_a12.seq
ATGGCAGAGTGATTTTAACTGGAACTGCTTTTGAATCAGCCAGGATGGAA Majority
6460 6470 6480 6490 6500
6324 ATGGAAGAGTGATGTTAACTGGAACTGCTTTTGAAGCAGCCAGGAGGAA 2603_a12.seq
6370 ATGGCAGAGTGATTTTAACTGGAACTGCTTTTGAATCAGCCAAGATAGAA nem316_a12.seq

FIGURE 20I

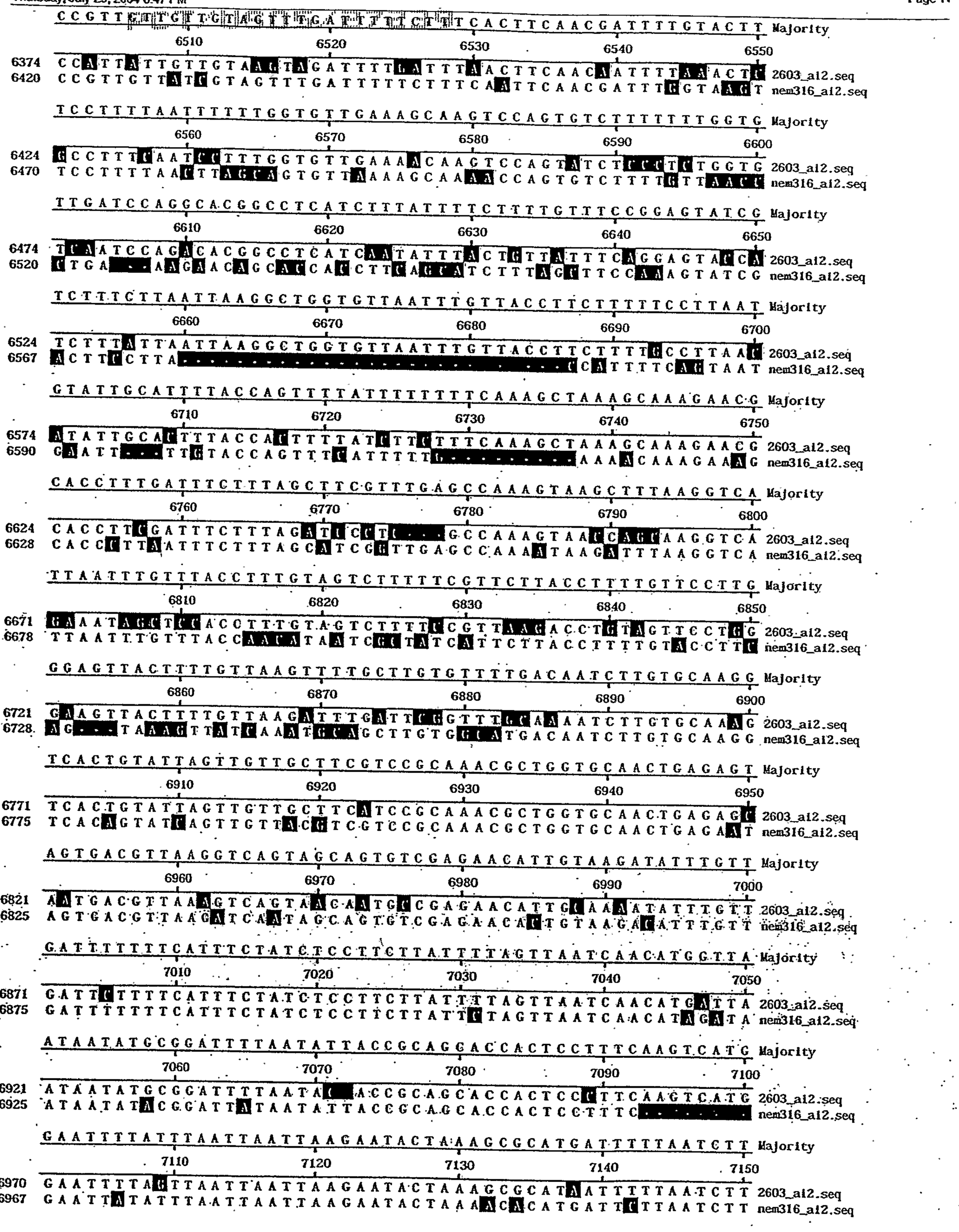

CCGTTGTTGTTAGTTTGATTTTCTTTCACTTCAACGATTTTGTACTT Majority
6510 6520 6530 6540 6550
6374 CCATTATTGTTGTAAGTAGATTTTGATTTAACTTCAACATTTTAAACTC 2603_a12.seq
6420 CCGTTGTTATGGTAGTTTGATTTTTCTTTCAATTCAACGATTTGGTAAGT nem316_a12.seq
TCCTTTTAATTTTTTGGTGTTGAAAGCAAGTCCAGTGTCTTTTTTTGGTG Majority
6560 6570 6580 6590 6600
6424 GCCTTTGAATCCTTTGGTGTTGAAAACAAGTCCAGTATCTCCTCTGGTG 2603_a12.seq
6470 TCCTTTTAACTTAGCAGTGTTAAAAGCAAACCAGTGTCTTTTGTTAACT nem316_a12.seq
TTGATCCAGGCACGGCCTCATCTTTATTTTCTTTTGTTTCCGGAGTATCG Majority
6610 6620 6630 6640 6650
6474 TCAATCCAGACACGGCCTCATCAATATTTACTGTTATTTCAGGAGTACCA 2603_a12.seq
6520 CTGA---AAGAACAGCACCACTTCAGCATCTTTAGCTTCCAAAGTATCG nem316_a12.seq
TCTTTCTTAATTAAGGCTGGTGTTAATTTGTTACCTTCTTTTTCCTTAAT Majority
6660 6670 6680 6690 6700
6524 TCTTTATTAATTAAGGCTGGTGTTAATTTGTTACCTTCTTTTGCCTTAAC 2603_a12.seq
6567 ACTTCCTTA----------CATTTTCAGTAAT nem316_a12.seq
GTATTGCATTTTACCAGTTTTATTTTTTTTTCAAAGCTAAAGCAAAGAACG Majority
6710 6720 6730 6740 6750
6574 ATATTGCACTTTACCAGTTTTATGTTCTTTCAAAGCTAAAGCAAAGAACG 2603_a12.seq
6590 GAATT---TTGTACCAGTTTCATTTTTT----------AAAACAAAGAAAG nem316_a12.seq
CACCTTTGATTTCTTTAGCTTCGTTTGAGCCAAAGTAAGCTTTAAGGTCA Majority
6760 6770 6780 6790 6800
6624 CACCTTCGATTTCTTTAGATCCTC---GCCAAAGTAACCAGCAAGGTCA 2603_a12.seq
6628 CACCGTTAATTTCTTTAGCATCGGTTGAGCCAAAATAAGATTTAAGGTCA nem316_a12.seq
TTAATTTGTTTACCTTTGTAGTCTTTTTCGTTCTTACCTTTTGTTCCTTG Majority
6810 6820 6830 6840 6850
6671 GAAATAGCTCCACCTTTGTAGTCTTTTTCGTTAAGACCTGTAGTTCCTGG 2603_a12.seq
6678 TTAATTTGTTTACCAACATAATCGTATCATTCTTACCTTTTGTACCTTC nem316_a12.seq
GGAGTTACTTTTGTTAAGTTTTGCTTGTGTTTTGACAATCTTGTGCAAGG Majority
6860 6870 6880 6890 6900
6721 GAAGTTACTTTTGTTAAGATTTGATTCGGTTTTGCAAAATCTTGTGCAAAG 2603_a12.seq
6728 AG---TAAAGTTATCAATGCAGCTTGTGGCATGACAATCTTGTGCAAGG nem316_a12.seq
TCACTGTATTAGTTGTTGCTTCGTCCGCAAACGCTGGTGCAACTGAGAGT Majority
6910 6920 6930 6940 6950
6771 TCACTGTATTAGTTGTTGCTTCATCCGCAAACGCTGGTGCAACTGAGAGC 2603_a12.seq
6775 TCACGTATCAGTTGTTACGTCGTCCGCAAACGCTGGTGCAACTGAGAAT nem316_a12.seq
AGTGACGTTAAGGTCAGTAGCAGTGTCGAGAACATTGTAAGATATTTGTT Majority
6960 6970 6980 6990 7000
6821 AATGACGTTAAAGTCAGTAACAATGTCGAGAACATTGCAAATATTTGTT 2603_a12.seq
6825 AGTGACGTTAAGATCAATAGCAGTGTCGAGAACATTGTAAGACATTTGTT nem316_a12.seq
GATTTTTTTCATTTCTATCTCCTTCTTATTTTAGTTAATCAACATGGTTA Majority
7010 7020 7030 7040 7050
6871 GATTCTTTTCATTTCTATCTCCTTCTTATTTTAGTTAATCAACATGATTA 2603_a12.seq
6875 GATTTTTTTCATTTCTATCTCCTTCTTATTCTAGTTAATCAACATAGATA nem316_a12.seq
ATAATATGCGGATTTTAATATTACCGCAGGACCACTCCTTTCAAGTCATG Majority
7060 7070 7080 7090 7100
6921 ATAATATGCGGATTTTAATA-ACCGCAGCACCACTCCCTTCAAGTCATG 2603_a12.seq
6925 ATAATATACGGATTATAATATTACCGCAGCACCACTCCTTTC---------- nem316_a12.seq
GAATTTTATTTAATTAATTAAGAATACTAAAGCGCATGATTTTTAATCTT Majority
7110 7120 7130 7140 7150
6970 GAATTTTAGTTAATTAATTAAGAATACTAAAGCGCATAATTTTTAATCTT 2603_a12.seq
6967 GAATTATTATTTAATTAATTAAGAATACTAAAACACATGATTCTTAATCTT nem316_a12.seq

FIGURE 20J

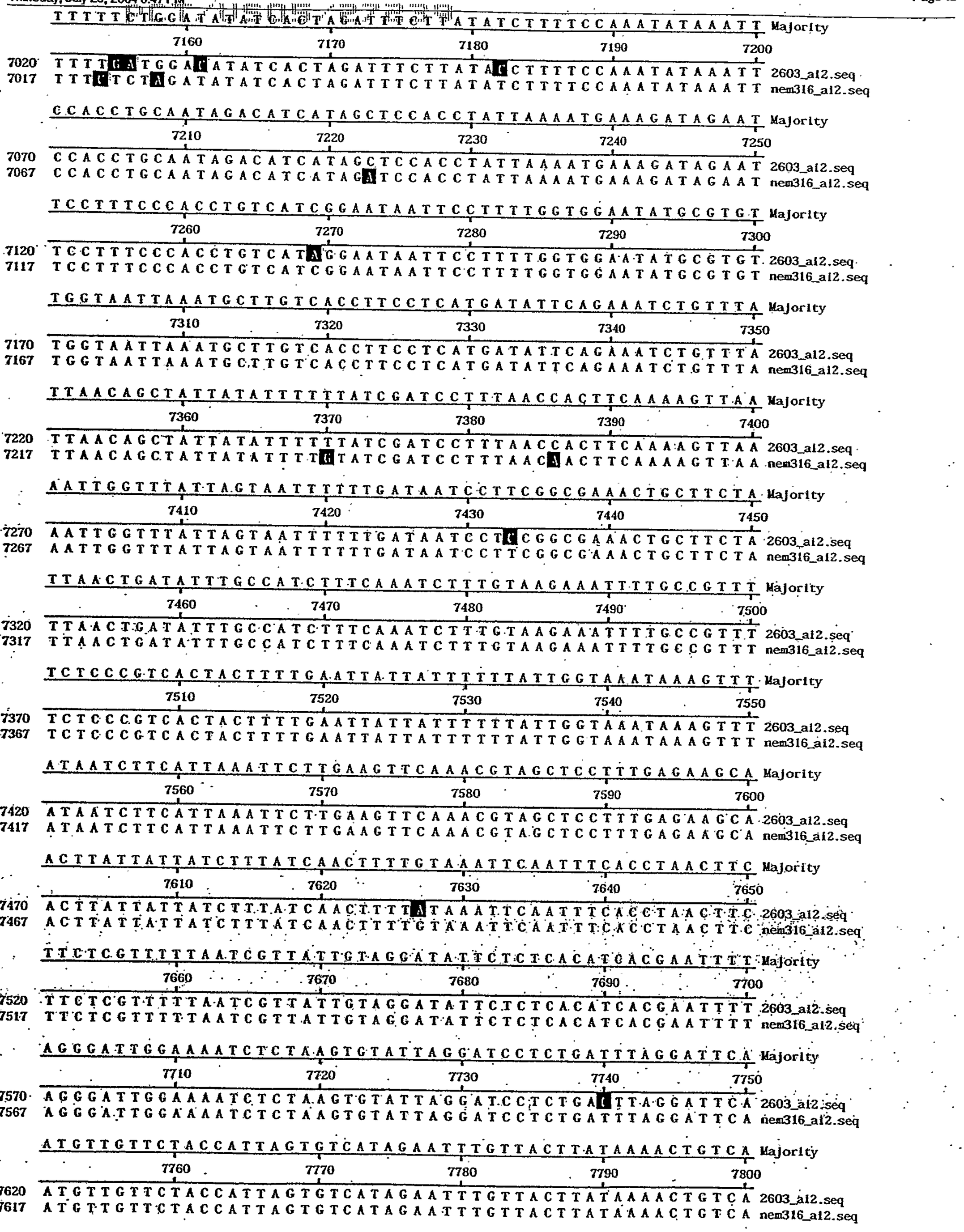

TTTTTCTGGATATATCACTAGATTTCTTATATCTTTTCCAAATATAAATT Majority
7160 7170 7180 7190 7200
7020 TTTTGATGGATATATCACTAGATTTCTTATACTTTTCCAAATATAAATT 2603_al2.seq
7017 TTTCTCTAGATATATCACTAGATTTCTTATATCTTTTCCAAATATAAATT nem316_al2.seq
CCACCTGCAATAGACATCATAGCTCCACCTATTAAAATGAAAGATAGAAT Majority
7210 7220 7230 7240 7250
7070 CCACCTGCAATAGACATCATAGCTCCACCTATTAAAATGAAAGATAGAAT 2603_al2.seq
7067 CCACCTGCAATAGACATCATAGATCCACCTATTAAAATGAAAGATAGAAT nem316_al2.seq
TCCTTTCCCACCTGTCATCGGAATAATTCCTTTTGGTGGAATATGCGTGT Majority
7260 7270 7280 7290 7300
7120 TCCTTTCCCACCTGTCATAGGAATAATTCCTTTTGGTGGAATATGCGTGT 2603_al2.seq
7117 TCCTTTCCCACCTGTCATCGGAATAATTCCTTTTGGTGGAATATGCGTGT nem316_al2.seq
TGGTAATTAAATGCTTGTCACCTTCCTCATGATATTCAGAAATCTGTTTA Majority
7310 7320 7330 7340 7350
7170 TGGTAATTAAATGCTTGTCACCTTCCTCATGATATTCAGAAATCTGTTTA 2603_al2.seq
7167 TGGTAATTAAATGCTTGTCACCTTCCTCATGATATTCAGAAATCTGTTTA nem316_al2.seq
TTAACAGCTATTATATTTTTTATCGATCCTTTAACCACTTCAAAAGTTAA Majority
7360 7370 7380 7390 7400
7220 TTAACAGCTATTATATTTTTTATCGATCCTTTAACCACTTCAAAAGTTAA 2603_al2.seq
7217 TTAACAGCTATTATATTTTGTATCGATCCTTTAACAACTTCAAAAGTTAA nem316_al2.seq
AATTGGTTTATTAGTAATTTTTTGATAATCCTTCGGCGAAACTGCTTCTA Majority
7410 7420 7430 7440 7450
7270 AATTGGTTTATTAGTAATTTTTTGATAATCCTCCGGCGAAACTGCTTCTA 2603_al2.seq
7267 AATTGGTTTATTAGTAATTTTTTGATAATCCTTCGGCGAAACTGCTTCTA nem316_al2.seq
TTAACTGATATTTGCCATCTTTCAAATCTTTGTAAGAAATTTTGCCGTTT Majority
7460 7470 7480 7490 7500
7320 TTAACTGATATTTGCCATCTTTCAAATCTTTGTAAGAAATTTTGCCGTTT 2603_al2.seq
7317 TTAACTGATATTTGCCATCTTTCAAATCTTTGTAAGAAATTTTGCCGTTT nem316_al2.seq
TCTCCCGTCACTACTTTTGAATTATTATTTTTTATTGGTAAATAAAGTTT Majority
7510 7520 7530 7540 7550
7370 TCTCCCGTCACTACTTTTGAATTATTATTTTTTATTGGTAAATAAAGTTT 2603_al2.seq
7367 TCTCCCGTCACTACTTTTGAATTATTATTTTTTATTGGTAAATAAAGTTT nem316_al2.seq
ATAATCTTCATTAAATTCTTGAAGTTCAAACGTAGCTCCTTTGAGAAGCA Majority
7560 7570 7580 7590 7600
7420 ATAATCTTCATTAAATTCTTGAAGTTCAAACGTAGCTCCTTTGAGAAGCA 2603_al2.seq
7417 ATAATCTTCATTAAATTCTTGAAGTTCAAACGTAGCTCCTTTGAGAAGCA nem316_al2.seq
ACTTATTATTATCTTTATCAACTTTTGTAAATTCAATTTCACCTAACTTC Majority
7610 7620 7630 7640 7650
7470 ACTTATTATTATCTTTATCAACTTTTATAAATTCAATTTCACCTAACTTC 2603_al2.seq
7467 ACTTATTATTATCTTTATCAACTTTTGTAAATTCAATTTCACCTAACTTC nem316_al2.seq
TTCTCGTTTTTAATCGTTATTGTAGGATATTCTCTCACATCACGAATTTT Majority
7660 7670 7680 7690 7700
7520 TTCTCGTTTTTAATCGTTATTGTAGGATATTCTCTCACATCACGAATTTT 2603_al2.seq
7517 TTCTCGTTTTTAATCGTTATTGTAGGATATTCTCTCACATCACGAATTTT nem316_al2.seq
AGGGATTGGAAAATCTCTAAGTGTATTAGGATCCTCTGATTTAGGATTCA Majority
7710 7720 7730 7740 7750
7570 AGGGATTGGAAAATCTCTAAGTGTATTAGGATCCTCTGAGTTAGGATTCA 2603_al2.seq
7567 AGGGATTGGAAAATCTCTAAGTGTATTAGGATCCTCTGATTTAGGATTCA nem316_al2.seq
ATGTTGTTCTACCATTAGTGTCATAGAATTTGTTACTTATAAAACTGTCA Majority
7760 7770 7780 7790 7800
7620 ATGTTGTTCTACCATTAGTGTCATAGAATTTGTTACTTATAAAACTGTCA 2603_al2.seq
7617 ATGTTGTTCTACCATTAGTGTCATAGAATTTGTTACTTATAAAACTGTCA nem316_al2.seq

FIGURE 20K

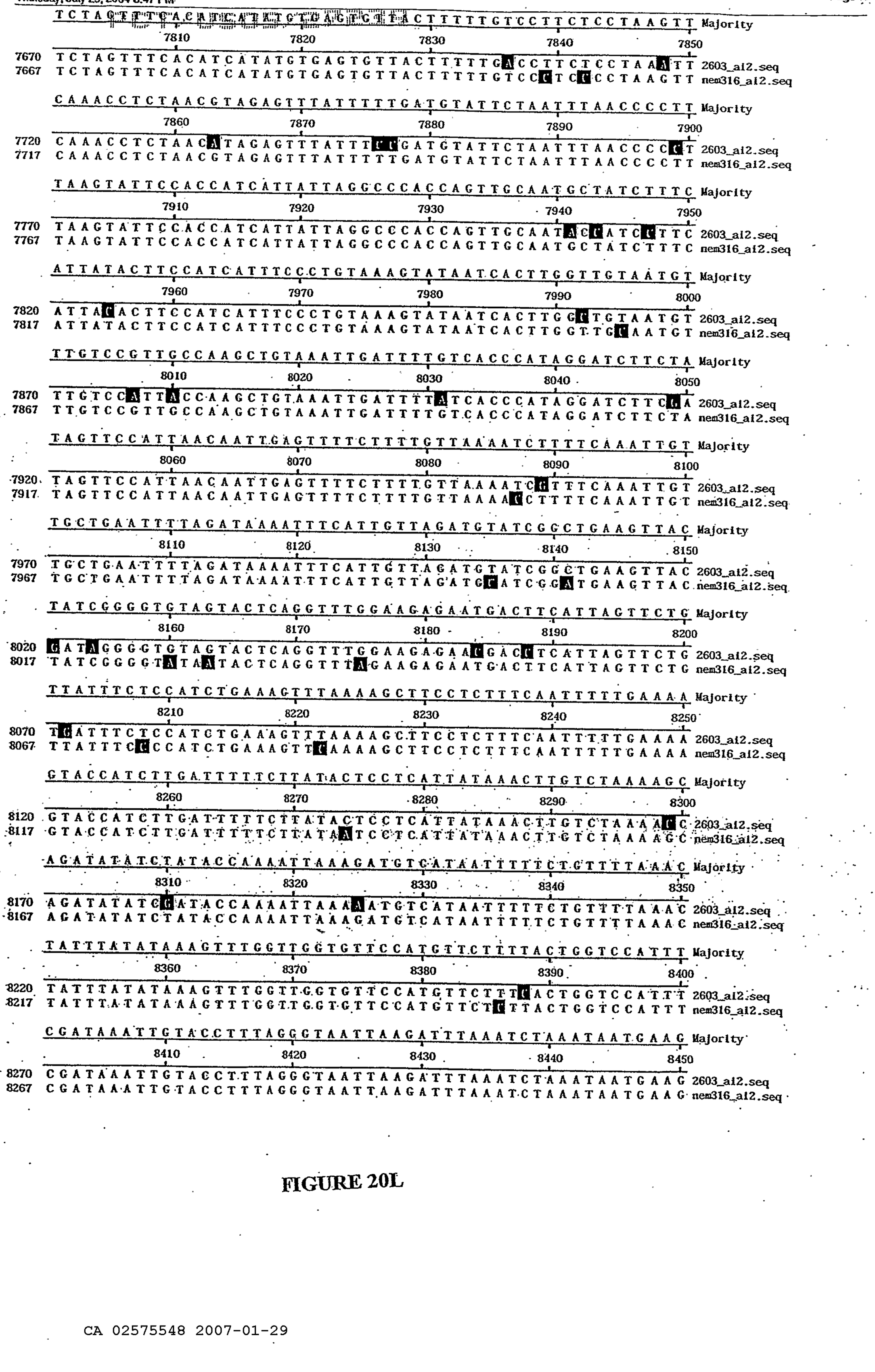

TCTAGTTTCACATCATATGTGAGTGTTACTTTTTGTCCTTCTCCTAAGTT Majority
7810 7820 7830 7840 7850
7670 TCTAGTTTCACATCATATGTGAGTGTTACTTTTTGACCTTCTCCTAAATT 2603_a12.seq
7667 TCTAGTTTCACATCATATGTGAGTGTTACTTTTTGTCCCTCCCTAAGTT nem316_a12.seq
CAAACCTCTAACGTAGAGTTTATTTTTGATGTATTCTAATTTAACCCCTT Majority
7860 7870 7880 7890 7900
7720 CAAACCTCTAACATAGAGTTTATTTCCGATGTATTCTAATTTAACCCCCT 2603_a12.seq
7717 CAAACCTCTAACGTAGAGTTTATTTTTGATGTATTCTAATTTAACCCCTT nem316_a12.seq
TAAGTATTCCACCATCATTATTAGGCCCACCAGTTGCAATGCTATCTTTC Majority
7910 7920 7930 7940 7950
7770 TAAGTATTCCACCATCATTATTAGGCCCACCAGTTGCAATACTATCCTTC 2603_a12.seq
7767 TAAGTATTCCACCATCATTATTAGGCCCACCAGTTGCAATGCTATCTTTC nem316_a12.seq
ATTATACTTCCATCATTTCCCTGTAAAGTATAATCACTTGGTTGTAATGT Majority
7960 7970 7980 7990 8000
7820 ATTACACTTCCATCATTTCCCTGTAAAGTATAATCACTTGGCTGTAATGT 2603_a12.seq
7817 ATTATACTTCCATCATTTCCCTGTAAAGTATAATCACTTGGTTGCAATGT nem316_a12.seq
TTGTCCGTTGCCAAGCTGTAAATTGATTTTGTCACCCATAGGATCTTCTA Majority
8010 8020 8030 8040 8050
7870 TTGTCCATTACCAAGCTGTAAATTGATTTTATCACCCATAGGATCTTCGA 2603_a12.seq
7867 TTGTCCGTTGCCAAGCTGTAAATTGATTTTGTCACCCATAGGATCTTCTA nem316_a12.seq
TAGTTCCATTAACAATTGAGTTTTCTTTTGTTAAAATCTTTTCAAATTGT Majority
8060 8070 8080 8090 8100
7920 TAGTTCCATTAACAATTGAGTTTTCTTTTGTTAAAATCGTTTCAAATTGT 2603_a12.seq
7917 TAGTTCCATTAACAATTGAGTTTTCTTTTGTTAAAACCTTTTCAAATTGT nem316_a12.seq
TGCTGAATTTTAGATAAAATTTCATTGTTAGATGTATCGGCTGAAGTTAC Majority
8110 8120 8130 8140 8150
7970 TGCTGAATTTTAGATAAAATTTCATTGTTAGATGTATCGGCTGAAGTTAC 2603_a12.seq
7967 TGCTGAATTTTAGATAAAATTTCATTGTTAGATGTATCGGATGAAGTTAC nem316_a12.seq
TATCGGGGTGTAGTACTCAGGTTTGGAAGAGAATGACTTCATTAGTTCTG Majority
8160 8170 8180 8190 8200
8020 GATAGGGGTGTAGTACTCAGGTTTGGAAGAGAACGACTTCATTAGTTCTG 2603_a12.seq
8017 TATCGGGGTATAATACTCAGGTTTAGAAGAGAATGACTTCATTAGTTCTG nem316_a12.seq
TTATTTCTCCATCTGAAAGTTTAAAAGCTTCCTCTTTCAATTTTTGAAAA Majority
8210 8220 8230 8240 8250
8070 TGATTTCTCCATCTGAAAGTTTAAAAGCTTCCTCTTTCAATTTTTGAAAA 2603_a12.seq
8067 TTATTTCCCCATCTGAAAGTTTAAAAGCTTCCTCTTTCAATTTTTGAAAA nem316_a12.seq
GTACCATCTTGATTTTTCTTATACTCCTCATTATAAACTTGTCTAAAAGC Majority
8260 8270 8280 8290 8300
8120 GTACCATCTTGATTTTTCTTATACTCCTCATTATAAACTTGTCTAAAACC 2603_a12.seq
8117 GTACCATCTTGATTTTTCTTATAATCCTCATTATAAACTTGTCTAAAAGC nem316_a12.seq
AGATATATCTATACCAAAATTAAAGATGTCATAATTTTTCTGTTTTAAAC Majority
8310 8320 8330 8340 8350
8170 AGATATATCGATACCAAAATTAAAAATGTCATAATTTTTCTGTTTTAAAC 2603_a12.seq
8167 AGATATATCTATACCAAAATTAAAGATGTCATAATTTTTCTGTTTTAAAC nem316_a12.seq
TATTTATATAAAGTTTGGTTGGTGTTCCATGTTCTTTTACTGGTCCATTT Majority
8360 8370 8380 8390 8400
8220 TATTTATATAAAGTTTGGTTGGTGTTCCATGTTCTTTCACTGGTCCATTT 2603_a12.seq
8217 TATTTATATAAAGTTTGGTTGGTGTTCCATGTTCTCTTACTGGTCCATTT nem316_a12.seq
CGATAAATTGTACCTTTAGGGTAATTAAGATTTAAATCTAAATAATGAAG Majority
8410 8420 8430 8440 8450
8270 CGATAAATTGTACCTTTAGGGTAATTAAGATTTAAATCTAAATAATGAAG 2603_a12.seq
8267 CGATAAATTGTACCTTTAGGGTAATTAAGATTTAAATCTAAATAATGAAG nem316_a12.seq

FIGURE 20L

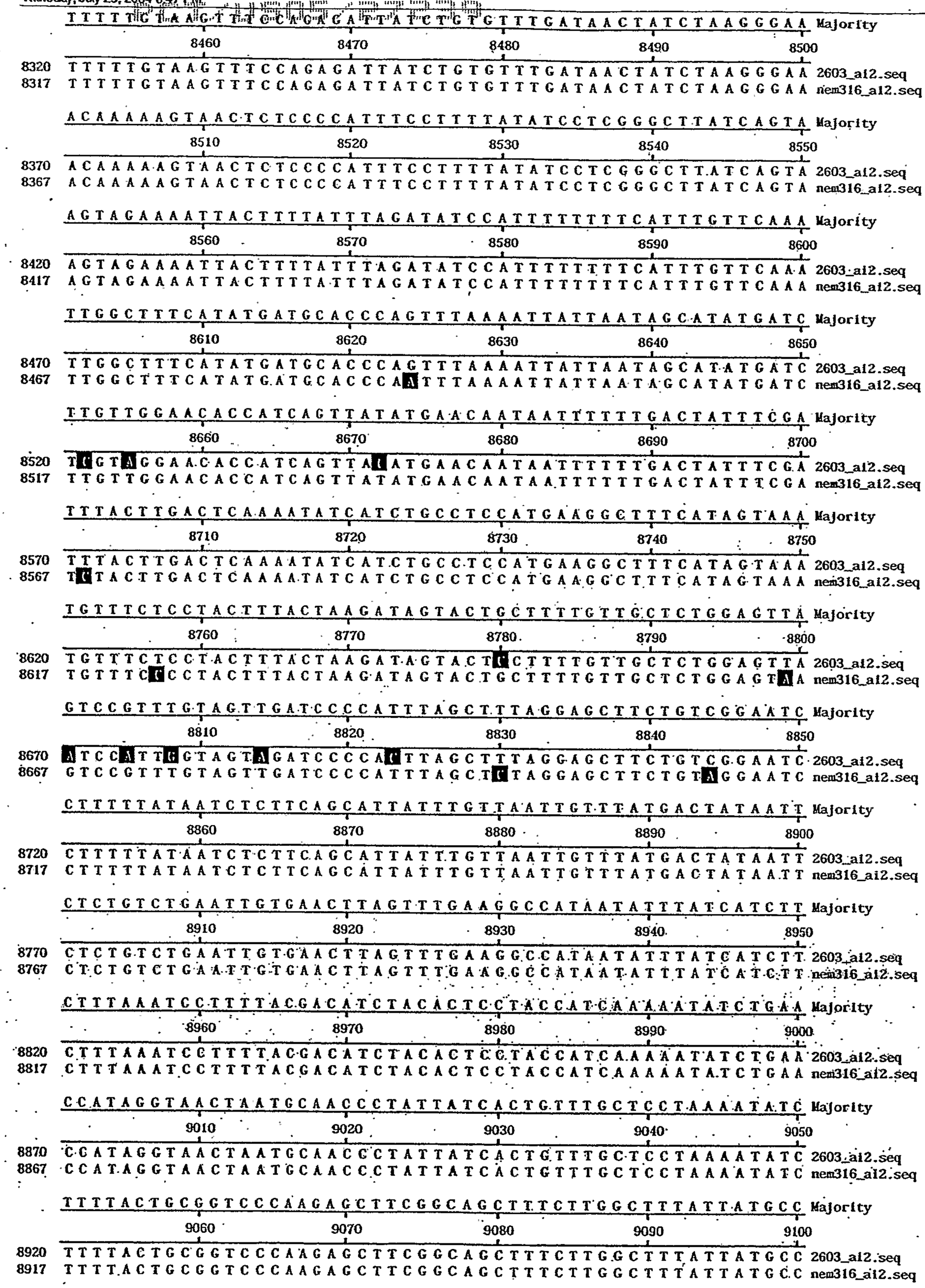

TTTTTGTAAGTTTCCAGAGATTATCTGTGTTTGATAACTATCTAAGGGAA Majority
8460 8470 8480 8490 8500
8320 TTTTTGTAAGTTTCCAGAGATTATCTGTGTTTGATAACTATCTAAGGGAA 2603_a12.seq
8317 TTTTTGTAAGTTTCCAGAGATTATCTGTGTTTGATAACTATCTAAGGGAA nem316_a12.seq
ACAAAAAGTAACTCTCCCCATTTCCTTTTATATCCTCGGGCTTATCAGTA Majority
8510 8520 8530 8540 8550
8370 ACAAAAAGTAACTCTCCCCATTTCCTTTTATATCCTCGGGCTTATCAGTA 2603_a12.seq
8367 ACAAAAAGTAACTCTCCCCATTTCCTTTTATATCCTCGGGCTTATCAGTA nem316_a12.seq
AGTAGAAAATTACTTTTATTTAGATATCCATTTTTTTTCATTTGTTCAAA Majority
8560 8570 8580 8590 8600
8420 AGTAGAAAATTACTTTTATTTAGATATCCATTTTTTTTCATTTGTTCAAA 2603_a12.seq
8417 AGTAGAAAATTACTTTTATTTAGATATCCATTTTTTTTCATTTGTTCAAA nem316_a12.seq
TTGGCTTTCATATGATGCACCCAGTTTAAAATTATTAATAGCATATGATC Majority
8610 8620 8630 8640 8650
8470 TTGGCTTTCATATGATGCACCCAGTTTAAAATTATTAATAGCATATGATC 2603_a12.seq
8467 TTGGCTTTCATATGATGCACCCANTTTAAAATTATTAATAGCATATGATC nem316_a12.seq
TTGTTGGAACACCATCAGTTATATGAACAATAATTTTTTGACTATTTCGA Majority
8660 8670 8680 8690 8700
8520 TCGTAGGAACACCATCAGTTACATGAACAATAATTTTTTGACTATTTCGA 2603_a12.seq
8517 TTGTTGGAACACCATCAGTTATATGAACAATAATTTTTTGACTATTTCGA nem316_a12.seq
TTTACTTGACTCAAAATATCATCTGCCTCCATGAAGGCTTTCATAGTAAA Majority
8710 8720 8730 8740 8750
8570 TTTACTTGACTCAAAATATCATCTGCCTCCATGAAGGCTTTCATAGTAAA 2603_a12.seq
8567 TCTACTTGACTCAAAATATCATCTGCCTCCATGAAGGCTTTCATAGTAAA nem316_a12.seq
TGTTTCTCCTACTTTACTAAGATAGTACTGCTTTTGTTGCTCTGGAGTTA Majority
8760 8770 8780 8790 8800
8620 TGTTTCTCCTACTTTACTAAGATAGTACTCCTTTTGTTGCTCTGGAGTTA 2603_a12.seq
8617 TGTTTCCCCTACTTTACTAAGATAGTACTGCTTTTGTTGCTCTGGAGTAA nem316_a12.seq
GTCCGTTTGTAGTTGATCCCCATTTAGCTTTAGGAGCTTCTGTCGGAATC Majority
8810 8820 8830 8840 8850
8670 ATCCATTGGTAGTAGATCCCCACTTAGCTTTAGGAGCTTCTGTCGGAATC 2603_a12.seq
8667 GTCCGTTTGTAGTTGATCCCCATTTAGCTCTAGGAGCTTCTGTAGGAATC nem316_a12.seq
CTTTTTATAATCTCTTCAGCATTATTTGTTAATTGTTTATGACTATAATT Majority
8860 8870 8880 8890 8900
8720 CTTTTTATAATCTCTTCAGCATTATTTGTTAATTGTTTATGACTATAATT 2603_a12.seq
8717 CTTTTTATAATCTCTTCAGCATTATTTGTTAATTGTTTATGACTATAATT nem316_a12.seq
CTCTGTCTGAATTGTGAACTTAGTTTGAAGGCCATAATATTTATCATCTT Majority
8910 8920 8930 8940 8950
8770 CTCTGTCTGAATTGTGAACTTAGTTTGAAGGCCATAATATTTATCATCTT 2603_a12.seq
8767 CTCTGTCTGAATTGTGAACTTAGTTTGAAGGCCATAATATTTATCATCTT nem316_a12.seq
CTTTAAATCCTTTTACGACATCTACACTCCTACCATCAAAAATATCTGAA Majority
8960 8970 8980 8990 9000
8820 CTTTAAATCCTTTTACGACATCTACACTCCTACCATCAAAAATATCTGAA 2603_a12.seq
8817 CTTTAAATCCTTTTACGACATCTACACTCCTACCATCAAAAATATCTGAA nem316_a12.seq
CCATAGGTAACTAATGCAACCCTATTATCACTGTTTGCTCCTAAAATATC Majority
9010 9020 9030 9040 9050
8870 CCATAGGTAACTAATGCAACCCTATTATCACTGTTTGCTCCTAAAATATC 2603_a12.seq
8867 CCATAGGTAACTAATGCAACCCTATTATCACTGTTTGCTCCTAAAATATC nem316_a12.seq
TTTTACTGCGGTCCCAAGAGCTTCGGCAGCTTTCTTGGCTTTATTATGCC Majority
9060 9070 9080 9090 9100
8920 TTTTACTGCGGTCCCAAGAGCTTCGGCAGCTTTCTTGGCTTTATTATGCC 2603_a12.seq
8917 TTTTACTGCGGTCCCAAGAGCTTCGGCAGCTTTCTTGGCTTTATTATGCC nem316_a12.seq

FIGURE 20M

```
TTTGAAAATTTGGCCATCGTTATTCATTGAGTTAGAATTATCGAGTACG Majority
          9110      9120      9130      9140      9150
8970 TTTGAAAATTTGGCCATCGTTATTCATTGAGTTAGAATTATCGAGTACG 2603_a12.seq
8967 TTTGAAAATTTGGCCATCGTTATTCATTGAGTTAGAATTATCGAGTACG nem316_a12.seq

AAGACAACATCTAACGGCTTTTGTTTGTCCACTGGTTTTACTATGGTTTT Majority
          9160      9170      9180      9190      9200
9020 AAGACAACATCTAACGGCTTTTGTTTGTCCACTGGTTTTACTATGGTTTT 2603_a12.seq
9017 AAGACAACATCTAACGGCTTTTGTTTGTCCACTGGTTTTACTATGGTTTT nem316_a12.seq

TCCACTGACAGTTAACTCAATTTTATATTTATTATGAGCTAAATCACCTA Majority
          9210      9220      9230      9240      9250
9070 TCCACTGACAGTTAACTCAATTTTATATTTATTATGAGCTAAATCACCTA 2603_a12.seq
9067 TCCACTGACAGTTAACTCAATTTTATATTTATTATGAGCTAAATCACCTA nem316_a12.seq

CTTCTGAAATACGTTTAGATAATGTTCCCTCTGGAATTTCTCTTATATGC Majority
          9260      9270      9280      9290      9300
9120 CTTCTGAAATACGTTTAGATAATGTTCCCTCTGGAATTTCTCTTATATGC 2603_a12.seq
9117 CTTCTGAAATACGTTTAGATAATGTTCCCTCTGGAATTTCTCTTATATGC nem316_a12.seq

TCACCTTCACTTGAATATGGGTTAACTGCTTTTGCCTCTGACTTTCCATT Majority
          9310      9320      9330      9340      9350
9170 TCACCTTCACTTGAATATGGGTTAACTGCTTTTGCCTCTGACTTTCCATT 2603_a12.seq
9167 TCACCTTCACTTGAATATGGGTTAACTGCTTTTGCCTCTGACTTTCCATT nem316_a12.seq

TGGAACTGAACCTTTAACATGCTCAAGTTTATAAGATTCCTTTGTATCTT Majority
          9360      9370      9380      9390      9400
9220 TGGAACTGAACCTTTAACATGCTCAAGTTTATAAGATTCCTTTGTATCTT 2603_a12.seq
9217 TGGAACTGAACCTTTAACATGCTCAAGTTTATAAGATTCCTTTGTATCTT nem316_a12.seq

CATAAATTCCTGTGGGGGGATACTGCTTATCTAGTTCTTCGTGATTTTGT Majority
          9410      9420      9430      9440      9450
9270 CATAAATTCCTGTGGGGGGATACTGCTTATCTAGTTCTTCGTGATTTTGT 2603_a12.seq
9267 CATAAATTCCTGTGGGGGGATACTGCTTATCTAGTTCTTCGTGATTTTGT nem316_a12.seq

CCAATTGTGGAATTTTTTATCACCACTATTTTGTATCGTAGTTTTTCCATT Majority
          9460      9470      9480      9490      9500
9320 CCAATTGTGGAATTTTTTATCACCACTATTTTGTATCGTAGTTTTTCCATT 2603_a12.seq
9317 CCAATTGTGGAATTTTTTATCACCACTATTTTGTATCGTAGTTTTTCCATT nem316_a12.seq

ACTCTCAACCTTAACTTGCCAAGTCTGGTTAGTCTTTTTTATAACCTTCGG Majority
          9510      9520      9530      9540      9550
9370 ACTCTCAACCTTAACTTGCCAAGTCTGGTTAGTCTTTTTTATAACCTTCGG 2603_a12.seq
9367 ACTCTCAACCTTAACTTGCCAAGTCTGGTTAGTCTTTTTTATAACCTTCGG nem316_a12.seq

GCGCTGTTTCTTCTGATAAAGTATAATCTCCAGGTATGAGATTATCAAAA Majority
          9560      9570      9580      9590      9600
9420 GCGCTGTTTCTTCTGATAAAGTATAATCTCCAGGTATGAGATTATCAAAA 2603_a12.seq
9417 GCGCTGTTTCTTCTGATAAAGTATAATCTCCAGGTATGAGATTATCAAAA nem316_a12.seq

GTAGCTTCACCTGTTAGCTCAGCAGTTACTTTTTCTATTTTACTTTCTGG Majority
          9610      9620      9630      9640      9650
9470 GTAGCTTCACCTGTTAGCTCAGCAGTTACTTTTTCTATTTTACTTTCTGG 2603_a12.seq
9467 GTAGCTTCACCTGTTAGCTCAGCAGTTACTTTTTCTATTTTACTTTCTGG nem316_a12.seq

ATGAGCAGTAGTTTTTAAAACAAAGGTAGCTTTTGAAAGTGGTTTGTTCT Majority
          9660      9670      9680      9690      9700
9520 ATGAGCAGTAGTTTTTAAAACAAAGGTAGCTTTTGAAAGTGGTTTGTTCT 2603_a12.seq
9517 ATGAGCAGTAGTTTTTAAAACAAAGGTAGCTTTTGAAAGTGGTTTGTTCT nem316_a12.seq

GGTCATCTGTCTTTTTTAACAACTAACTTTCCTTTAGCACCATTTTCCGGT Majority
          9710      9720      9730      9740      9750
9570 GGTCATCTGTCTTTTTTAACAACTAACTTTCCTTTAGCACCATTTTCCGGT 2603_a12.seq
9567 GGTCATCTGTCTTTTTTAACAACTAACTTTCCTTTAGCACCATTTTCCGGT nem316_a12.seq
```

FIGURE 20N

```
                                                
```

```
                ACAACTTTTAAAATTTTTGATCACGAATGATTCAGGCAGGTTACCTTTAA Majority
                         10410      10420      10430      10440      10450
10268 ACAACTTTTAAAATTTTTGATCACGAATGATTCAGGCAGGTTACCTTTAA 2603_a12.seq
10267 ACAACTTTTAAAATTTTTTGATCACGAATGATTCAGGCAGGTTACCTTTAA nem316_a12.seq

                CTTACTTTAGTGAAAAATTTGGACTATCTTGTGCAACTGCTTATCGCATA Majority
                         10460      10470      10480      10490      10500
10318 CTTACTTTAGTGAAAAATTTGGACTATCTTGTGCAACTGCTTATCGCATA 2603_a12.seq
10317 CTTACTTTAGTGAAAAATTTGGACTATCTTGTGCAACTGCTTATCGCATA nem316_a12.seq

                CGAAAACATATTAGTCCGTTACTAGAAAAACTTGGATTTCAGATTTTCAA Majority
                         10510      10520      10530      10540      10550
10368 CGAAAACATATTAGTCCATTACTAGAAAAACTTGGATTTCAGATTGTCAA 2603_a12.seq
10367 CGAAAACATATTAGTCCGTTACTAGAAAAACTTGGATTTAAGATTTTCAA nem316_a12.seq

                AAATACTATTACCGGTGACGAGTATCGAATTCGCTATTTAATCGCATTTT Majority
                         10560      10570      10580      10590      10600
10418 AAATACTATTACCGGTGACGAGTATCGAATTCGCTATTTAATCGCATTTT 2603_a12.seq
10417 AAATACTATTACCGGTGACGAGTATCGAATTCGCTATTTAATCGCACATT nem316_a12.seq

                TAAATGCTCGATTTGGTATAGAAGTTTATCCCTTGTCTAAGATGGATAAA Majority
                         10610      10620      10630      10640      10650
10468 TAAATGCTCAATTTGGTATAGAAGTTTATCCCATGTCTAAGATGGATAAA 2603_a12.seq
10467 TAAATGCTCGATTTGGTATAGAAGTTTATCCCTTATCTAAGATGGATAAA nem316_a12.seq

                TTGCTTATCAAACGATTGTTATTAGAATACTCAACTACTTTTACTGCTTC Majority
                         10660      10670      10680      10690      10700
10518 TTGCTGATCAAACGATTGTTATTAGAACACTCAACTACTTTTACTGCTTC 2603_a12.seq
10517 TTGCTTATCAAACGATTGTTATTAGAATACTCAACTACTTTTACTGCTTC nem316_a12.seq

                TCATTACTTCCCAAATACATTTATTTTCTTTGATACATTGTTGTCTCTAT Majority
                         10710      10720      10730      10740      10750
10568 TCATTACTTCCCAAATACATTTATTTTCTTTGATACATTGTTGTCTCTAT 2603_a12.seq
10567 TCATTACTTCCCAAATACATTTATTTTCTTTGATACATTATTGTCTCTAT nem316_a12.seq

                CATGGAAACGTATTAATTATAATGTAGTTGTCCCTTACTCATCCCTTTTC Majority
                         10760      10770      10780      10790      10800
10618 CATGGAAACGTATTAATTATAATGTAGTTGTCCCTTACTCATCCCTTTTC 2603_a12.seq
10617 CATGGAAACGTATTAATTATAATGTAGTTGTCCCTTACTCACCCCTTTTC nem316_a12.seq

                ATTGAACTACAAAATATATTTATCTATGATACATTACAATATTGTGTCAA Majority
                         10810      10820      10830      10840      10850
10668 ACTGAACTACAAAATATATTTATCTATGATACATTACAATATTGTGTCAA 2603_a12.seq
10667 ATTGAACTACAAAATATATTTATCTATGATACATTACAATATTGTGTCAA nem316_a12.seq

                AAATGTTATTATAGATTCCTTTAAAATTAATTTAAAAAAAGACGATATAG Majority
                         10860      10870      10880      10890      10900
10718 AAATGTTATTATAGATTCCTTTAAAATTAATTTAAAAAAAGACGATATAG 2603_a12.seq
10717 AAATGTTATTATAGATTCCTTTAAAATTAATTTAAAAAAAGACGATATAG nem316_a12.seq

                ACTATATTTTTCTTGCTTACCTTACTTCCCATAACTCTTTTTCTAATCCA Majority
                         10910      10920      10930      10940      10950
10768 ACTATATTTTTCTTGCTTACCTTACTTCCCATAACTCTTTTTCTAATCCA 2603_a12.seq
10767 ACTATATTTTTCTTGCTTACCTTACTTCCCATAACTCTTTTTCTAATCCA nem316_a12.seq

                AATTGGACTGAGAAGCGTATCGATAATGTAATAGCTATTTTCGAAAATTA Majority
                         10960      10970      10980      10990      11000
10818 AATTGGACTGAGAAGCGTATCGATAATGTAATAGCTATTTTCGAAAATTA 2603_a12.seq
10817 AATTGGACTGAGAAGCGTATCGATAATGTAATAGCTATTTTCGAAAATTA nem316_a12.seq

                TCCCAAATTCCAAAAATTATTACAGCCACTCAAAGATGCTCTTCCCTTAT Majority
                         11010      11020      11030      11040      11050
10868 TCCCAAATTCCAAAAATTATTACAGCCACTCAAAGATGCTCTTCCCTTAT 2603_a12.seq
10867 TCCCAAATTCCAAAAATTATTACAGCCACTCAAAGATGCTCTTCCCTTAT nem316_a12.seq
```

FIGURE 20P

```
CTGGCTCCTATCATGATGAGTTGGTAAAAGTTGCTATCATTTTTTTCCGA Majority
          11060     11070     11080     11090     11100
10918 CTGGCTCCTATCATGATGAGTTGGTAAAAGTTGCTATC-TTTTTTTCCGA 2603_a12.seq
10917 CTGACTCCTATCATGACGAGTTGGTAAAAGTTGCTATCATTTTTTGCCGA nem316_a12.seq

ACATTTATTTTAGGATTAAATCAATTAATCCCTGAAACAATTTCATTTCC Majority
          11110     11120     11130     11140     11150
10967 ACATTTATTTTAGGATTAAATCAATTAATCCCTGAAACAATTTCATTTCC 2603_a12.seq
10967 ACATTTATTTTAGGATTAAATCAATTAATCCCTGAAACAATTTCATTTCC nem316_a12.seq

TTCATGGAACTATCATAGACATGATAAATTAACTACTATTCTCCGACCGA Majority
          11160     11170     11180     11190     11200
11017 TTCATGGAACTATCATAGACATGATAAATTAACTACTATTCTCCGACCGA 2603_a12.seq
11017 TTCATGGAACTATCATAGACATGATAAATTAACTACTATTCTCCGACCGA nem316_a12.seq

TAATTACAAATTGGTTAAGTGAAATTGGAGAATACACGTTTAAGGAACAA Majority
          11210     11220     11230     11240     11250
11067 TAATTACAAATTGGTTAAGTAAAATTGGAGAATACACGTTTAAAGAACAA 2603_a12.seq
11067 TAATTACAAATTGGTTAAGTGAAATTGGAGAATACACGTTTAAGGAACAA nem316_a12.seq

CATTTTCTTCTCCTTTGTGCTCATCTAGAAAGAATTATCAAAAATCATAT Majority
          11260     11270     11280     11290     11300
11117 CATTTTCTTCTCCTTTGTACTCATCTAGAAAGAATTATCAAAAATCATAT 2603_a12.seq
11117 CATTTTCTTCTCCTTTGTGCTCACCTAGAAAGAATTATCAAAAATCACAT nem316_a12.seq

TCCTCCGATACAGATAGCCGTACTAACTACAGACTTTATTAATAACCAAA Majority
          11310     11320     11330     11340     11350
11167 TCCTCCGATACAGATAGCCGTACTAACTACAGACTTTATTAATAACCAAA 2603_a12.seq
11167 TCCTCCGATACAAATAGCCGTACTAACTACAGACTTTATTAATAACAAAA nem316_a12.seq

TTTTAACAGAATGTTTATTACAGAGGTTTTCTTCTAAACAGATTCATTTC Majority
          11360     11370     11380     11390     11400
11217 TTTTAACAGAATGCTTATTACAGAGATTTTCTTCTAAAAAGATTCATTTC 2603_a12.seq
11217 TTTTAACAGAATGTTTATCACAGAGGTTTTCTTCTAAACAGATTCATTTC nem316_a12.seq

CACCCTTACTATTTATTAACTGATGATCTTTCCAATATTACTAATCTTAA Majority
          11410     11420     11430     11440     11450
11267 CACCCTTACTATTTATTAACTGATGATCTTTCCAATATTACTAATCTTAA 2603_a12.seq
11267 CACCCTTACTATTTATTAACTGATGATCTTTCCAATATTACTAATCTTAA nem316_a12.seq

TCCAGATATTATTATTACCAATCCAAAGCTTTCTTCCTTTATCAAACATG Majority
          11460     11470     11480     11490     11500
11317 TCCAGACATTATTATTACCAATAAAAAGCTTTCTCCTTTATCAAACATG 2603_a12.seq
11317 TCCAGATATTATTATTACCAACCAAAGCTTTCTTCCTTTATAAAACATG nem316_a12.seq

AGATTTCTTCAGAGAGTTTAATTACATATATTGATTTGATTCATACTTCA Majority
          11510     11520     11530     11540     11550
11367 AGATTTCTTCAGAGAGTTTAATTACATATATTGATCTGATTAATACTTCA 2603_a12.seq
11367 AGATTTCTTCAGAGAGTTTAATTACATATATTGATTTGATTCATACTCCA nem316_a12.seq

GACCAGATCAATCAAATCCAAGAAATTATTTCATCAATACAGGAAGAAAA Majority
          11560     11570     11580     11590     11600
11417 GACCAAATCAATCAAATCCAAAAAATTATTTCATCAATACAGGAAGAAAA 2603_a12.seq
11417 GACCAGATCAATCAAATCCAAGAAATTATTTCATCAATACAAGAAGAAAA nem316_a12.seq

ATATTGTAAACTTTTTGCAAAAACTAATGAAATAACTACTCGTAGCTCCT Majority
          11610     11620     11630     11640     11650
11467 ATATTGCAAACTTTTTGCAAAAACTAATGAAATAACTAATCGTAGCTCCT 2603_a12.seq
11467 ATATTGTAAACTTTTTGCAAAAAACAATGAAATAACTACCGTAGCTCCT nem316_a12.seq

ATAACTCTTAAAAATTAACATTAAAAAGCTAGAGCATTGTGTAATGCTCT Majority
          11660     11670     11680     11690     11700
11517 ATAACTCTTAAAAATTAACATTAAAAAGCTAGAGCATTGTGTAATGCTCT 2603_a12.seq
11517 ATAACTCTTAAAAATTAACATTAAAAAGCTAGAGCATTGTGTAATGCTCT nem316_a12.seq
```

FIGURE 20Q

```
                                                
```

```
                TTCTCTTGGATACCGCATAAAAATCTGGACGATAATGCTTAACACGCGCT Majority
                        12360     12370     12380     12390     12400
12217 TTCTCTTGGATACCGCATAAAAATCTGGACGATAATGCTTAACACGCGCT 2603_a12.seq
12217 TTCTCTTGGATACCGCATAAAAATCTGGACGATAATGCTTAACACGCGCT nem316_a12.seq

                GTGAGAAGATGTTCATAGATAGCTCCAAAGAAATCTAAAAAACGATTATT Majority
                        12410     12420     12430     12440     12450
12267 GTGAGAAGATGTTCATAGATAGCTCCAAAGAAATCTAAAAAACGATTATT 2603_a12.seq
12267 GTGAGAAGATGTTCATAGATAGCTCCAAAGAAATCTAAAAAACGATTATT nem316_a12.seq

                GACAGAAAAATGACTTGACCCATGGTCTAAAACAATACTAGGTAAATGGT Majority
                        12460     12470     12480     12490     12500
12317 GACAGAAAAATGACTTGACCCATGGTCTAAAACAATACTAGGTAAATGGT 2603_a12.seq
12317 GACAGAAAAATGACTTGACCCATGGTCTAAAACAATACTAGGTAAATGGT nem316_a12.seq

                GATTCTTTGCAAAAGATAGCCCTTCTAGCGTTGTTAACTGAAAACGTGTA Majority
                        12510     12520     12530     12540     12550
12367 GATTCTTTGCAAAAGATAGCCCTTCTAGCGTTGTTAACTGAAAACGTGTA 2603_a12.seq
12367 GATTCTTTGCAAAAGATAGCCCTTCTAGCGTTGTTAACTGAAAACGTGTA nem316_a12.seq

                TTACAAATCACAAAATCAATATTTTCATCTGAAACATATTTCATCAGCGT Majority
                        12560     12570     12580     12590     12600
12417 TTACAAATCACAAAATCAATATTTTCATCTGAAACATATTTCATCAGCGT 2603_a12.seq
12417 TTACAAATCACAAAATCAATATTTTCATCTGAAACATATTTCATCAGCGT nem316_a12.seq

                GTTGTATTCTCGATTTTTGTTAATAATAGGATAGCGCTGCTTGACAATAT Majority
                        12610     12620     12630     12640     12650
12467 GTTGTATTCTCGATTTTTGTTAATAATAGGATAGCGCTGCTTGACAATAT 2603_a12.seq
12467 GTTGTATTCTCGATTTTTGTTAATAATAGGATAGCGCTGCTTGACAATAT nem316_a12.seq

                TTTTGGTCGGTAAACGGTAAATTTTTCTACCCTTGTCTTCATCTATAATC Majority
                        12660     12670     12680     12690     12700
12517 TTTTGGTCGGTAAACGGTAAATTTTTCTACCCTTGTCTTCATCTATAATC 2603_a12.seq
12517 TTTTGGTCGGTAAACGGTAAATTTTTCTACCCTTGTCTTCATCTATAATC nem316_a12.seq

                GGTAAATCACCATGATTAGTTGTTACAATAACAACACGGTAGCCACGCTT Majority
                        12710     12720     12730     12740     12750
12567 GGTAAATCACCATGATTAGTTGTTACAATAACAACACGGTAGCCACGCTT 2603_a12.seq
12567 GGTAAATCACCATGATTAGTTGTTACAATAACAACACGGTAGCCACGCTT nem316_a12.seq

                AACCAAATCTGCTGTCATTTTATCTGTATAACGTTCAATACCTCCGAGGA Majority
                        12760     12770     12780     12790     12800
12617 AACCAAATCTGCTGTCATTTTATCTGTATAACGTTCAATACCTCCGAGGA 2603_a12.seq
12617 AACCAAATCTGCTGTCATTTTATCTGTATAACGTTCAATACCTCCGAGGA nem316_a12.seq

                AGGGTAGATAATATCCTGAGAAAACAGCAACTGTTTTTACCTTATTTTCC Majority
                        12810     12820     12830     12840     12850
12667 AGGGTAGATAATATCCTGAGAAAACAGCAACTGTTTTTACCTTATTTTCC 2603_a12.seq
12667 AGGGTAGATAATATCCTGAGAAAACAGCAACTGTTTTTACCTTATTTTCC nem316_a12.seq

                ATATTTATCCACTTTCATCAATAAGCCATCTTTTAAGCCTTTAATCATAG Majority
                        12860     12870     12880     12890     12900
12717 ATATTTATCCACTTTCATCAATAAGCCATCTTTTAAGCCTTTAATCATAG 2603_a12.seq
12717 ATATTTATCCACTTTCATCAATAAGCCATCTTTTAAGCCTTTAATCATAG nem316_a12.seq

                CAACTATTTTTTTGCTCTTTTGCTCTTCTGCTACCAACACTCGAACAAAT Majority
                        12910     12920     12930     12940     12950
12767 CAACTATTTTTTTGCTCTTTTGCTCTTCTGCTACCAACACTCGAACAAAT 2603_a12.seq
12767 CAACTATTTTTTTGCTCTTTTGCTCTTCTGCTACCAACACTCGAACAAAT nem316_a12.seq

                TCATTTCGCATAAATACTAAATATTTGTATCGCTTCTTCTTACCATATTT Majority
                        12960     12970     12980     12990     13000
12817 TCATTTCGCATAAATACTAAATATTTGTATCGCTTCTTCTTACCATATTT 2603_a12.seq
12817 TCATTTCGCATAAATACTAAATATTTGTATCGCTTCTTCTTACCATATTT nem316_a12.seq
```

FIGURE 20S

108/487

```
TTTTATAATATAGATCGCATTGCGTATCATGTAATATTTTCGAAATGGTG Majority
          13010     13020     13030     13040     13050
12867 TTTTATAATATAGATCGCATTGCGTATCATGTAATATTTTCGAAATGGTG 2603_a12.seq
12867 TTTTATAATATAGATCGCATTGCGTATCATGTAATATTTTCGAAATGGTG nem316_a12.seq

AATGATTCAATACATGAAAAACATGGCCAAATTTTTTAACTCGTGAAGAG Majority
          13060     13070     13080     13090     13100
12917 AATGATTCAATACATGAAAAACATGGCCAAATTTTTTAACTCGTGAAGAG 2603_a12.seq
12917 AATGATTCAATACATGAAAAACATGGCCAAATTTTTTAACTCGTGAAGAG nem316_a12.seq

TGTCCAATTTCGTGTAACAGACCAATAAAATTAACCTGATAAGTCTTATA Majority
          13110     13120     13130     13140     13150
12967 TGTCCAATTTCGTGTAACAGACCAATAAAATTAACCTGATAAGTCTTATA 2603_a12.seq
12967 TGTCCAATTTCGTGTAACAGACCAATAAAATTAACCTGATAAGTCTTATA nem316_a12.seq

TCCCATCTCTGACAGACGATAATTCATTTCAGAGTCAACAAAATCAATAA Majority
          13160     13170     13180     13190     13200
13017 TCCCATCTCTGACAGACGATAATTCATTTCAGAGTCAACAAAATCAATAA 2603_a12.seq
13017 TCCCATCTCTGACAGACGATAATTCATTTCAGAGTCAACAAAATCAATAA nem316_a12.seq

ACATCTCTTCTGCAAAGCCAGATGTTTCTTCGAAAACGCTCGTTTTCATT Majority
          13210     13220     13230     13240     13250
13067 ACATCTCTTCTGCAAAGCCAGATGTTTCTTCGAAAACGCTCGTTTTCATT 2603_a12.seq
13067 ACATCTCTTCTGCAAAGCCAGATGTTTCTTCGAAAACGCTCGTTTTCATT nem316_a12.seq

AAAGCAGCCGAAGTAATACACTCTTCAATTTCTTTATAGTCAAATTCTTG Majority
          13260     13270     13280     13290     13300
13117 AAAGCAGCCGAAGTAATACACTCTTCAATTTCTTTATAGTCAAATTCTTG 2603_a12.seq
13117 AAAGCAGCCGAAGTAATACACTCTTCAATTTCTTTATAGTCAAATTCTTG nem316_a12.seq

CATCACTAAATTTTCACGGTTCATATCTTGATACAAACAAGATAACATAC Majority
          13310     13320     13330     13340     13350
13167 CATCACTAAATTTTCACGGTTCATATCTTGATACAAACAAGATAACATAC 2603_a12.seq
13167 CATCACTAAATTTTCACGGTTCATATCTTGATACAAACAAGATAACATAC nem316_a12.seq

CGACCTTAGGTAAATGAAGGTAATTTTCATAATTATCTATCAAATCACCT Majority
          13360     13370     13380     13390     13400
13217 CGACCTTAGGTAAATGAAGGTAATTTTCATAATTATCTATCAAATCACCT 2603_a12.seq
13217 CGACCTTAGGTAAATGAAGGTAATTTTCATAATTATCTATCAAATCACCT nem316_a12.seq

AGGACAACCGAATCTTGATCTAAAGTCAAGAACCAATCAAATTCTTGTGC Majority
          13410     13420     13430     13440     13450
13267 AGGACAACCGAATCTTGATCTAAAGTCAAGAACCAATCAAATTCTTGTGC 2603_a12.seq
13267 AGGACAACCGAATCTTGATCTAAAGTCAAGAACCAATCAAATTCTTGTGC nem316_a12.seq

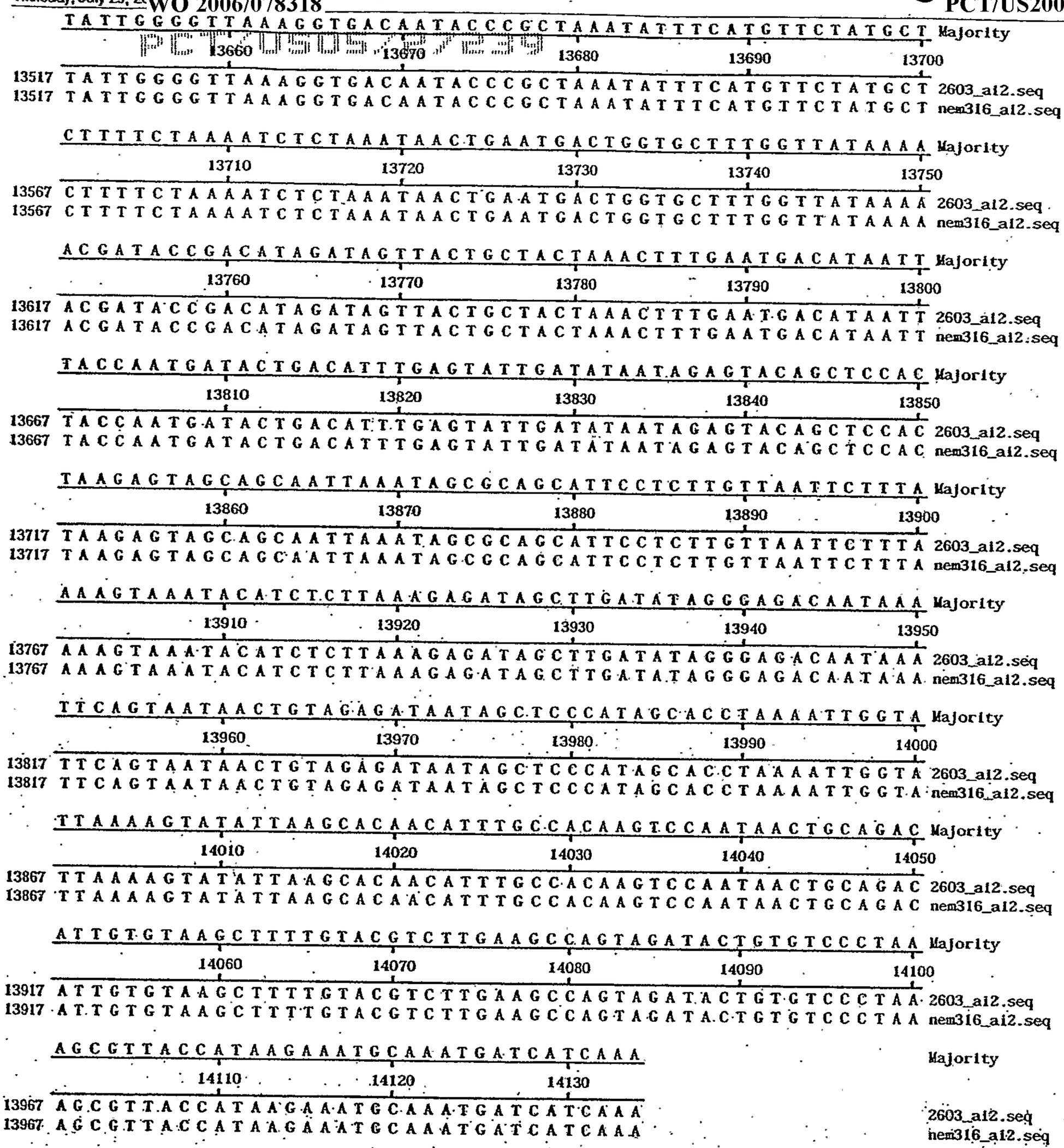
TACTGCAAATTGACCGATACAGTTCAAAGCATATGCAATCCCTTTATTTT Majority
          13460     13470     13480     13490     13500
13317 TACTGCAAATTGACCGATACAGTTCAAAGCATATGCAATCCCTTTATTTT 2603_a12.seq
13317 TACTGCAAATTGACCGATACAGTTCAAAGCATATGCAATCCCTTTATTTT nem316_a12.seq

CTGTTAAATAATCAACAGTTAGGTGCCCCTCTTCATTATAATCGGCTACT Majority
          13510     13520     13530     13540     13550
13367 CTGTTAAATAATCAACAGTTAGGTGCCCCTCTTCATTATAATCGGCTACT 2603_a12.seq
13367 CTGTTAAATAATCAACAGTTAGGTGCCCCTCTTCATTATAATCGGCTACT nem316_a12.seq

AATTGAGAAATTTCTTCCTTATTTTTCGAGCCATTATCTACGATATAGAT Majority
          13560     13570     13580     13590     13600
13417 AATTGAGAAATTTCTTCCTTATTTTTCGAGCCATTATCTACGATATAGAT 2603_a12.seq
13417 AATTGAGAAATTTCTTCCTTATTTTTCGAGCCATTATCTACGATATAGAT nem316_a12.seq

GTGGCTTACTTGAGGATAAATTGCTCGAATGTTCTGATCTAAGCGTTCAA Majority
          13610     13620     13630     13640     13650
13467 GTGGCTTACTTGAGGATAAATTGCTCGAATGTTCTGATCTAAGCGTTCAA 2603_a12.seq
13467 GTGGCTTACTTGAGGATAAATTGCTCGAATGTTCTGATCTAAGCGTTCAA nem316_a12.seq
```

FIGURE 20T



Decoration 'Decoration #1': Shade (with solid black) residues that differ from the Consensus.

FIGURE 20U

```
TCCACATCGGTCCAATTAACATATGACGTGGCGCATCACCAGTAATTCGG Majority
         10        20        30        40        50
1   TCCACATCGGTCCAATTAACATATGACGTGGCGCATCACCAGTAATTCGG cohl_al2.seq
1   TCCACATCGGTCCAATTAACATATGACGTGGCGCATCACCAGTAATTCGG a909_al2.seq

TGAATAACAATATGTTTTGGAATAATCTCCAGTTGGTCACAAATAATCGA Majority
         60        70        80        90       100
51  TGAATAACAATATGTTTTGGAATAATCTCCAGTTGGTCACAAATAATCGA cohl_al2.seq
51  TGAATAACAATATGTTTTGGAATAATCTCCAGTTGGTCACAAATAATCGA a909_al2.seq

AATATAGTCTTCTTGACTTAACAAACGTAAACGACCTTCATGGTAATCTC Majority
        110       120       130       140       150
101 AATATAGTCTTCTTGACTTAACAAACGTAAACGACCTTCATGGTAATCTC cohl_al2.seq
101 AATATAGTCTTCTTGACTTAACAAACGTAAACGACCTTCATGGTAATCTC a909_al2.seq

TCTGCATTCTTGTATTAGTCATAAGATGCAGAAGGTGTAATTTTATACCC Majority
        160       170       180       190       200
151 TCTGCATTCTTGTATTAATCATAAGATGCAGAAGGTGTAATTTTATACCC cohl_al2.seq
151 TCTGCATTCTTGTATTAGTCATAAGATGCAGAAGGTGTAATTTTATACCC a909_al2.seq

TGAATATCATTATCCGTAACACATCGACGAACATTTTCCACCATCATATC Majority
        210       220       230       240       250
201 TGAATATCATTATCCGTAACACATCGACGAACATTTTCCACCATCATATC cohl_al2.seq
201 TGAATATCATTATCCGTAACACATCGACGAACATTTTCCACCATCATATC a909_al2:seq

ATGTGTCTCCCCTGGGAGACCATTTATTAGGTGAGAAACGATTTCTACTT Majority
        260       270       280       290       300
251 ATGTGTCTCCCCTGGGAGACCATTTATTAGGTGAGAAACGATTTCTACTT cohl_al2.seq
251 ATGTGTCTCCCCTGGGAGACCATTTATTAGGTGAGAAACGATTTCTACTT a909_al2:seq

TAGGAGCTAATTCTCGTATTCTCTTAACAGTTTTTTTGTAAAGGTCATAT Majority
        310       320       330       340       350
301 TAGGAGCTAATTCTCGTACTCTCTTAACAGTTTTTTTGTAAAGGTCATAT cohl_al2.seq
301 TAGGAGCTAATTCTCGTATTCTCTTAACAGTTTTTTTGTAAAGGTCATAT a909_al2.seq

GAATGTGCTCTATTTATTAATGCAGAAGTTGCTTCATAAGTTGTCTGAAG Majority
        360       370       380       390       400
351 GAATGTGCTCTATTTATTAATGCAGAAGTTGCTTCATAAGTTGTCTGAAG cohl_al2.seq
351 GAATGTGCTCTATTTATTAATGCAGAAGTTGCTTCATAAGTTGTCTGAAG a909_al2.seq

GCCTAATTCTAAAGTCACATGCATTCTTTCAGAAAGTTCAGCGAGATAGT Majority
        410       420       430       440       450
401 GCCTAATTCTAAAGTCACATGCATTCTTTCAGAAAGTTCAGCGAGATAGT cohl_al2.seq
401 GCCTAATTCTAAAGTCACATGCATTCTTTCAGAAAGTTCAGCGAGATAGT a909_al2.seq

ATATAGTTTCATCAGGTAAGCAATCCGGCCTTGTTCCGATGTTGATCCCG Majority
        460       470       480       490       500
451 ATATAGTTTCATCAGGTAAGCAATCCGGCCTTGTTCCGATGTTGATCCCG cohl_al2.seq
451 ATATAGTTTCATCAGGTAAGCAATCCGGCCTTGTTCCGATGTTGATCCCG a909_al2.seq

ATAACTCCTGGCTCATTAATAGCCTGTTCGTAACGCTCTTTAATTATCTC Majority
        510       520       530       540       550
501 ATAACTCCTGGCTCATTAATAGCCTGTTCGTAACGCTCTTTAATTATCTC cohl_al2.seq
501 ATAACTCCTGGCTCATTAATAGCCTGTTCGTAACGCTCTTTAATTATCTC a909_al2.seq

TAACTTAGCATGGGTATTGGTAAAATTTTGAAAATAGACTAAGTATTTAT Majority
        560       570       580       590       600
551 TAACTTAGCATGGGTATTGGTAAAATTTTGAAAATAGACTAAGTATTTAT cohl_al2.seq
551 TAACTTAGCATGGGTATTGGTAAAATTTTGAAAATAGACTAAGTATTTAT a909_al2.seq

TAACCTCAGGCCACTTTCTATGCATGAAATCAATTTCTTTATAGAATTGT Majority
        610       620       630       640       650
601 TAACCTCAGGCCACTTTCTATGCATGAAATCAATTTCTTTATAGAATTGT cohl_al2.seq
601 TAACCTCAGGCCACTTTCTATGCATGAAATCAATTTCTTTATAGAATTGT a909_al2.seq
```

Figure 21

```
                                                
```

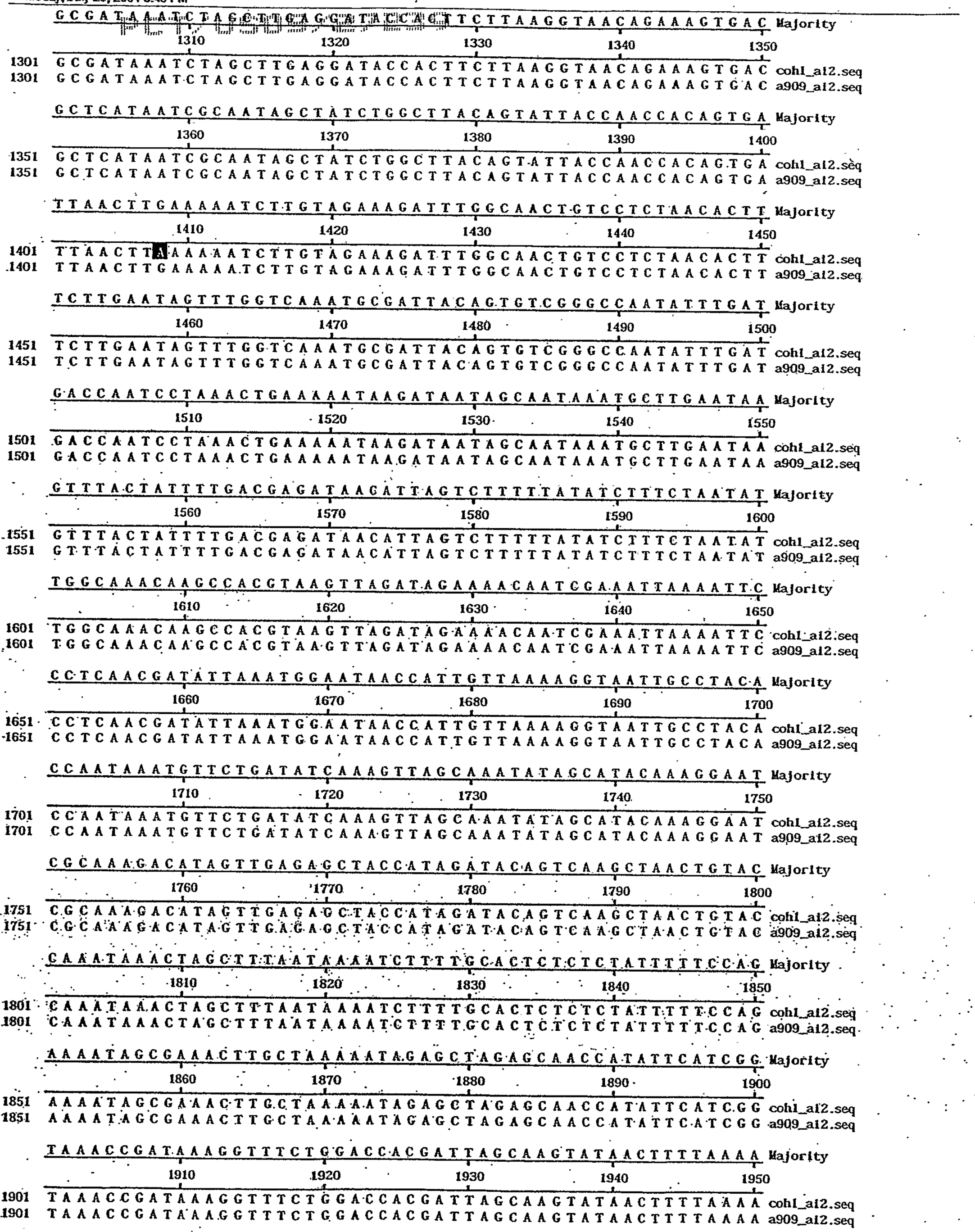

GCGATAAATCTAGCTTGAGGATACCACTTCTTAAGGTAACAGAAAGTGAC Majority
1310 1320 1330 1340 1350
1301 GCGATAAATCTAGCTTGAGGATACCACTTCTTAAGGTAACAGAAAGTGAC cohl_a12.seq
1301 GCGATAAATCTAGCTTGAGGATACCACTTCTTAAGGTAACAGAAAGTGAC a909_a12.seq
GCTCATAATCGCAATAGCTATCTGGCTTACAGTATTACCAACCACAGTGA Majority
1360 1370 1380 1390 1400
1351 GCTCATAATCGCAATAGCTATCTGGCTTACAGTATTACCAACCACAGTGA cohl_a12.seq
1351 GCTCATAATCGCAATAGCTATCTGGCTTACAGTATTACCAACCACAGTGA a909_a12.seq
TTAACTTGAAAAATCTTGTAGAAAGATTTGGCAACTGTCCTCTAACACTT Majority
1410 1420 1430 1440 1450
1401 TTAACTTGAAAAATCTTGTAGAAAGATTTGGCAACTGTCCTCTAACACTT cohl_a12.seq
1401 TTAACTTGAAAAATCTTGTAGAAAGATTTGGCAACTGTCCTCTAACACTT a909_a12.seq
TCTTGAATAGTTTGGTCAAATGCGATTACAGTGTCGGGCCAATATTTGAT Majority
1460 1470 1480 1490 1500
1451 TCTTGAATAGTTTGGTCAAATGCGATTACAGTGTCGGGCCAATATTTGAT cohl_a12.seq
1451 TCTTGAATAGTTTGGTCAAATGCGATTACAGTGTCGGGCCAATATTTGAT a909_a12.seq
GACCAATCCTAAACTGAAAAATAAGATAATAGCAATAAATGCTTGAATAA Majority
1510 1520 1530 1540 1550
1501 GACCAATCCTAAACTGAAAAATAAGATAATAGCAATAAATGCTTGAATAA cohl_a12.seq
1501 GACCAATCCTAAACTGAAAAATAAGATAATAGCAATAAATGCTTGAATAA a909_a12.seq
GTTTACTATTTTGACGAGATAAGATTAGTCTTTTTATATCTTTCTAATAT Majority
1560 1570 1580 1590 1600
1551 GTTTACTATTTTGACGAGATAACATTAGTCTTTTTATATCTTTCTAATAT cohl_a12.seq
1551 GTTTACTATTTTGACGAGATAACATTAGTCTTTTTATATCTTTCTAATAT a909_a12.seq
TGGCAAACAAGCCACGTAAGTTAGATAGAAAACAATCGAAATTAAAATTC Majority
1610 1620 1630 1640 1650
1601 TGGCAAACAAGCCACGTAAGTTAGATAGAAAACAATCGAAATTAAAATTC cohl_a12.seq
1601 TGGCAAACAAGCCACGTAAGTTAGATAGAAAACAATCGAAATTAAAATTC a909_a12.seq
CCTCAACGATATTAAATGGAATAACCATTGTTAAAAGGTAATTGCCTACA Majority
1660 1670 1680 1690 1700
1651 CCTCAACGATATTAAATGGAATAACCATTGTTAAAAGGTAATTGCCTACA cohl_a12.seq
1651 CCTCAACGATATTAAATGGAATAACCATTGTTAAAAGGTAATTGCCTACA a909_a12.seq
CCAATAAATGTTCTGATATCAAAGTTAGCAAATATAGCATACAAAGGAAT Majority
1710 1720 1730 1740 1750
1701 CCAATAAATGTTCTGATATCAAAGTTAGCAAATATAGCATACAAAGGAAT cohl_a12.seq
1701 CCAATAAATGTTCTGATATCAAAGTTAGCAAATATAGCATACAAAGGAAT a909_a12.seq
CGCAAAGACATAGTTGAGAGCTACCATAGATACAGTCAAGCTAACTGTAC Majority
1760 1770 1780 1790 1800
1751 CGCAAAGACATAGTTGAGAGCTACCATAGATACAGTCAAGCTAACTGTAC cohl_a12.seq
1751 CGCAAAGACATAGTTGAGAGCTACCATAGATACAGTCAAGCTAACTGTAC a909_a12.seq
CAAATAAACTAGCTTTAATAAAATCTTTTGCACTCTCTCTATTTTTCCAG Majority
1810 1820 1830 1840 1850
1801 CAAATAAACTAGCTTTAATAAAATCTTTTGCACTCTCTCTATTTTTCCAG cohl_a12.seq
1801 CAAATAAACTAGCTTTAATAAAATCTTTTGCACTCTCTCTATTTTTCCAG a909_a12.seq
AAAATAGCGAAACTTGCTAAAAATAGAGCTAGAGCAACCATATTCATCGG Majority
1860 1870 1880 1890 1900
1851 AAAATAGCGAAACTTGCTAAAAATAGAGCTAGAGCAACCATATTCATCGG cohl_a12.seq
1851 AAAATAGCGAAACTTGCTAAAAATAGAGCTAGAGCAACCATATTCATCGG a909_a12.seq
TAAACCGATAAAGGTTTCTGGACCACGATTAGCAAGTATAACTTTTAAAA Majority
1910 1920 1930 1940 1950
1901 TAAACCGATAAAGGTTTCTGGACCACGATTAGCAAGTATAACTTTTAAAA cohl_a12.seq
1901 TAAACCGATAAAGGTTTCTGGACCACGATTAGCAAGTATAACTTTTAAAA a909_a12.seq

FIGURE 21B

```
                                                
```

```
                                                
```

```
GACATTAGTTCATAACCTTGAGCTGTTTTAGTCTGAATAAATAGATAAAT Majority
          3260      3270      3280      3290      3300
3251 GACATTAGTTCATAACCTTGAGCTGTTTTAGTCTGAATAAATAGATAAAT cohl_al2.seq
3251 GACATTAGTTCATAACCTTGAGCTGTTTTAGTCTGAATAAATAGATAAAT a909_al2.seq

CCCTTGAGGAAGATTGTTCGCAACAATACCTTCAGCCGGTAAATTATCAA Majority
          3310      3320      3330      3340      3350
3301 CCCTTGAGGAAGATTGTTCGCAACAATACCTTCAGCCGGTAAATTATCAA cohl_al2.seq
3301 CCCTTGAGGAAGATTGTTCGCAACAATACCTTCAGCCGGTAAATTATCAA a909_al2.seq

ACGTTTGTAAAGGTTGAGTTTTATGAACAGCTTTTGTTAGTAGATTGACG Majority
          3360      3370      3380      3390      3400
3351 ACGTTTGTAAAGGTTGAGTTTTATGAACAGCTTTTGTTAGTAGATTGACG cohl_al2.seq
3351 ACGTTTGTAAAGGTTGANTTTTATGAACAGCTTTTGTTAGTAGATTGACN a909_al2.seq

TATTTGGCTTGGTTACTATCAAGGTTTACTTGTGTTAGATCATCGTCTTT Majority
          3410      3420      3430      3440      3450
3401 TATTTGGCTTGGTTACTATCAAGGTTTACTTGTGTTAGATCATCGTCTTT cohl_al2.seq
3401 TATTTGGCTTGGTTACTATCAAGGTTTACTTGTGTTANATCATCGTCTTT a909_al2.seq

TATTCCAATACCTTGAAATGGGGTAGTTAGAGTAAAAACTTGGTTACCAT Majority
          3460      3470      3480      3490      3500
3451 TATTCCAATACCTTGAAATGGGGTAGTTAGAGTAAAAACTTGNTTACCAT cohl_al2.seq
3451 TATTCCAATACCTTGAAATGGGGTAGTTAGAGTAAAAACTTGGTTACCAT a909_al2.seq

CAACATCTTTAGCTTGTGCTACTTGGTAAACAAGTAAATTACCGCCAGCG Majority
          3510      3520      3530      3540      3550
3501 CAACATCTTTAGCTTGTGCTACTTGGTAAACAAGTAAATTACCGCCAGCG cohl_al2.seq
3501 CAACATCTTTAGCTTGTGCTACTTGGTAAACAAGTAAATTACCGCCAGCG a909_al2.seq

ATACCTTGATTATTATACTTATTTTGTATAGTAATAGAACCCGTTTTCAT Majority
          3560      3570      3580      3590      3600
3551 ATACCTTGATTATTATACTTATTTTGTATAGTAATAGAACCCGTTTTCAT cohl_al2.seq
3551 ATACCTTGATTATTATACTTATTTTGTATAGTAATAGAACCCGTTTTCAT a909_al2.seq

CTGATCATTGGTATCAGCAGACACAAGTTGAGTACTTAGACTAAATAATA Majority
          3610      3620      3630      3640      3650
3601 CTGATCATTGGTATCAGCAGACACAAGTTGAGTACTTAGACTAAATAATA cohl_al2.seq
3601 CTGATCATTGGTATCAGCAGACACAAGTTGAGTACTTAGACTAAATAATA a909_al2.seq

AGAGAAGAGTTATCTTTAGGATCTTTTTTATAAATCATTGTTCTCTTCCTT Majority
          3660      3670      3680      3690      3700
3651 AGAGAAGAGTTATCTTTAGGATCTTTTTTATAAATCATTGTTCTCTTCCTT cohl_al2.seq
3651 AGAGAAGAGTTATNTTTAGGATCTTTTTTATAAATCATTGTTCTCTTCCTT a909_al2.seq

TCTCATTGCTTGTTTTAAAATTTTCTTACGTTGACGTGCTCTCCTAGTTA Majority
          3710      3720      3730      3740      3750
3701 TCTCATTGCTTGTTTTAAAATTTTCTTACGTTGACGTGCTCTCCTAGTTA cohl_al2.seq
3701 TCTCATTGCTTGTTTTAAAATTTTCTTACGTTGACGTGCTCTCCTAGTTA a909_al2.seq

CTTCTAAAGAGATTAAAAGTAAAATCAAAGTAAGGAAAATAGCGATAAAT Majority
          3760      3770      3780      3790      3800
3751 CTTCTAAAGAGATTAAAAGTAAAATCAAAGTAAGGAAAATAGCGATAAAT cohl_al2.seq
3751 CTTCTAAANAGATTAAAAGTAAAATCAAAGTAAGGAAAATAGCGATAAAT a909_al2.seq

GGTGCGATATAAATAGGCTCTATTTGTATTGCCTCTGCTACTACCAAAGC Majority
          3810      3820      3830      3840      3850
3801 GGTGCGATATAAATAGGCTCTATTTGTATTGCCTCTGCTACTACCAAAGC cohl_al2.seq
3801 GGTGCGATATAAATAGGCTCTATTTGTATTGCCTCTGCTACNACCAAAGC a909_al2.seq

GTTACCATTATCGTTTGGTACACGATGTCCTCTCACTAGTAACCGATGGG Majority
          3860      3870      3880      3890      3900
3851 GTTACCATTATCGTTTGGTACACGATGTCCTCTCACTAGTAACCGATGGG cohl_al2.seq
3851 GTTACCATTATCGTTTGGTACACGATGTCCTCTCACTAGTAACCGATGGG a909_al2.seq
```

FIGURE 21E

```
TATTAACGCCATATGGTGTACACGTCACCAAAGTTTGGTAGTCTTTACCT Majority
        3910      3920      3930      3940      3950
3901 TATTAACGCCATATGGTGTACACGTCACCAAAGTTTGGTAGTCTTTACCT cohl_al2.seq
3901 TATTAACACCATATGGTGTACACGTCACCAAAGTTTGGTAGTCTTTACCT a909_al2.seq

TTAACAATTTGTAAATCCCTCAAATCATCCGGTTTAACTGTTCTGATTTG Majority
        3960      3970      3980      3990      4000
3951 TTAACAATTTGTAAATCCCTCAAATCATCCGGTTTAACTGTTCTGATTTG cohl_al2.seq
3951 TTAACAATTTGTAAATCCCTCAAATCATCCGGTTTAACTGTTCTGATTTG a909_al2.seq

ATCCACTTGATAAGTATATGTTTCATTTAAGATACTGACTGTCCAGTGGT Majority
        4010      4020      4030      4040      4050
4001 ATCCACTTGATAAGTATATGTTTCATTTAAGATACTGACTGTCCAGTGGT cohl_al2.seq
4001 ATCCACTTGATAAGTATATGTTTCATTTAAGATACTGACTGTCCAGTGGT a909_al2.seq

CTCCAGCTTTTAACTTATCCAAATCAGAAAAAAGCCTTGAAGAGGGTAAA Majority
        4060      4070      4080      4090      4100
4051 CTCCAACTTTTAACTTATCCAAATCAGAAAAAAGCCTTGAAGAGGGTAAA cohl_al2.seq
4051 CTCCAGCTTTTAACTTATCCAAATCAGAAAAAAGCCTTGAAGAGGGTAAA a909_al2.seq

CCTCTATGTCCTGATAAAATAGAATGAGTTGAGTCTCCTCCAATTGGAAG Majority
        4110      4120      4130      4140      4150
4101 CCTCTATGTCCTGATAAAATAGAATGAGTTGAGTCTCCTCCAATTGGAAG cohl_al2.seq
4101 CCTCTATGTCCTGATAAAATAGAATGAGTTGAGTCTCCTCCAATTGGAAG a909_al2.seq

ACTACTTCCTTCTAAATGACCAATAGAAGTTTGAAGCACTTTTTCACTTG Majority
        4160      4170      4180      4190      4200
4151 ACTACTTCCTTCTAAATGACCAATAGAAGTTTGAAGCACTTTTTCACTTG cohl_al2.seq
4151 ACTACTTCCTTCTAAATGACCAATAGAAGTTTGAAGCACCTTTTCACTTG a909_al2.seq

TACCATGATAAAGTGGTAATTTTATGTTTATCTTTGGAATTGAAATATAA Majority
        4210      4220      4230      4240      4250
4201 TACCATGATAAAGTGGTAATTTTATGTTTATCTTTGGAATTGAAATATAA cohl_al2.seq
4201 TACCATGATAAAGTGGTAATTTTATATTTATCTTTGGAATTGAAATATAA a909_al2.seq

CCCATATTACCCGTTTTATCGATAGCCAGTTGTGAATTATAATCCAAACG Majority
        4260      4270      4280      4290      4300
4251 CCCATATTACCCGTTTTATCGATAGCCAGTTGTGAATTATAATCCAAACG cohl_al2.seq
4251 CCCATATTACCCGTTTTATCGATAGCCAGTTGTGAATTATAATCCAAACG a909_al2.seq

CTCTTGGTTAGTCATGTGCCACTTCATTCCTGAAGTTTTAAATTGCTTAT Majority
        4310      4320      4330      4340      4350
4301 CTCTTGGCTAGTCATGTGCCACTTCATTCCTGAAGTTTTAAATTGCTTAT cohl_al2.seq
4301 CTCTTGGTCAGTCATGTGCCACTTCATTCCTGAAGTTTTAAATTGCTTAT a909_al2.seq

TATATTCTTTGGCTCGGTTAATAATTTTTTTATAGTCGTTTTCATCCATA Majority
        4360      4370      4380      4390      4400
4351 TATATTCTTTGGCTCGGTTAATAATTTTTTTATAATCGTTTTCATCCATA cohl_al2.seq
4351 TATATTCTTTGGCTCGGTTAATAATTTTTTTATAGTCGTTTTCATCCATA a909_al2.seq

TGCGTTACGCGGTCTTGGTAATCCATAATCGCTCGAGATTGGTGAAATGA Majority
        4410      4420      4430      4440      4450
4401 TGCGTTACGCGGTCTTGGTAATCCATAATCGCTCGAGATTGGTGAAATGA cohl_al2.seq
4401 TGCGTTACGCGGTCTTGGTAATCCATAATCGCTCGAGATTGGTGAAATGA a909_al2.seq

ATTCCAATAATTAGCAAGTGAAGGATAAGCCATTAAGCCTACCCCCACTG Majority
        4460      4470      4480      4490      4500
4451 ATTCCAATAATTAGCAAGTGAAGGATAAGCCATTAAGCCTACCCCCACTG cohl_al2.seq
4451 ATTCCAATAATTAGCAAGTGAAGGATAAGCCATTAAGCCTACCCCCACTG a909_al2.seq

CAATTATAGTGACAAGCAAAATGGATACTAAATGTTGTCTTATTTTTTTTC Majority
        4510      4520      4530      4540      4550
4501 CAATTATAGTGACAAGCAAAATGGATACTAAATGTTGTCTTATTTTTTTTC cohl_al2.seq
4501 CAATTATAGTGACAAGCAAAATGGATACTAAATGTTGTCTTATTTTTTTTC a909_al2.seq
```

FIGURE 21F

```
ATATATTTAAATCTGTACCACTTTGCTAGCCCATCTTATTAAGAACGTA Majority
        4560      4570      4580      4590      4600
4551 ATATATTTAAATCTGTACCACTTTGCTAGCCCATCTTATTAACAACGTA cohl_al2.seq
4551 ATATATTTAAATCTGTACCACTTTGCTAGCCCATCTTATTAAGAACGTA a909_al2.seq

AACGACGACGAGCAACAAGCACGATACCTGCTCCTATTACTAAAATTGCA Majority
        4610      4620      4630      4640      4650
4601 AACGACGACGAGCAACAAGCACGATACCTGCTCCTATTACTAAAATTGCA cohl_al2.seq
4601 AACGACGACGAGCAACAAGCACGATACCTGCTCCTATTACTAAAATTGCA a909_al2.seq

CCTATAATGTAGAAAATTGTTGTACCAATACCACCTGTTGAAGGCAACTC Majority
        4660      4670      4680      4690      4700
4651 CCTATAATGTAGAAAATTGTTGTACCAATACCACCTGTTGAAGGCAACTC cohl_al2.seq
4651 CCTATAATGTAGAAAATTGTTGTACCAATACCACCTGTTGAAGGCAACTC a909_al2.seq

AGTACCTTTGTTATTTTCAACAGTTGGGTTAACTAAAAGGTTATCTGAAT Majority
        4710      4720      4730      4740      4750
4701 AGTACCTTTGTTATTTTCAACAGTTGGGTTAACTAAAAGGTTATCTGAAT cohl_al2.seq
4701 AGTACCTTTGTTATTTTCAACAGTTGGGTTAACTAAAAGGTTATCTGAAT a909_al2.seq

TAGTCGTATCAGTGGCTCCATCTCCTAAAATAACCTTCTGAGAGTTATCT Majority
        4760      4770      4780      4790      4800
4751 TAGTCGTATCAGTGGCTCCATCTCCTAAAATAACCTTCTGAGAGTTATCT cohl_al2.seq
4751 TAGTCGTATCAGTGGCTCCATCTCCTAAAATAACCTTCTGAGAGTTATCT a909_al2.seq

AACAAATTGTAACCTAAGGGAGCCTTTTTCTCAACTAGATAGTATGTACC Majority
        4810      4820      4830      4840      4850
4801 AACAAATTGTAACCTAAGGGAGCCTTTTTCTCAACTAGATAGTATGTACC cohl_al2.seq
4801 AACAAATTGTAACCTAAGGGAGCCTTTTTCTCAACTAGATAGTATGTACC a909_al2.seq

TTCTTTCAAGCCTGTAATGGTAATTATACCATCTGCTCCTGTTGTATATT Majority
        4860      4870      4880      4890      4900
4851 TTCTTTCAAGCCTGTAATGGTAATTATACCATCTGCTCCTGTTGTATATT cohl_al2.seq
4851 TTCTTTCAAGCCTGTAATGGTAATTATACCATCTGCTCCTGTTGTATATT a909_al2.seq

CTGTTGCATTAGCTTCTGTGCCCCATTCAACGTTATTTGTATCGTTAAAG Majority
        4910      4920      4930      4940      4950
4901 CTGTTGCATTAGCTTCTGTGCCCCATTCAACGTTATTTGTATCGTTAAAG cohl_al2.seq
4901 CTGTTGCATTAGCTTCTGTGCCCCATTCAACGTTATTTGTATCGTTAAAG a909_al2.seq

TTTAGAAATTGACCCGTAGCATTCTTTAAAACAAATATAGCACCTTGTAA Majority
        4960      4970      4980      4990      5000
4951 TTTAGAAATTGACCCGTAGCATTCTTTAAAACAAATATAGCACCTTGTAA cohl_al2.seq
4951 TTTAGAAATTGACCCGTAGCATTCTTTAAAACAAATATAGCACCTTGTAA a909_al2.seq

TGAAGCTTTTGTGGAACCATCAATTTTTTTTATAGTAATTTGACCATCCC Majority
        5010      5020      5030      5040      5050
5001 TGAAGCTTTTGTGGAACCATCAATTTTTTTTATAGTAATTTGACCATCCC cohl_al2.seq
5001 TGAAGCTTTTGTGGAACCATCAATTTTTTTTATAGTAATTTGACCATCCC a909_al2.seq

TCACTGTTACTTTTTGACCTGGGTCATCATTGCTAGTATTGGGGTTGATG Majority
        5060      5070      5080      5090      5100
5051 TCACTGTTACTTTTTGACCTGGGTCATCATTGCTAGTATTGGGGTTGATG cohl_al2.seq
5051 TCACTGTTACTTTTTGACCTGGGTCATCATTGCTAGTATTGGGGTTGATG a909_al2.seq

GTCGCAATGTTTGTATTTTCTGGTAAATCAGCTGAACCTGGTTTAGCTCC Majority
        5110      5120      5130      5140      5150
5101 GTCGCAATGTTTGTATTTTCTGGTAAATCAGCTGAACCTGGTTTAGCTCC cohl_al2.seq
5101 GTCGCAATGTTTGTATTTTCTGGTAAATCAGCTGAACCTGGTTTAGCTCC a909_al2.seq

ACTCTTTAATACTCCTGTATAAGTGACTGTGATTGTATTTATTCCCTTAT Majority
        5160      5170      5180      5190      5200
5151 ACTCTTTAATACTCCTGTATAAGTGACTGTGATTGTATTTATTCCCTTAT cohl_al2.seq
5151 ACTCTTTAATACTCCTGTATAAGTGACTGTGATTGTATTTATTCCCTTAT a909_al2.seq
```

FIGURE 21G

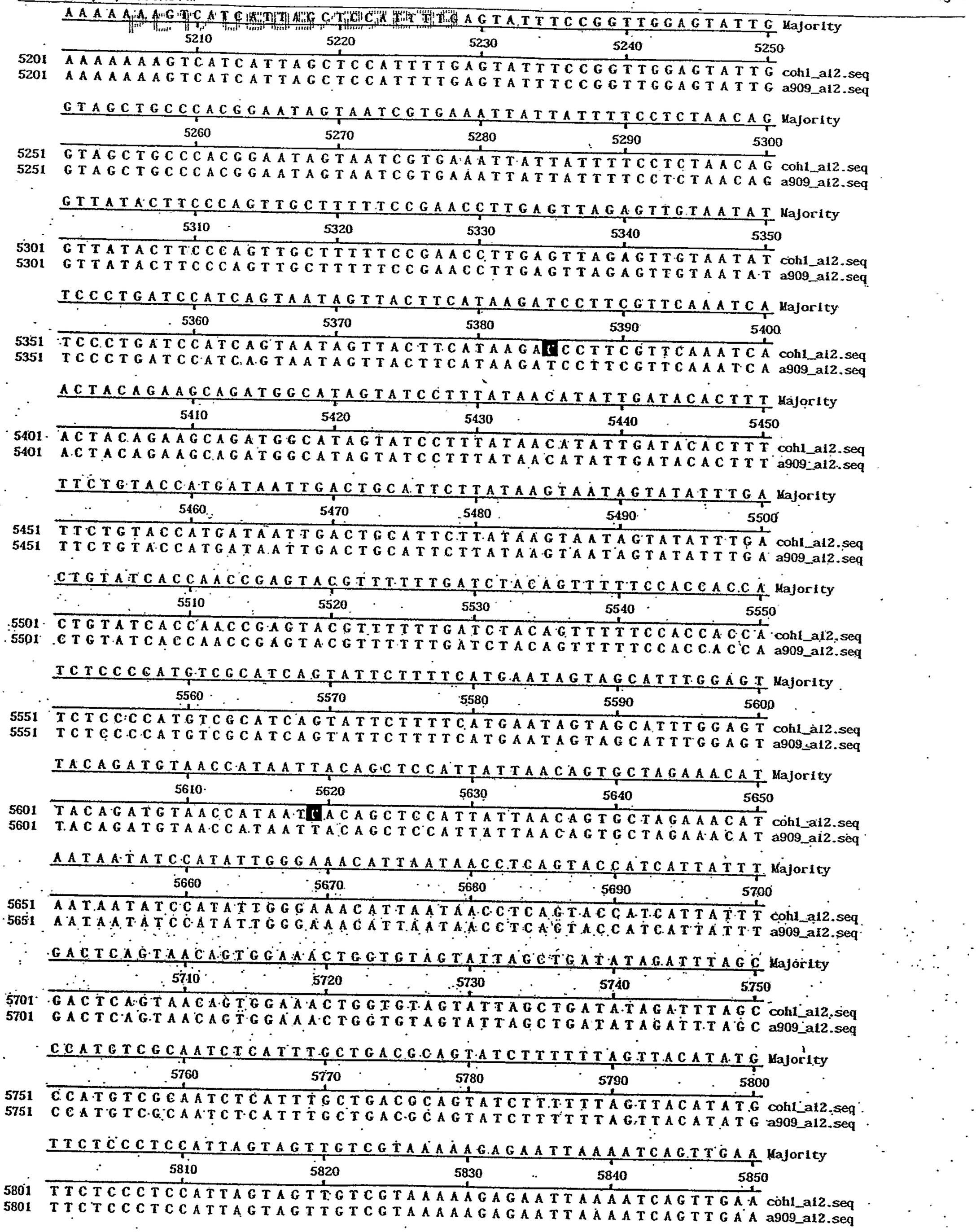

AAAAAAAGTCATCATTAGCTCCATTTTGAGTATTTCCGGTTGGAGTATTG Majority
5210 5220 5230 5240 5250
5201 AAAAAAAGTCATCATTAGCTCCATTTTGAGTATTTCCGGTTGGAGTATTG cohl_al2.seq
5201 AAAAAAAGTCATCATTAGCTCCATTTTGAGTATTTCCGGTTGGAGTATTG a909_al2.seq
GTAGCTGCCCACGGAATAGTAATCGTGAAATTATTATTTTCCTCTAACAG Majority
5260 5270 5280 5290 5300
5251 GTAGCTGCCCACGGAATAGTAATCGTGAAATTATTATTTTCCTCTAACAG cohl_al2.seq
5251 GTAGCTGCCCACGGAATAGTAATCGTGAAATTATTATTTTCCTCTAACAG a909_al2.seq
GTTATACTTCCCAGTTGCTTTTTCCGAACCTTGAGTTAGAGTTGTAATAT Majority
5310 5320 5330 5340 5350
5301 GTTATACTTCCCAGTTGCTTTTTCCGAACCTTGAGTTAGAGTTGTAATAT cohl_al2.seq
5301 GTTATACTTCCCAGTTGCTTTTTCCGAACCTTGAGTTAGAGTTGTAATAT a909_al2.seq
TCCCTGATCCATCAGTAATAGTTACTTCATAAGATCCTTCGTTCAAATCA Majority
5360 5370 5380 5390 5400
5351 TCCCTGATCCATCAGTAATAGTTACTTCATAAGATCCTTCGTTCAAATCA cohl_al2.seq
5351 TCCCTGATCCATCAGTAATAGTTACTTCATAAGATCCTTCGTTCAAATCA a909_al2.seq
ACTACAGAAGCAGATGGCATAGTATCCTTTATAACATATTGATACACTTT Majority
5410 5420 5430 5440 5450
5401 ACTACAGAAGCAGATGGCATAGTATCCTTTATAACATATTGATACACTTT cohl_al2.seq
5401 ACTACAGAAGCAGATGGCATAGTATCCTTTATAACATATTGATACACTTT a909_al2.seq
TTCTGTACCATGATAATTGACTGCATTCTTATAAGTAATAGTATATTTGA Majority
5460 5470 5480 5490 5500
5451 TTCTGTACCATGATAATTGACTGCATTCTTATAAGTAATAGTATATTTGA cohl_al2.seq
5451 TTCTGTACCATGATAATTGACTGCATTCTTATAAGTAATAGTATATTTGA a909_al2.seq
CTGTATCACCAACCGAGTACGTTTTTTGATCTACAGTTTTTCCACCACCA Majority
5510 5520 5530 5540 5550
5501 CTGTATCACCAACCGAGTACGTTTTTTGATCTACAGTTTTTCCACCACCA cohl_al2.seq
5501 CTGTATCACCAACCGAGTACGTTTTTTGATCTACAGTTTTTCCACCACCA a909_al2.seq
TCTCCCCATGTCGCATCAGTATTCTTTTCATGAATAGTAGCATTTGGAGT Majority
5560 5570 5580 5590 5600
5551 TCTCCCCATGTCGCATCAGTATTCTTTTCATGAATAGTAGCATTTGGAGT cohl_al2.seq
5551 TCTCCCCATGTCGCATCAGTATTCTTTTCATGAATAGTAGCATTTGGAGT a909_al2.seq
TACAGATGTAACCATAATTACAGCTCCATTATTAACAGTGCTAGAAACAT Majority
5610 5620 5630 5640 5650
5601 TACAGATGTAACCATAATCACAGCTCCATTATTAACAGTGCTAGAAACAT cohl_al2.seq
5601 TACAGATGTAACCATAATTACAGCTCCATTATTAACAGTGCTAGAAACAT a909_al2.seq
AATAATATCCATATTGGGAAACATTAATAACCTCAGTACCATCATTATTT Majority
5660 5670 5680 5690 5700
5651 AATAATATCCATATTGGGAAACATTAATAACCTCAGTACCATCATTATTT cohl_al2.seq
5651 AATAATATCCATATTGGGAAACATTAATAACCTCAGTACCATCATTATTT a909_al2.seq
GACTCAGTAACAGTGGAAACTGGTGTAGTATTAGCTGATATAGATTTAGC Majority
5710 5720 5730 5740 5750
5701 GACTCAGTAACAGTGGAAACTGGTGTAGTATTAGCTGATATAGATTTAGC cohl_al2.seq
5701 GACTCAGTAACAGTGGAAACTGGTGTAGTATTAGCTGATATAGATTTAGC a909_al2.seq
CCATGTCGCAATCTCATTTGCTGACGCAGTATCTTTTTTAGTTACATATG Majority
5760 5770 5780 5790 5800
5751 CCATGTCGCAATCTCATTTGCTGACGCAGTATCTTTTTTAGTTACATATG cohl_al2.seq
5751 CCATGTCGCAATCTCATTTGCTGACGCAGTATCTTTTTTAGTTACATATG a909_al2.seq
TTCTCCCTCCATTAGTAGTTGTCGTAAAAAGAGAATTAAAATCAGTTGAA Majority
5810 5820 5830 5840 5850
5801 TTCTCCCTCCATTAGTAGTTGTCGTAAAAAGAGAATTAAAATCAGTTGAA cohl_al2.seq
5801 TTCTCCCTCCATTAGTAGTTGTCGTAAAAAGAGAATTAAAATCAGTTGAA a909_al2.seq

FIGURE 21H

```
GCTTTATACTCAGCTTCTTTACCTTGAGGAATTAAATAAGAAGCTCCATC Majority
          5860      5870      5880      5890      5900
5851 GCTTTATACTCAGCTTCTTTACCTTGAGGAATTAAATAAGAAGCTCCATC cohl_al2.seq
5851 GCTTTATACTCAGCTTCTTTACCTTGAGGAATTAAATAAGAAGCTCCATC a909_al2.seq

TTTATTCGAATCAGATACATTTGCATTATCTATTTCTGCATCAAAAACTT Majority
          5910      5920      5930      5940      5950
5901 TTTATTCGAATCAGATACATTTGCATTATCTATTTCTGCATCAAAAACTT cohl_al2.seq
5901 TTTATTCGAATCAGATACATTTGCATTATCTATTTCTGCATCAAAAACTT a909_al2.seq

TGTATGCTTTATAGGTTGCGCCTTTTTGAGTATCTTGAACTGTAATTGTC Majority
          5960      5970      5980      5990      6000
5951 TATATGCTTTATAGGTTGCGCCTTTTTGAGTATCTTGAACTGTAATTGTC cohl_al2.seq
5951 TGTATGCTTTATAGGTTGCGCCTTTTTGAGTATCTTGAACTGTAATTGTC a909_al2.seq

CCTGTCTCAGCGGCAAAAGCTATCGGCGTAACTGGTGATACAGCCATACC Majority
          6010      6020      6030      6040      6050
6001 CCTGTCTCAGCGGCAAAAGCTATCGGCGTAACTGGTGATACAGCCATACC cohl_al2.seq
6001 CCTGTCTCAGCGGCAAAAGCTATCGGCGTAACTGGTGATACAGCCATACC a909_al2.seq

AAATGCTAAACTCGCCACTAACAGCGATTGAATCATTTTCTTTTTTCATTG Majority
          6060      6070      6080      6090      6100
6051 AAATGCTAAACTCGCCACTAACAGCGATTGAATCATTTTCTTTTTTCATTG cohl_al2.seq
6051 AAATGCTAAACTCGCCACTAACAGCGATTGAATCATTTTCTTTTTTCATTG a909_al2.seq

AAATCTTTCTCCTAAAATCATATTGATGAATGATTAATTCATATTTTTTTT Majority
          6110      6120      6130      6140      6150
6101 AAATCTTTCTCCTAAAATCATATTGATGAATGATTAATTCATATTTTTTT cohl_al2.seq
6101 AAATCTTTCTCCTAAAATCATATTGATGAATGATTAATTCATATTTTTTT a909_al2.seq

TCGATAGTATAATATTAATCCTGATGGTAGAGCTAAAGCTAAACCAACTA Majority
          6160      6170      6180      6190      6200
6151 TCGATAGTATAATATTAATCCTGATGGTAGAGCTAAAGCTAAACCAACTA cohl_al2.seq
6151 TCGATAGTATAATATTAATCCTGATGGTAGAGCTAAAGCTAAACCAACTA a909_al2.seq

GGATATAAATGTGTGTTCCAATACCTCCAGTACTAGGCAATTCTGTTCCT Majority
          6210      6220      6230      6240      6250
6201 GGATATAAATGTGTGTTCCAATACCTCCAGTACTAGGCAATTCTGTTCCT cohl_al2.seq
6201 GGATATAAATGTGTGTTCCAATACCTCCAGTACTAGGCAATTCTGTTCCT a909_al2.seq

TTACTGTTAGTAATTTTAAAAGTATATACTGTACTTCCATCTACTAAATT Majority
          6260      6270      6280      6290      6300
6251 TTACTGTTAGTAATTTTAAAAGTATATACTGTACTTCCATCTACTAAATT cohl_al2.seq
6251 TTACTGTTAGTAATTTTAAAAGTATATACTGTACTTCCATCTACTAAATT a909_al2.seq

CTCTTTTATTGGTGTCGCATTATTACCATTTTGTTCAAAGGTAACTCCCG Majority
          6310      6320      6330      6340      6350
6301 CTCTTTTATTGGTGTCGCATTATTACCATTTTGTTCAAAGGTAACTCCCG cohl_al2.seq
6301 CTCTTTTATTGGTGTCGCATTATTACCATTTTGTTCAAAGGTAACTCCCG a909_al2.seq

TAGAAATCACTAATACTGATATATCATTTTTAGGTAGTAGGTACCCTGGA Majority
          6360      6370      6380      6390      6400
6351 TAGAAATCACTAATACTGATATATCATTTTTAGGTAGTAGGTACCCTGGA cohl_al2.seq
6351 TAGAAATCACTAATACTGATATATCATTTTTAGGTAGTAGGTACCCTGGA a909_al2.seq

GGGGCCTTTGTCTCTGTTAGGTAGTATTTTCCTACTGGCAAACTGAGGTA Majority
          6410      6420      6430      6440      6450
6401 GGGGCCTTTGTCTCTGTTAGGTAGTATTTTCCTACTGGCAAACTGAGGTA cohl_al2.seq
6401 GGGGCCTTTGTCTCTGTTAGGTAGTATTTTCCTACTGGCAAACTGAGGTA a909_al2.seq

GTTATTAGCATCCACTAATAACAAGCCTTTATCGTTTGTCACCAGCCCTG Majority
          6460      6470      6480      6490      6500
6451 GTTATTAGCATCCACTAATAACAAGCCTTTATCGTTTGTCACCAGCCCTG cohl_al2.seq
6451 GTTATTAGCATCCACTAATAACAAGCCTTTATCGTTTGTCACCAGCCCTG a909_al2.seq
```

FIGURE 21I

```
                                                
```

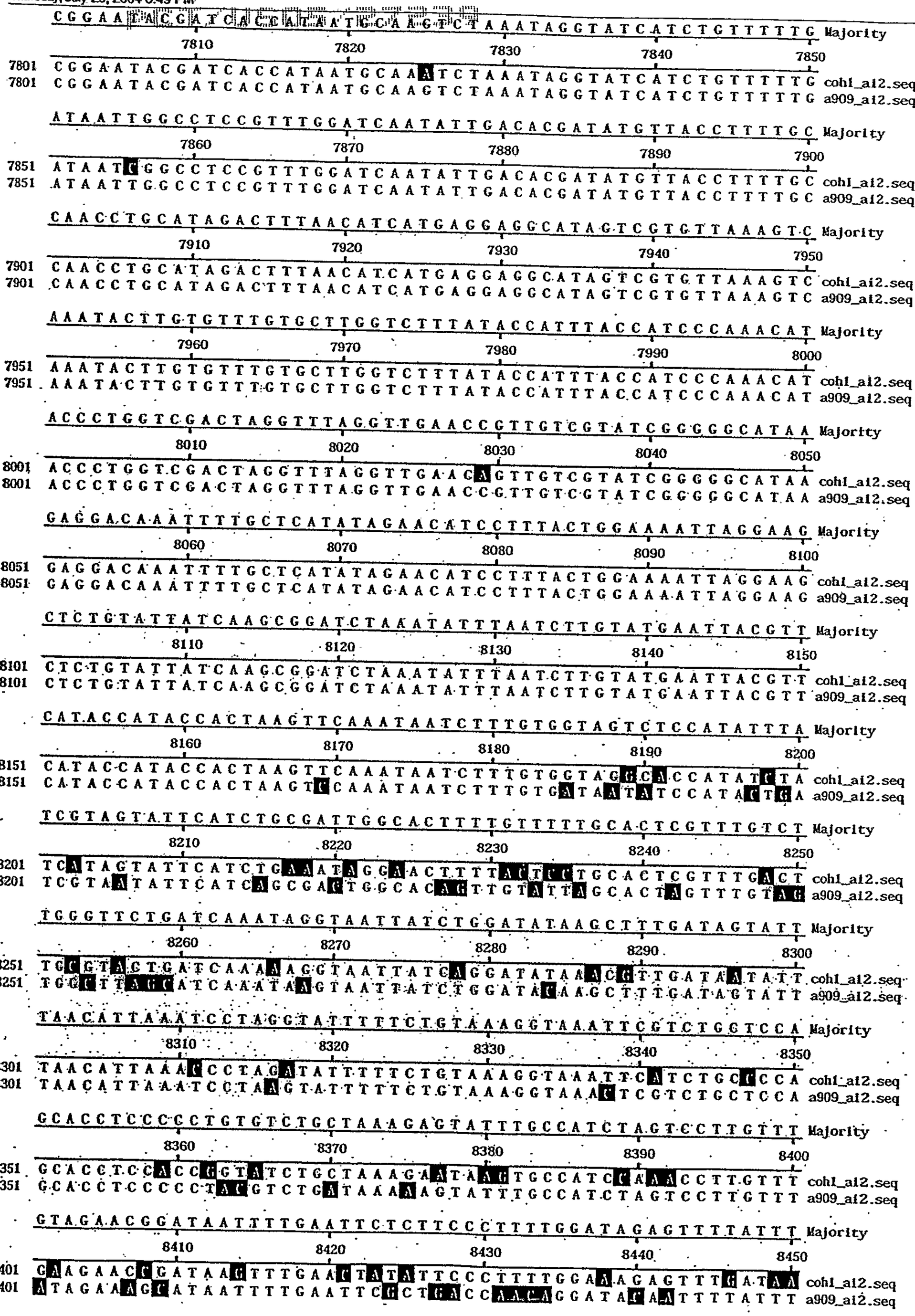

CGGAATACGATCACCATAATGCAAGTCTAAATAGGTATCATCTGTTTTTG Majority
7810 7820 7830 7840 7850
7801 CGGAATACGATCACCATAATGCAAATCTAAATAGGTATCATCTGTTTTTG cohl_ai2.seq
7801 CGGAATACGATCACCATAATGCAAGTCTAAATAGGTATCATCTGTTTTTG a909_ai2.seq
ATAATTGGCCTCCGTTTGGATCAATATTGACACGATATGTTACCTTTTGC Majority
7860 7870 7880 7890 7900
7851 ATAATGGGCCTCCGTTTGGATCAATATTGACACGATATGTTACCTTTTGC cohl_ai2.seq
7851 ATAATTGGCCTCCGTTTGGATCAATATTGACACGATATGTTACCTTTTGC a909_ai2.seq
CAACCTGCATAGACTTTAACATCATGAGGAGGCATAGTCGTGTTAAAGTC Majority
7910 7920 7930 7940 7950
7901 CAACCTGCATAGACTTTAACATCATGAGGAGGCATAGTCGTGTTAAAGTC cohl_ai2.seq
7901 CAACCTGCATAGACTTTAACATCATGAGGAGGCATAGTCGTGTTAAAGTC a909_ai2.seq
AAATACTTGTGTTTGTGCTTGGTCTTTATACCATTTACCATCCCAAACAT Majority
7960 7970 7980 7990 8000
7951 AAATACTTGTGTTTGTGCTTGGTCTTTATACCATTTACCATCCCAAACAT cohl_ai2.seq
7951 AAATACTTGTGTTTGTGCTTGGTCTTTATACCATTTACCATCCCAAACAT a909_ai2.seq
ACCCTGGTCGACTAGGTTTAGGTTGAACCGTTGTCGTATCGGGGGCATAA Majority
8010 8020 8030 8040 8050
8001 ACCCTGGTCGACTAGGTTTAGGTTGAACAGTTGTCGTATCGGGGGCATAA cohl_ai2.seq
8001 ACCCTGGTCGACTAGGTTTAGGTTGAACCGTTGTCGTATCGGGGGCATAA a909_ai2.seq
GAGGACAAATTTTGCTCATATAGAACATCCTTTACTGGAAAATTAGGAAG Majority
8060 8070 8080 8090 8100
8051 GAGGACAAATTTTGCTCATATAGAACATCCTTTACTGGAAAATTAGGAAG cohl_ai2.seq
8051 GAGGACAAATTTTGCTCATATAGAACATCCTTTACTGGAAAATTAGGAAG a909_ai2.seq
CTCTGTATTATCAAGCGGATCTAAATATTTAATCTTGTATGAATTACGTT Majority
8110 8120 8130 8140 8150
8101 CTCTGTATTATCAAGCGGATCTAAATATTTAATCTTGTATGAATTACGTT cohl_ai2.seq
8101 CTCTGTATTATCAAGCGGATCTAAATATTTAATCTTGTATGAATTACGTT a909_ai2.seq
CATACCATACCACTAAGTTCAAATAATCTTTGTGGTAGTCTCCATATTTA Majority
8160 8170 8180 8190 8200
8151 CATACCATACCACTAAGTTCAAATAATCTTTGTGGTAGGCACCATATCTA cohl_ai2.seq
8151 CATACCATACCACTAAGTCCAAATAATCTTTGTGATAATATCCATACTGA a909_ai2.seq
TCGTAGTATTCATCTGCGATTGGCACTTTTGTTTTTGCACTCGTTTGTCT Majority
8210 8220 8230 8240 8250
8201 TCATAGTATTCATCTGAAATAGGAACTTTTACTCCTGCACTCGTTTGACT cohl_ai2.seq
8201 TCGTAATATTCATCAGCGAGTGGCACAGTTGTATTAGCACTAGTTTGTAG a909_ai2.seq
TGGGTTCTGATCAAATAGGTAATTATCTGGATATAAGCTTTGATAGTATT Majority
8260 8270 8280 8290 8300
8251 TGCGTACTGATCAAAAAGGTAATTATCAGGATATAAACGTTGATAATATT cohl_ai2.seq
8251 TGGCTTAGCATCAAATAAGTAATTATCTGGATAGAAGCTTTGATAGTATT a909_ai2.seq
TAACATTAAATCCTAGGTATTTTTCTGTAAAGGTAAATTCGTCTGGTCCA Majority
8310 8320 8330 8340 8350
8301 TAACATTAAACCCTAGATATTTTTCTGTAAAGGTAAATTCATCTGCCCA cohl_ai2.seq
8301 TAACATTAAATCCTAGTATTTTTCTGTAAAGGTAAACTCGTCTGCTCCA a909_ai2.seq
GCACCTCCCCCTGTGTCTGCTAAAGAGTATTTGCCATCTAGTCCTTGTTT Majority
8360 8370 8380 8390 8400
8351 GCACCTCCACCGGTATCTGCTAAAGAATAAGTGCCATCCAAACCTTGTTT cohl_ai2.seq
8351 GCACCTCCCCCTACGTCTGATAAAAAGTATTTGCCATCTAGTCCTTGTTT a909_ai2.seq
GTAGAACGGATAATTTTGAATTCTCTTCCCTTTTGGATAGAGTTTTATTT Majority
8410 8420 8430 8440 8450
8401 GAAGAACCGATAAGTTTGAACTATATTCCCTTTTGGAAAGAGTTTGATAA cohl_ai2.seq
8401 ATAGAAAGCATAATTTTGAATTCGCTGACCA-A-AGGATAGAATTTTATTT a909_ai2.seq

FIGURE 21L

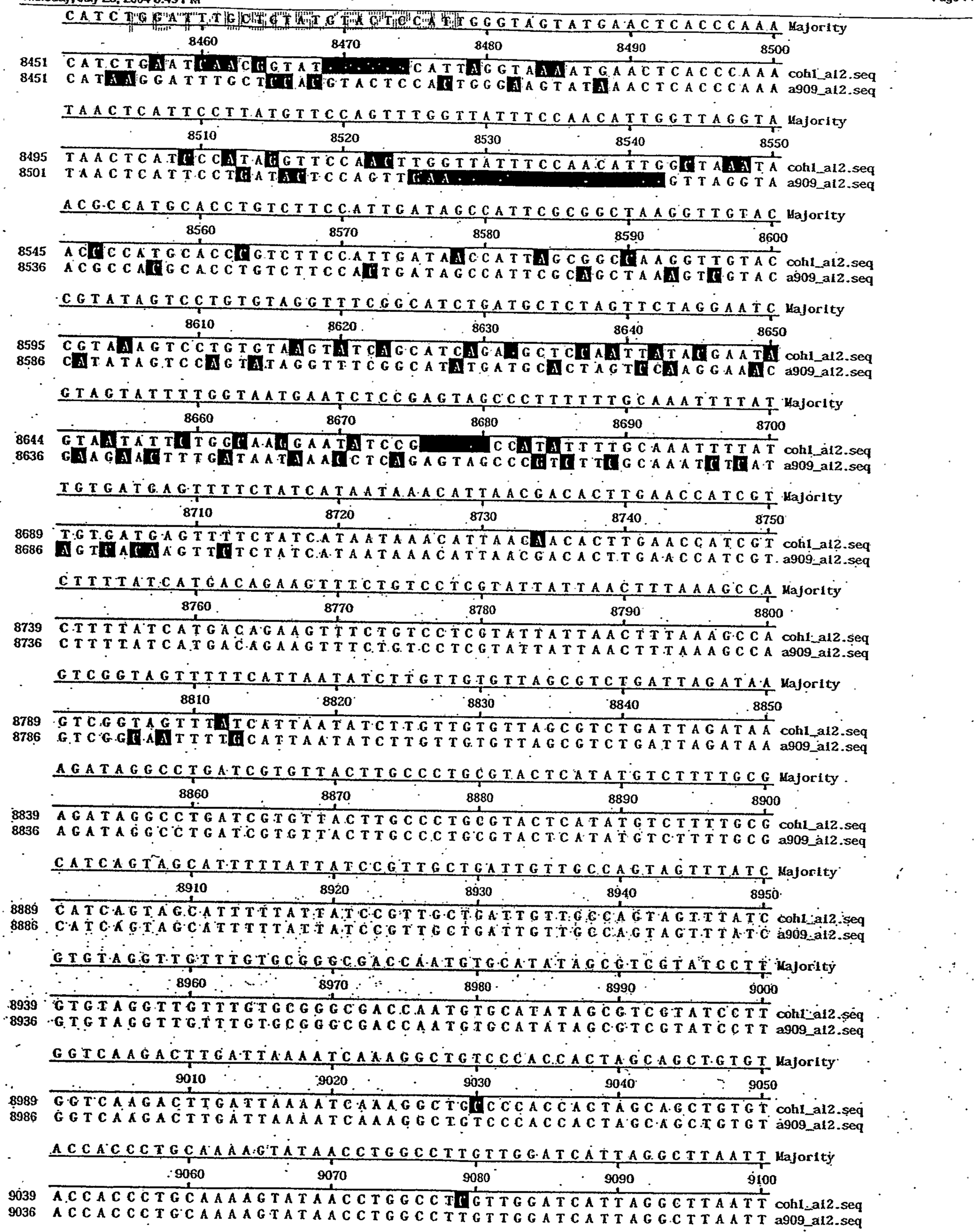

CATCTGGATTTGCTGTATGTACTCCATTGGGTAGTATGAACTCACCCAAA Majority
8460 8470 8480 8490 8500
8451 CATCTGAATCAACGGTAT.....CATTAGGTAAAATGAACTCACCCAAA cohl_al2.seq
8451 CATAAGGATTTGCTCCACGTACTCCACTGGGAGTATAAACTCACCCAAA a909_al2.seq
TAACTCATTCCTTATGTTCCAGTTTGGTTATTTCCAACATTGGTTAGGTA Majority
8510 8520 8530 8540 8550
8495 TAACTCATCCCATAGGTTCCAACTTGGTTATTTCCAACATTGGCTAAATA cohl_al2.seq
8501 TAACTCATTCCTGATACTCCAGTTGAA.....GTTAGGTA a909_al2.seq
ACGCCATGCACCTGTCTTCCATTGATAGCCATTCGCGGCTAAGGTTGTAC Majority
8560 8570 8580 8590 8600
8545 ACCCCATGCACCGGTCTTCCATTGATAACCATTAGCGGCCAAGGTTGTAC cohl_al2.seq
8536 ACGCCAGGCACCTGTCTTCCATTGATAGCCATTCGCAGCTAAAGTGGTAC a909_al2.seq
CGTATAGTCCTGTGTAGGTTTCGGCATCTGATGCTCTAGTTCTAGGAATC Majority
8610 8620 8630 8640 8650
8595 CGTAAAGTCCTGTGTAAGTATCAGCATCAGA-GCTCCAATTATACGAATA cohl_al2.seq
8586 CATATAGTCCAGTATAGGTTTCGGCATATGATGCACTAGTTCAAGGAAAC a909_al2.seq
GTAGTATTTTGGTAATGAATCTCCGAGTAGCCCTTTTTTGCAAATTTTAT Majority
8660 8670 8680 8690 8700
8644 GTAATATTTCTGGCAAGGAATATCCG.....CCATATTTTGCAAATTTTAT cohl_al2.seq
8636 GAAGAACTTTGATAATAAACCTCAGAGTAGCCCGTCTTTGCAAATCTTAT a909_al2.seq
TGTGATGAGTTTTCTATCATAATAAACATTAACGACACTTGAACCATCGT Majority
8710 8720 8730 8740 8750
8689 TGTGATGAGTTTTCTATCATAATAAACATTAACAACACTTGAACCATCGT cohl_al2.seq
8686 AGTTACAAGTTCTCTATCATAATAAACATTAACGACACTTGAACCATCGT. a909_al2.seq
CTTTTATCATGACAGAAGTTTCTGTCCTCGTATTATTAACTTTAAAGCCA Majority
8760 8770 8780 8790 8800
8739 CTTTTATCATGACAGAAGTTTCTGTCCTCGTATTATTAACTTTAAAGCCA cohl_al2.seq
8736 CTTTTATCATGACAGAAGTTTCTGTCCTCGTATTATTAACTTTAAAGCCA a909_al2.seq
GTCGGTAGTTTTTCATTAATATCTTGTTGTGTTAGCGTCTGATTAGATAA Majority
8810 8820 8830 8840 8850
8789 GTCGGTAGTTTATCATTAATATCTTGTTGTGTTAGCGTCTGATTAGATAA cohl_al2.seq
8786 GTCGGAATTTTGCATTAATATCTTGTTGTGTTAGCGTCTGATTAGATAA a909_al2.seq
AGATAGGCCTGATCGTGTTACTTGCCCTGCGTACTCATATGTCTTTTGCG Majority
8860 8870 8880 8890 8900
8839 AGATAGGCCTGATCGTGTTACTTGCCCTGCGTACTCATATGTCTTTTGCG cohl_al2.seq
8836 AGATAGGCCTGATCGTGTTACTTGCCCTGCGTACTCATATGTCTTTTGCG a909_al2.seq
CATCAGTAGCATTTTTATTATCCGTTGCTGATTGTTGCCAGTAGTTTATC Majority
8910 8920 8930 8940 8950
8889 CATCAGTAGCATTTTTATTATCCGTTGCTGATTGTTGCCAGTAGTTTATC cohl_al2.seq
8886 CATCAGTAGCATTTTTATTATCCGTTGCTGATTGTTGCCAGTAGTTTATC a909_al2.seq
GTGTAGGTTGTTTGTGCGGGCGACCAATGTGCATATAGCGTCGTATCCTT Majority
8960 8970 8980 8990 9000
8939 GTGTAGGTTGTTTGTGCGGGCGACCAATGTGCATATAGCGTCGTATCCTT cohl_al2.seq
8936 GTGTAGGTTGTTTGTGCGGGCGACCAATGTGCATATAGCGTCGTATCCTT a909_al2.seq
GGTCAAGACTTGATTAAAATCAAAGGCTGTCCCACCACTAGCAGCTGTGT Majority
9010 9020 9030 9040 9050
8989 GGTCAAGACTTGATTAAAATCAAAGGCTGCCCCACCACTAGCAGCTGTGT cohl_al2.seq
8986 GGTCAAGACTTGATTAAAATCAAAGGCTGTCCCACCACTAGCAGCTGTGT a909_al2.seq
ACCACCCTGCAAAAGTATAACCTGGCCTTGTTGGATCATTAGGCTTAATT Majority
9060 9070 9080 9090 9100
9039 ACCACCCTGCAAAAGTATAACCTGGCCTTGTTGGATCATTAGGCTTAATT cohl_al2.seq
9036 ACCACCCTGCAAAAGTATAACCTGGCCTTGTTGGATCATTAGGCTTAATT a909_al2.seq

FIGURE 21M

```
GTCGAAGCAGGTTGGTCTGTTAACACACGACGAGGTGCAATATAGGTAAC Majority
          9110      9120      9130      9140      9150
9089 GTCGAAGCAGGTTGGTCTGTTAACACACGACGAGGTGCAATATAGGTAAC cohl_al2.seq
9086 GTCGAAGCAGGTTGGTCTGTTAACACACGACGAGGTGCAATATAGGTAAC a909_al2.seq

TCCTGTTGATAAGTCGCCTGTGTTGAATTCAACACCGTACGATTCTTTAAA Majority
          9160      9170      9180      9190      9200
9139 TCCTGTTGATAAGTCGCCTGTGTTGAATTCAACACCGTACGATTCTTTAA cohl_al2.seq
9136 TCCTGTTGATAAGTCGCCTGTGTTGAATTCAACACCGTACGATTCTTTAA a909_al2.seq

AGACAGGATAAAGACTTATGTCTGAAGATACAGGTAATGCTTGAATTTCT Majority
          9210      9220      9230      9240      9250
9189 AGACAGGATAAAGACTTATGTCTGAAGATACAGGTAATGCTTGAATTTCT cohl_al2.seq
9186 AGACAGGATAAAGACTTATGTCTGAAGATACAGGTAATGCTTGAATTTCT a909_al2.seq

GATTCAGAAAGTGGAGCTCCATTTTGAGTTTTAGACCAACCTACAAATAA Majority
          9260      9270      9280      9290      9300
9239 GATTCAGAAAGTGGAGCTCCATTTTGAGTTTTAGACCAACCTACAAATAA cohl_al2.seq
9236 GATTCAGAAAGTGGAGCTCCATTTTGAGTTTTAGACCAACCTACAAATAA a909_al2:seq

TAATGTTGAAGTGGGTGGTGAAACTTTAAATGAACTTAGATCCGCGGTTC Majority
          9310      9320      9330      9340      9350
9289 GAATGTTGAAGTGGGTGGTGAAACTTTAAATGAACTTAAATCCGCGGTTC cohl_al2.seq
9286 TAATGTTGAAGTGGGTGGTGAAACTTTAAATGAACTTAGATCCGCGGTTC a909_al2.seq

CGTTTCCTGAACTATCCAATGGTACTTGATAACGCTCCAAAATACTTTTA Majority
          9360      9370      9380      9390      9400
9339 CGTTTCCTGAACTATCCAATGGTACTTGATAACGCTCCAAAATACTTTTA cohl_al2.seq
9336 CGTTTCCTGAACTATCCAATGGTACTTGATAACGCTCCAAAATACTTTTA a909_al2.seq

CCAGCAGAATCATCATATAAAGTTACTGTCGCCACTTTACTCATAAATGG Majority
          9410      9420      9430      9440      9450
9389 CCAGCAGAATCATCATATAAAGTTACTGTCGCCACTTTACTCATAAATGG cohl_al2.seq
9386 CCAGCAGAATCATCATATAAAGTTACTGTCGCCACTTTACTCATAAATGG a909_al2.seq

ACGAACATAAATTTCTTTTGTCTCAGTTACAGTTATTGGCTCACCAAATT Majority
          9460      9470      9480      9490      9500
9439 ACGAACATAAATTTCTTTTGTCTCAGTTACAGTTATTGGCTCACCAAATT cohl_al2.seq
9436 ACGAACATAAATTTCTTTTGTCTCAGTTACAGTTATTGGCTCACCAAATT a909_al2.seq

TAACAGGGTCACCATACTTTCCAGTAGTAGGATCATAGGTATACCAACCA Majority
          9510      9520      9530      9540      9550
9489 TAACAGGGTCACCATACTTTCCAGTAGTAGGATCATAGGTATACCAACCA cohl_al2.seq
9486 TAACAGGGTCACCATACTTTCCAGTAGTAGGATCATAGGTATACCAACCA a909_al2.seq

TTAAAATGCTCTCCTGCTTTAATCGTCGGAATCCCAACTTCTCCTAGAGA Majority
          9560      9570      9580      9590      9600
9539 TTAAAATGCTCTCCTGCTTTAATCGTCGGAATCCCAACTTCTCCTAGAGA cohl_al2.seq
9536 TTAAAATGCTCTCCTGCTTTAATCGTCGGAATCCCAACTTCTCCTAGAGA a909_al2.seq

TTCTCCATCTTTTATAATTTGATGATGAACTTGCATACCTGAAGCTGTCA Majority
          9610      9620      9630      9640      9650
9589 TTCTCCATCTTTTATAATTTGATGATGAACTTGCATACCTGAAGCTGTCA cohl_al2.seq
9586 TTCTCCATCTTTTATAATTTGATGATGAACTTGCATACCTGAAGCTGTCA a909_al2.seq

GGAAATTATAATCAGTTCCGTCATTATTTTGAAAATGGTAAGTTAACCTA Majority
          9660      9670      9680      9690      9700
9639 GGAAATTATAATCAGTTCCGTCATTATTTTGAAAATGGTAAGTTAACCTA cohl_al2.seq
9636 GGAAATTATAATCAGTTCCGTCATTATTTTGAAAATGGTAAGTTAACCTA a909_al2.seq

GGAACTTCTGTATTATCCTCTTGAACAATTGCATAAATGGAGAATGAATC Majority
          9710      9720      9730      9740      9750
9689 GGAACTTCTGTATTATCCTCTTGAACAATTGCATAAATGGAGAATGAATC cohl_al2.seq
9686 GGAACTTCTGTATTATCCTCTTGAACAATTGCATAAATGGAGAATGAATC a909_al2.seq
```

FIGURE 21N

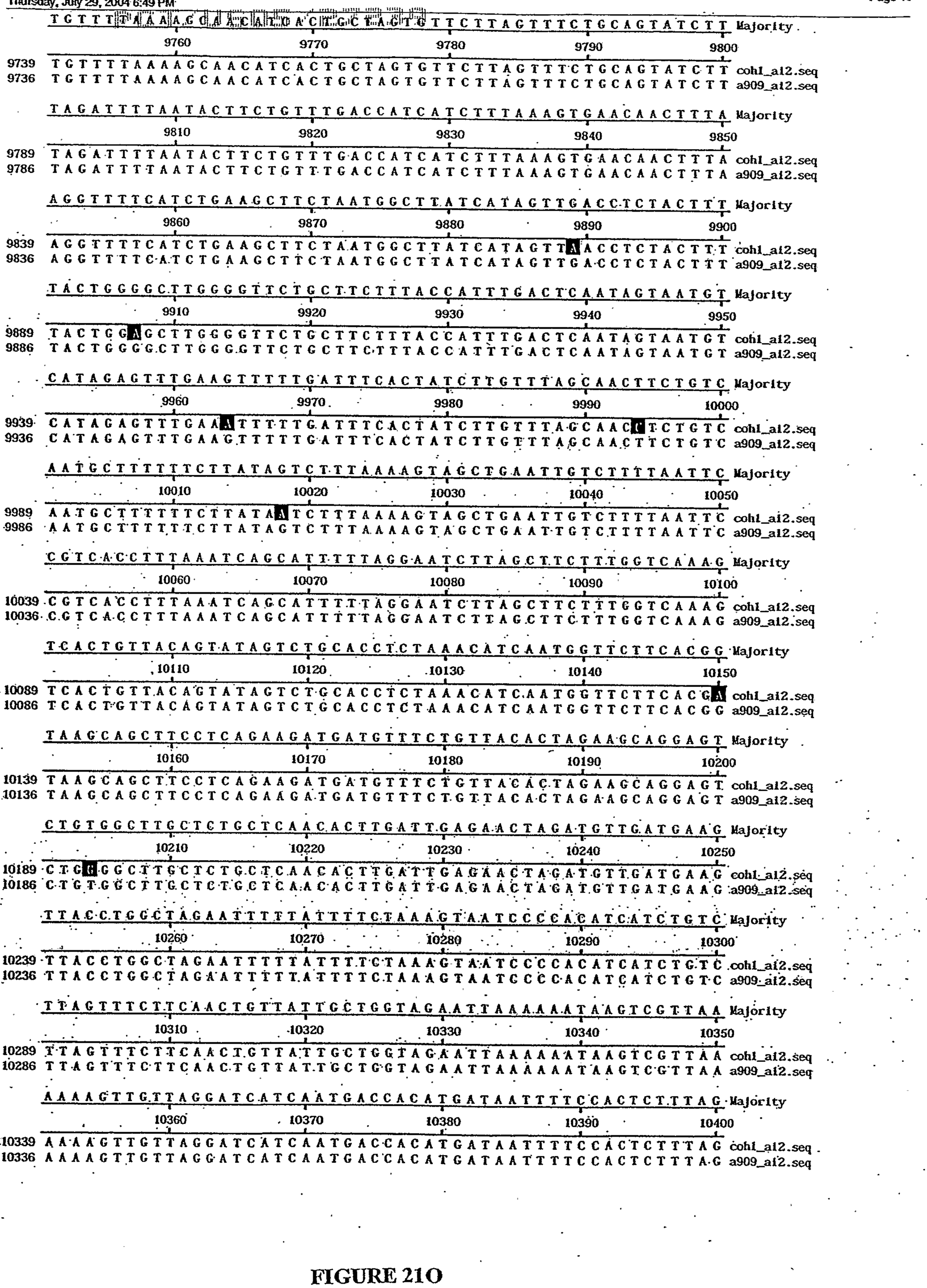

TGTTTTAAAAGCAACATCACTGCTAGTGTTCTTAGTTTCTGCAGTATCTT Majority
9760 9770 9780 9790 9800
9739 TGTTTTAAAAGCAACATCACTGCTAGTGTTCTTAGTTTCTGCAGTATCTT cohl_al2.seq
9736 TGTTTTAAAAGCAACATCACTGCTAGTGTTCTTAGTTTCTGCAGTATCTT a909_al2.seq
TAGATTTTAATACTTCTGTTTGACCATCATCTTTAAAGTGAACAACTTTA Majority
9810 9820 9830 9840 9850
9789 TAGATTTTAATACTTCTGTTTGACCATCATCTTTAAAGTGAACAACTTTA cohl_al2.seq
9786 TAGATTTTAATACTTCTGTTTGACCATCATCTTTAAAGTGAACAACTTTA a909_al2.seq
AGGTTTTCATCTGAAGCTTCTAATGGCTTATCATAGTTGACCTCTACTTT Majority
9860 9870 9880 9890 9900
9839 AGGTTTTCATCTGAAGCTTCTAATGGCTTATCATAGTTAACCTCTACTTT cohl_al2.seq
9836 AGGTTTTCATCTGAAGCTTCTAATGGCTTATCATAGTTGACCTCTACTTT a909_al2.seq
TACTGGGGCTTGGGGTTCTGCTTCTTTACCATTTGACTCAATAGTAATGT Majority
9910 9920 9930 9940 9950
9889 TACTGGAGCTTGGGGTTCTGCTTCTTTACCATTTGACTCAATAGTAATGT cohl_al2.seq
9886 TACTGGGGCTTGGGGTTCTGCTTCTTTACCATTTGACTCAATAGTAATGT a909_al2.seq
CATAGAGTTTGAAGTTTTTTGATTTCACTATCTTGTTTAGCAACTTCTGTC Majority
9960 9970 9980 9990 10000
9939 CATAGAGTTTGAAATTTTTTGATTTCACTATCTTGTTTAGCAACTTCTGTC cohl_al2.seq
9936 CATAGAGTTTGAAGTTTTTTGATTTCACTATCTTGTTTAGCAACTTCTGTC a909_al2.seq
AATGCTTTTTTTCTTATAGTCTTTAAAAGTAGCTGAATTGTCTTTTTAATTC Majority
10010 10020 10030 10040 10050
9989 AATGCTTTTTTTCTTATAATCTTTAAAAGTAGCTGAATTGTCTTTTTAATTC cohl_al2.seq
9986 AATGCTTTTTTTCTTATAGTCTTTAAAAGTAGCTGAATTGTCTTTTTAATTC a909_al2.seq
CGTCACCTTTAAATCAGCATTTTTAGGAATCTTAGCTTCTTTGGTCAAAG Majority
10060 10070 10080 10090 10100
10039 CGTCACCTTTAAATCAGCATTTTTAGGAATCTTAGCTTCTTTGGTCAAAG cohl_al2.seq
10036 CGTCACCTTTAAATCAGCATTTTTAGGAATCTTAGCTTCTTTGGTCAAAG a909_al2.seq
TCACTGTTACAGTATAGTCTGCACCTCTAAACATCAATGGTTCTTCACGG Majority
10110 10120 10130 10140 10150
10089 TCACTGTTACAGTATAGTCTGCACCTCTAAACATCAATGGTTCTTCACGA cohl_al2.seq
10086 TCACTGTTACAGTATAGTCTGCACCTCTAAACATCAATGGTTCTTCACGG a909_al2.seq
TAAGCAGCTTCCTCAGAAGATGATGTTTCTGTTACACTAGAAGCAGGAGT Majority
10160 10170 10180 10190 10200
10139 TAAGCAGCTTCCTCAGAAGATGATGTTTCTGTTACACTAGAAGCAGGAGT cohl_al2.seq
10136 TAAGCAGCTTCCTCAGAAGATGATGTTTCTGTTACACTAGAAGCAGGAGT a909_al2.seq
CTGTGGCTTGCTCTGCTCAACACTTGATTGAGAACTAGATGTTGATGAAG Majority
10210 10220 10230 10240 10250
10189 CTGGGGCTTGCTCTGCTCAACACTTGATTGAGAACTAGATGTTGATGAAG cohl_al2.seq
10186 CTGTGGCTTGCTCTGCTCAACACTTGATTGAGAACTAGATGTTGATGAAG a909_al2.seq
TTACCTGGCTAGAATTTTTATTTTCTAAAGTAATCCCCACATCATCTGTC Majority
10260 10270 10280 10290 10300
10239 TTACCTGGCTAGAATTTTTATTTTCTAAAGTAATCCCCACATCATCTGTC cohl_al2.seq
10236 TTACCTGGCTAGAATTTTTATTTTCTAAAGTAATCCCCACATCATCTGTC a909_al2.seq
TTAGTTTCTTCAACTGTTATTGCTGGTAGAATTAAAAAATAAGTCGTTAA Majority
10310 10320 10330 10340 10350
10289 TTAGTTTCTTCAACTGTTATTGCTGGTAGAATTAAAAAATAAGTCGTTAA cohl_al2.seq
10286 TTAGTTTCTTCAACTGTTATTGCTGGTAGAATTAAAAAATAAGTCGTTAA a909_al2.seq
AAAAGTTGTTAGGATCATCAATGACCACATGATAATTTTCCACTCTTTAG Majority
10360 10370 10380 10390 10400
10339 AAAAGTTGTTAGGATCATCAATGACCACATGATAATTTTCCACTCTTTAG cohl_al2.seq
10336 AAAAGTTGTTAGGATCATCAATGACCACATGATAATTTTCCACTCTTTAG a909_al2.seq

FIGURE 21O

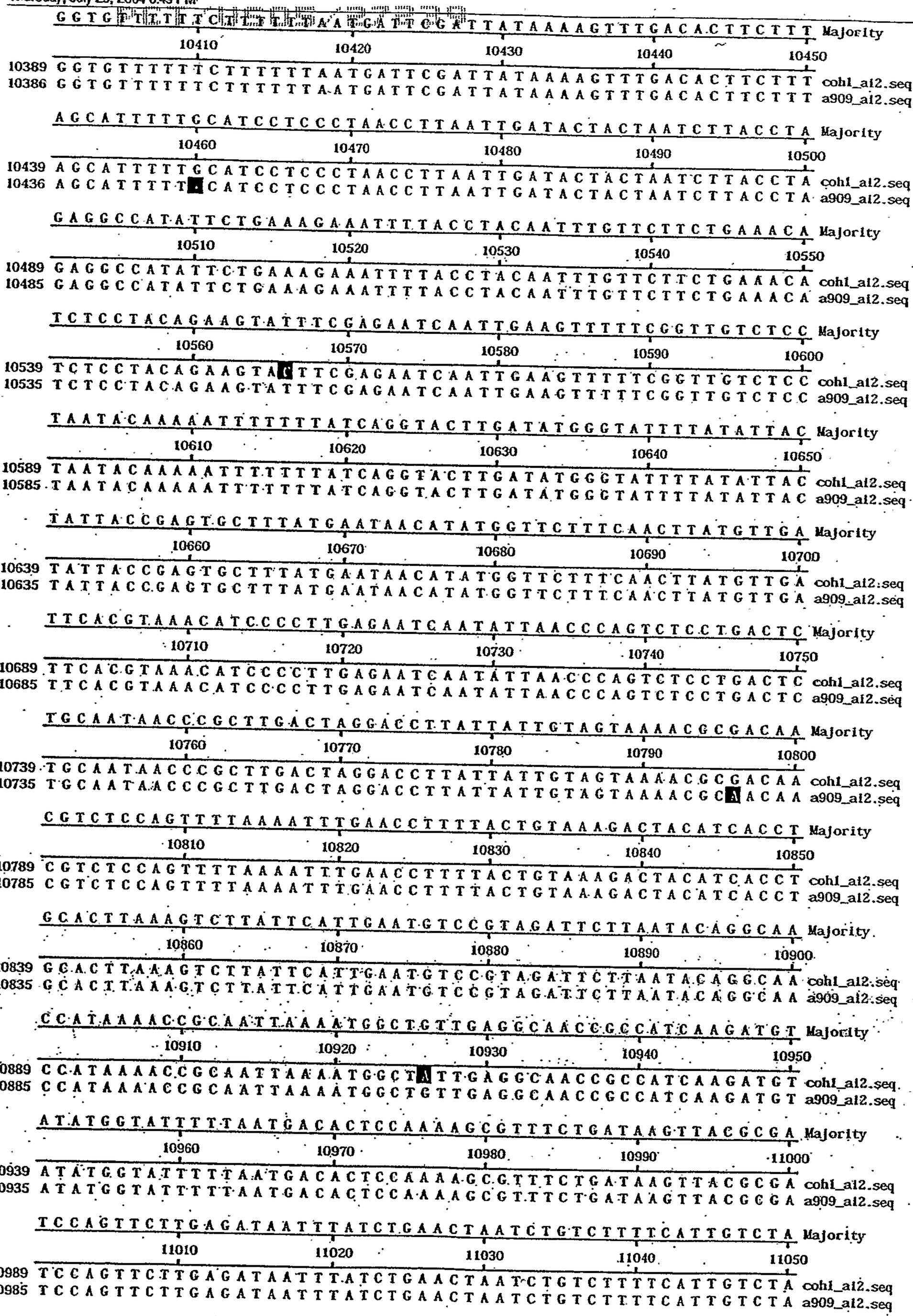

GGTGTTTTTTCTTTTTTAATGATTCGATTATAAAAGTTTGACACTTCTTT Majority
10410 10420 10430 10440 10450
10389 GGTGTTTTTTCTTTTTTAATGATTCGATTATAAAAGTTTGACACTTCTTT cohl_al2.seq
10386 GGTGTTTTTTCTTTTTTA-ATGATTCGATTATAAAAGTTTGACACTTCTTT a909_al2.seq
AGCATTTTTGCATCCTCCCTAACCTTAATTGATACTACTAATCTTACCTA Majority
10460 10470 10480 10490 10500
10439 AGCATTTTTGCATCCTCCCTAACCTTAATTGATACTACTAATCTTACCTA cohl_al2.seq
10436 AGCATTTTT.CATCCTCCCTAACCTTAATTGATACTACTAATCTTACCTA a909_al2.seq
GAGGCCATATTCTGAAAGAAATTTACCTACAATTTGTTCTTCTGAAACA Majority
10510 10520 10530 10540 10550
10489 GAGGCCATATTCTGAAAGAAATTTACCTACAATTTGTTCTTCTGAAACA cohl_al2.seq
10485 GAGGCCATATTCTGAAAGAAATTTACCTACAATTTGTTCTTCTGAAACA a909_al2.seq
TCTCCTACAGAAGTATTTCGAGAATCAATTGAAGTTTTTCGGTTGTCTCC Majority
10560 10570 10580 10590 10600
10539 TCTCCTACAGAAGTA.TTCGAGAATCAATTGAAGTTTTTCGGTTGTCTCC cohl_al2.seq
10535 TCTCCTACAGAAG.TATTTCGAGAATCAATTGAAGTTTTTCGGTTGTCTCC a909_al2.seq
TAATACAAAAATTTTTTTATCAGGTACTTGATATGGGTATTTTATATTAC Majority
10610 10620 10630 10640 10650
10589 TAATACAAAAATTTTTTTATCAGGTACTTGATATGGGTATTTTATATTAC cohl_al2.seq
10585 TAATACAAAAATTTTTTTATCAGGTACTTGATATGGCTATTTTATATTAC a909_al2.seq
TATTACCGAGTGCTTTATGAATAACATATGGTTCTTTCAACTTATGTTGA Majority
10660 10670 10680 10690 10700
10639 TATTACCGAGTGCTTTATGAATAACATATGGTTCTTTCAACTTATGTTGA cohl_al2.seq
10635 TATTACCGAGTGCTTTATGAATAACATATGGTTCTTTCAACTTATGTTGA a909_al2.seq
TTCACGTAAACATCCCCTTGAGAATCAATATTAACCCAGTCTCCTGACTC Majority
10710 10720 10730 10740 10750
10689 TTCACGTAAACATCCCCTTGAGAATCAATATTAACCCAGTCTCCTGACTC cohl_al2.seq
10685 TTCACGTAAACATCCCCTTGAGAATCAATATTAACCCAGTCTCCTGACTC a909_al2.seq
TGCAATAACCCGCTTGACTAGGACCTTATTATTGTAGTAAAACGCGACAA Majority
10760 10770 10780 10790 10800
10739 TGCAATAACCCGCTTGACTAGGACCTTATTATTGTAGTAAAACGCGACAA cohl_al2.seq
10735 TGCAATAACCCGCTTGACTAGGACCTTATTATTGTAGTAAAACGC.ACAA a909_al2.seq
CGTCTCCAGTTTTAAAATTTGAACCTTTTACTGTAAAGACTACATCACCT Majority
10810 10820 10830 10840 10850
10789 CGTCTCCAGTTTTAAAATTTGAACCTTTTACTGTAAAGACTACATCACCT cohl_al2.seq
10785 CGTCTCCAGTTTTAAAATTTGAACCTTTTACTGTAAAGACTACATCACCT a909_al2.seq
GCACTTAAAGTCTTATTCATTGAATGTCCGTAGATTCTTAATACAGGCAA Majority
10860 10870 10880 10890 10900
10839 GCACTTAAAGTCTTATTCATTGAATGTCCGTAGATTCTTAATACAGGCAA cohl_al2.seq
10835 GCACTTAAAGTCTTATTCATTGAATGTCCGTAGATTCTTAATACAGGCAA a909_al2.seq
CCATAAAACCGCAATTAAAATGGCTGTTGAGGCAACCGCCATCAAGATGT Majority
10910 10920 10930 10940 10950
10889 CCATAAAACCGCAATTAAAATGGCT.TTGAGGCAACCGCCATCAAGATGT cohl_al2.seq
10885 CCATAAAACCGCAATTAAAATGGCTGTTGAGGCAACCGCCATCAAGATGT a909_al2.seq
ATATGGTATTTTTAATGACACTCCAAAAGCGTTTCTGATAAGTTACGCGA Majority
10960 10970 10980 10990 11000
10939 ATATGGTATTTTTAATGACACTCCAAAAGCGTTTCTGATAAGTTACGCGA cohl_al2.seq
10935 ATATGGTATTTTTAATGACACTCCAAAAGCGTTTCTGATAAGTTACGCGA a909_al2.seq
TCCAGTTCTTGAGATAATTTATCTGAACTAATCTGTCTTTTCATTGTCTA Majority
11010 11020 11030 11040 11050
10989 TCCAGTTCTTGAGATAATTT.ATCTGAACTAAT.CTGTCTTTTCATTGTCTA cohl_al2.seq
10985 TCCAGTTCTTGAGATAATTTATCTGAACTAATCTGTCTTTTCATTGTCTA a909_al2.seq

FIGURE 21P

```
                CGCCTCTCTTATCTACTAAATTCTTTACGTTTTCTAAATAGGTATCTGCA Majority
                        11060     11070     11080     11090     11100
11039 CGCCTCTCTTATCTACTAAATTCTTTACGTTTTCTAAATAGGTATCTGCA cohl_al2.seq
11035 CGCCTCTCTTATCTACTAAATTCTTTACGTTTTCTAAATAGGTATCTGCA a909_al2.seq

                GCCTAACTGAGCAGTCTCAAAAATACCACTAAGCATTAAGGAAGCTTCGGC Majority
                        11110     11120     11130     11140     11150
11089 GCTAACTGAGCAGTCTCAAAAATACCACTAAGCATTAAGGAAGCTTCGGC cohl_al2.seq
11085 GCTAACTGAGCAGTCTCAAAAATACCACTAAGCATTAAGGAAGCTTCGGC a909_al2.seq

                AATAGAACCAGCTTTGCTAATTTTGATTTTTTTATCATCTAGCGCTTCTT Majority
                        11160     11170     11180     11190     11200
11139 AATAGAACCAGCTTTGCTAATTTTGATTTTTTTATCATCTAGCGCTTCTT cohl_al2.seq
11135 AATAGAACCAGCTTTGCTAATTTTGATTTTTTTATCATCTAGCGCTTCTT a909_al2.seq

                TAAGTTGCTGAATCTCTTTCTCTTGTTTTTCAATAAGAAGTTGCTGTTCT Majority
                        11210     11220     11230     11240     11250
11189 TAAGTTGCTGAATCTCTTTCTCTTGTTTTTCAATAAGAAGTTGCTGTTCT cohl_al2.seq
11185 TAAGTTGCTGAATCTCTTTCTCTTGTTTTTCAATAAGAAGTTGCTGTTCT a909_al2.seq

                AACATAATTTCTAGCAAGTCTTTTCTTTTTAATTTTTTTAAATCTTCCAT Majority
                        11260     11270     11280     11290     11300
11239 AACATAATTTCTAGCAAGTCTTTTCTTTTTAATTTTTTTAAATCTTCCAT cohl_al2.seq
11235 AACATAATTTCTAGCAAGTCTTTTCTTTTTAATTTTTTTAAATCTTCCAT a909_al2.seq

                CGCGATTACTTCCTTAACTGAACCTTAAATTATCGTTTAGATATTATATC Majority
                        11310     11320     11330     11340     11350
11289 CGCGATTACTTCCTTAACTGAACCTTAAATTATCGTTTAGATATTATATC cohl_al2.seq
11285 CGCGATTACTTCCTTAACTGAACCTTAAATTATCGTTTAGATATTATATC a909_al2.seq

                AAAGTTCTAACCTTTAAACTCATTTTTTGTCCTGTGTTTTTTCTCAAAAA Majority
                        11360     11370     11380     11390     11400
11339 AAAGTTCTAACCTTTAAACTCATTTTTTGTCCTGTGTTTTTTCTCAAAAA cohl_al2.seq
11335 AAAGTTCTAACCTTTAAACTCATTTTTTGTCCTGTGTTTTTTCTCAAAAA a909_al2.seq

                AGTCTATGCTAAATTAACATTTTTGATAATTTTTTGAAAAATCTCATCGA Majority
                        11410     11420     11430     11440     11450
11389 AGTCTATGCTAAATTAACATTTTTGATAATTTTTTGAAAAATCTCATCGA cohl_al2.seq
11385 AGTCTATGCTAAATTAACATTTTTGATAATTTTTTGAAAAATCTCATCGA a909_al2.seq

                AGTCATTTTCTTTTTGAAAGCTCGAATTCTAGGCATTAAAAAGCCATATA Majority
                        11460     11470     11480     11490     11500
11439 AGTCATTTTCTTTTTGAAAGCTCGAATTCTAGGCATTAAAAAGCCATATA cohl_al2.seq
11435 AGTCATTTTCTTTTTGAAAGCTCGAATTCTAGGCATTAAAA---------- a909_al2.seq

                TCAAATTGATATATGGCTTTTTTTATTATTTAAAACAAAAGAATCAATAG Majority
                        11510     11520     11530     11540     11550
11489 TCAAATTGATATATGGCTTTTTTTATTATTTAAAACAAAAGAATCAATAG cohl_al2.seq
11476 ----------------------------------ACAAAAGAATCAATAG a909_al2.seq

                GACAATAGCGTCAATTTAGTGACATAATCTATTACAGATTAAGTTCTTTT Majority
                        11560     11570     11580     11590     11600
11539 GACAATAGCGTCAATTTAGTGACATAATCTATTACAGATTAAGTTCTTTT cohl_al2.seq
11492 GACAATAGCATCAATTTAGTGACACAATCTATTACAGATTAAGTTCTTTT a909_al2.seq

                TGAATAATATAATCCAACTTTTCAACTGTTTTTTCCCATGTGAAATGTTC Majority
                        11610     11620     11630     11640     11650
11589 TGAATAATATAATCCAACTTTTCAACTGTTTTTTCCCATGTGAAATGTTC cohl_al2.seq
11542 TGAATAATATAATCCAACTTTTCAACTGTTTTTTCCCAGGTGAAATGTTC a909_al2.seq

                TTTAATTCTTTTAGCAATATTCTGTTGTAGTTTCTCTCTTAATGCCTTAT Majority
                        11660     11670     11680     11690     11700
11639 TTTAATTCTTTTAGCAATATTCTGTTGTAGTTTCTCTCTTAATGCCTTAT cohl_al2.seq
11592 TTTAATTCTTTTAGCAATATTCTGTTGTAATTTCTCTCTTAATGCCTTAT a909_al2.seq
```

FIGURE 21Q

128/487

```
                                                
```

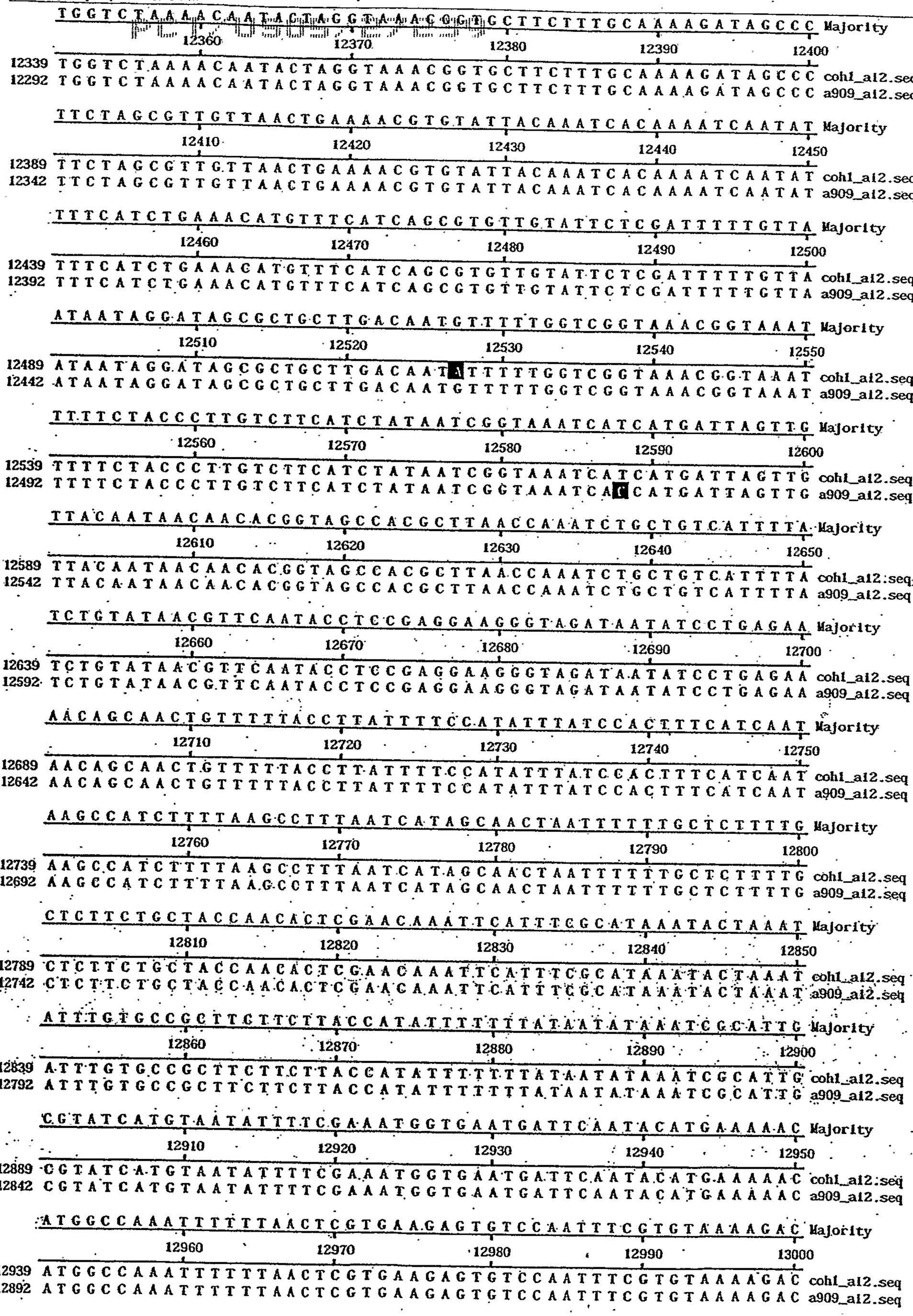

TGGTCTAAAACAATACTAGGTAAACGGTGCTTCTTTGCAAAAGATAGCCC Majority
12360 12370 12380 12390 12400
12339 TGGTCTAAAACAATACTAGGTAAACGGTGCTTCTTTGCAAAAGATAGCCC coh1_a12.seq
12292 TGGTCTAAAACAATACTAGGTAAACGGTGCTTCTTTGCAAAAGATAGCCC a909_a12.seq
TTCTAGCGTTGTTAACTGAAAACGTGTATTACAAATCACAAAATCAATAT Majority
12410 12420 12430 12440 12450
12389 TTCTAGCGTTGTTAACTGAAAACGTGTATTACAAATCACAAAATCAATAT coh1_a12.seq
12342 TTCTAGCGTTGTTAACTGAAAACGTGTATTACAAATCACAAAATCAATAT a909_a12.seq
TTTCATCTGAAACATGTTTCATCAGCGTGTTGTATTCTCGATTTTTGTTA Majority
12460 12470 12480 12490 12500
12439 TTTCATCTGAAACATGTTTCATCAGCGTGTTGTATTCTCGATTTTTGTTA coh1_a12.seq
12392 TTTCATCTGAAACATGTTTCATCAGCGTGTTGTATTCTCGATTTTTGTTA a909_a12.seq
ATAATAGGATAGCGCTGCTTGACAATGTTTTTGGTCGGTAAACGGTAAAT Majority
12510 12520 12530 12540 12550
12489 ATAATAGGATAGCGCTGCTTGACAATATTTTTGGTCGGTAAACGGTAAAT coh1_a12.seq
12442 ATAATAGGATAGCGCTGCTTGACAATGTTTTTGGTCGGTAAACGGTAAAT a909_a12.seq
TTTCTACCCTTGTCTTCATCTATAATCGGTAAATCATCATGATTAGTTG Majority
12560 12570 12580 12590 12600
12539 TTTTCTACCCTTGTCTTCATCTATAATCGGTAAATCATCATGATTAGTTG coh1_a12.seq
12492 TTTTCTACCCTTGTCTTCATCTATAATCGGTAAATCACCATGATTAGTTG a909_a12.seq
TTACAATAACAACACGGTAGCCACGCTTAACCAAATCTGCTGTCATTTTA Majority
12610 12620 12630 12640 12650
12589 TTACAATAACAACACGGTAGCCACGCTTAACCAAATCTGCTGTCATTTTA coh1_a12.seq
12542 TTACAATAACAACACGGTAGCCACGCTTAACCAAATCTGCTGTCATTTTA a909_a12.seq
TCTGTATAACGTTCAATACCTCCGAGGAAGGGTAGATAATATCCTGAGAA Majority
12660 12670 12680 12690 12700
12639 TCTGTATAACGTTCAATACCTCCGAGGAAGGGTAGATAATATCCTGAGAA coh1_a12.seq
12592 TCTGTATAACGTTCAATACCTCCGAGGAAGGGTAGATAATATCCTGAGAA a909_a12.seq
AACAGCAACTGTTTTTACCTTATTTTCCATATTTATCCACTTTCATCAAT Majority
12710 12720 12730 12740 12750
12689 AACAGCAACTGTTTTTACCTTATTTTCCATATTTATCCACTTTCATCAAT coh1_a12.seq
12642 AACAGCAACTGTTTTTACCTTATTTTCCATATTTATCCACTTTCATCAAT a909_a12.seq
AAGCCATCTTTTAAGCCTTTAATCATAGCAACTAATTTTTTGCTCTTTTG Majority
12760 12770 12780 12790 12800
12739 AAGCCATCTTTTAAGCCTTTAATCATAGCAACTAATTTTTTGCTCTTTTG coh1_a12.seq
12692 AAGCCATCTTTTAAGCCTTTAATCATAGCAACTAATTTTTTGCTCTTTTG a909_a12.seq
CTCTTCTGCTACCAACACTCGAACAAATTCATTTCGCATAAATACTAAAT Majority
12810 12820 12830 12840 12850
12789 CTCTTCTGCTACCAACACTCGAACAAATTCATTTCGCATAAATACTAAAT coh1_a12.seq
12742 CTCTTCTGCTACCAACACTCGAACAAATTCATTTCGCATAAATACTAAAT a909_a12.seq
ATTTGTGCCGCTTCTTCTTACCATATTTTTTTATAATATAAATCGCATTG Majority
12860 12870 12880 12890 12900
12839 ATTTGTGCCGCTTCTTCTTACCATATTTTTTTATAATATAAATCGCATTG coh1_a12.seq
12792 ATTTGTGCCGCTTCTTCTTACCATATTTTTTTATAATATAAATCGCATTG a909_a12.seq
CGTATCATGTAATATTTTCGAAATGGTGAATGATTCAATACATGAAAAAC Majority
12910 12920 12930 12940 12950
12889 CGTATCATGTAATATTTTCGAAATGGTGAATGATTCAATACATGAAAAAC coh1_a12.seq
12842 CGTATCATGTAATATTTTCGAAATGGTGAATGATTCAATACATGAAAAAC a909_a12.seq
ATGGCCAAATTTTTTTAACTCGTGAAGAGTGTCCAATTTCGTGTAAAAGAC Majority
12960 12970 12980 12990 13000
12939 ATGGCCAAATTTTTTTAACTCGTGAAGAGTGTCCAATTTCGTGTAAAAGAC coh1_a12.seq
12892 ATGGCCAAATTTTTTTAACTCGTGAAGAGTGTCCAATTTCGTGTAAAAGAC a909_a12.seq

FIGURE 21S

```
CAATAAATTACCTGATAAGTCTTATATCCCATCTCTGACAGACGATAA Majority
          13010     13020     13030     13040     13050
12989 CAATAAAATTAACCTGATAAGTCTTATATCCCATCTCTGACAGACGATAA cohl_al2.seq
12942 CAATAAAATTAACCTGATAAGTCTTATATCCCATCTCTGACAGACGATAA a909_al2.seq

TTCATTTCAGAGTCAACAAAATCAATAAACATCTCTTCTGCAAAGCCAGA Majority
          13060     13070     13080     13090     13100
13039 TTCATTTCAGAGTCAACAAAATCAATAAACATCTCTTCTGCAAAGCCAGA cohl_al2.seq
12992 TTCATTTCAGAGTCAACAAAATCAATAAACATCTCTTCTGCAAAGCCAGA a909_al2.seq

TGTTTCTTCAAAAACGCTCGTTTTCATTAAAGCAGCCGAAGTAATACACT Majority
          13110     13120     13130     13140     13150
13089 TGTTTCTTCAAAAACGCTCGTTTTCATTAAAGCAGCCGAAGTAATACACT cohl_al2.seq
13042 TGTTTCTTCAAAAACGCTCGTTTTCATTAAAGCAGCCGAAGTAATACACT a909_al2.seq

CTTCAATTTCTTTATAGTCAAATTCTTGCATCACTAAATCTTCACGGTTC Majority
          13160     13170     13180     13190     13200
13139 CTTCAATTTCTTTATAGTCAAATTCTTGCATAACTAAATCTTCACGGTTC cohl_al2.seq
13092 CTTCAATTTCTTTATAGTCAAATTCTTGCATCACTAAATCTTCACGGTTC a909_al2.seq

ATATCTTGATACAAACAAGATAACATACCGACCTTAGGTAAATGAAGGTA Majority
          13210     13220     13230     13240     13250
13189 ATATCTTGATACAAACAAGATAACATACCGACCTTAGGTAAAAGAAGGTA cohl_al2.seq
13142 ATATCTTGATACAAACAAGATAACATACCGACCTTAGGTAAATGAAGGTA a909_al2.seq

ATTTTCATAATTATCTATCAAATCACCTAGGACAACCGAATCTTGATCTA Majority
          13260     13270     13280     13290     13300
13239 ATTTTCATAATTATCTATCAAATCACCTAGGACAACCGAATCTTGATCTA cohl_al2.seq
13192 ATTTTCATAATTATCTATCAAATCACCTAGGACAACCGAATCTTGATCTA a909_al2.seq

AAGTCAAGAACCAATCAAATTCTTGTGCTACTGCAAATTGACCGATACAG Majority
          13310     13320     13330     13340     13350
13289 AAGTCAAGAACCAATCAAATTCTTGTGCTACTGCAAATTGACCGATACAG cohl_al2.seq
13242 AAGTCAAGAACCAATCAAATTCTTGTGCTACTGCAAATTGACCAATACAG a909_al2.seq

TTCAAAGCATATGCAATTCCTTTATTTTCTGTTAAATAATCAACAGTTAG Majority
          13360     13370     13380     13390     13400
13339 TTCAAAGCATATGCAATTCCTTTATTTTCTGTTAAATAATCAACAGTTAG cohl_al2.seq
13292 TTCAAAGCATATGCAATTCCTTTATTTTCTGTTAAATAATCAACAGTTAG a909_al2.seq

GTGCCCCTCTTCATTATAATCGGCCACTAATTGAGAAATTTCTTCCTTAT Majority
          13410     13420     13430     13440     13450
13389 GTGCCCCTCTTCATTATAATCGGCCACTAATTGAGAAATTTCTTCCTTAT cohl_al2.seq
13342 GTGCCCCTCTTCATTATAATCGGCCACTAATTGAGAAATTTCGTCCTTAT a909_al2.seq

TTTTCGAGCCATTATCTACGATGTAGATATGGCTTACTTGAGGATAAATT Majority
          13460     13470     13480     13490     13500
13439 TTTTCGAGCCATTATCTACGATGTAGATATGGCTTACTTGAGGATAAATT cohl_al2.seq
13392 TTTTCGAGCCATTATCTACGATGTAGATATGGCTTACTTGAGGATAAATT a909_al2.seq

GCTCGAATGTTCTGATCTAAGCGTTCAATATTGGGGTTAAAGGTGACAAT Majority
          13510     13520     13530     13540     13550
13489 GCTCGAATGTTCTGATCTAAGCGTTCAATATTGGGGTTAAAGGTGACAAT cohl_al2.seq
13442 GCTCGAATGTTCTGATCTAAGCGTTCAATATTGGGGTTAAAGGTGACAAT a909_al2.seq

ACCCGCTAAATATTTCATGTTCTATGCTCTTTTCTAAAATCTCTAAATAA Majority
          13560     13570     13580     13590     13600
13539 ACCCGCTAAATATTTCATGTTCTATGCTCTTTTCTAAAATCTCTAAATAA cohl_al2.seq
13492 ACCCGCTAAATATTTCATGTTCTATGCTCTTTTCTAAAATCTCTAAATAA a909_al2.seq

CTGAATGACTGGTGCTTTGGTTATAAAAACGATACCGACATAGATAGTTA Majority
          13610     13620     13630     13640     13650
13589 CTGAATGACTGGTGCTTTGGTTATAAAAACGATACCGACATAGATAGTTA cohl_al2.seq
13542 CTGAATGACTGGTGCTTTGGTTATAAAAACGATACCGACATAGATAGTTA a909_al2.seq
```

FIGURE 21T

```
CTGCTACTAAACTTTGAATGACATAATTTACCAATGATACTGACATTTGA Majority
          13660     13670     13680     13690     13700
13639 CTGCTACTAAACTTTGAATGACATAATTTACCAATGATACTGACATTTGA cohl_al2.seq
13592 CTGCTACTAAACTTTGAATGACATAATTTACCAATGATACTGACATTTGA a909_al2.seq

GTATTGATATAATAGAGTACAGCTCCACTAAGAGTAGCAGCAATTAAATA Majority
          13710     13720     13730     13740     13750
13689 GTATTGATATAATAAAGTACAGCTCCACTAAGAGTAGCAGCAATTAAATA cohl_al2.seq
13642 GTATTGATATAATAGAGTACAGCTCCACTAAGAGTAGCAGCAATTAAATA a909_al2.seq

GCGCAGCATTCCTCTTGTTAATTCTTTAAAAGTAAATACATCTCTTAAAG Majority
          13760     13770     13780     13790     13800
13739 GCGCAGCATTCCTCTTGTTAATTCTTTAAAAGTAAATACATCTCTTAAAG cohl_al2.seq
13692 GCGCAGCATTCCTCTTGTTAATTCTTTAAAAGTAAATACATCTCTTAAAG a909_al2.seq

AGATAGCTTGATATAGGGAGACAATAAATTCAGTAATAACTGTAGAGATA Majority
          13810     13820     13830     13840     13850
13789 AGATAGCTTGATAGAGGGAGACAATAAATTCAGTAATAACTGTAGAGATA cohl_al2.seq
13742 AGATAGCTTGATATAGAGAGACAATAAATTCAGTAATAACTGTAGAGATA a909_al2.seq

ATAGCTCCCATAGCACCTAAAATTGGTATTAAAAGTATATTAAGCACAAC Majority
          13860     13870     13880     13890     13900
13839 ATAGCTCCCATAGCACCTAAAATTGGTATTAAAAGTATATTAAGCACAAC cohl_al2.seq
13792 ATAGCTCCCATAGCACCTAAAATTGGTATTAAAAGTATATTAAGCACAAC a909_al2.seq

ATTTGCCACAAGTCCAATAACTGCAGACATTGTGTAAGCTTTTGTACGTC Majority
          13910     13920     13930     13940     13950
13889 ATTTGCCACAAGTCCAATAACTGCAGACATTGTGTAAGCTTTTGTACGTC cohl_al2.seq
13842 ATTTGCCACAAGTCCAATAACTGCAGACATTGTGTAAGCTTTTGTACGTC a909_al2.seq

TTGAAGCCAGTAGATACTGTGTCCCTAAAGCGTTACCATAAGAAATGCAA Majority
          13960     13970     13980     13990     14000
13939 TTGAAGCCAGTAGATACTGTGTCCCTAAAGCGTTACCATAAGAAATGCAA cohl_al2.seq
13892 TTGAAGCCAGTAGATACTGTGTCCCTAAAGCGTTACCATAAGAAATGCAA a909_al2.seq

ATGATCATCAAA Majority
          14010
13989 ATGATCATCAAA cohl_al2.seq
13942 ATGATCATCAAA a909_al2.seq
```

Decoration 'Decoration #1': Shade (with solid black) residues that differ from the Consensus.

FIGURE 21U

Alignment Report of gbs[illegible], using Clustal method with PAM250 residue weight table.
Thursday, July 29, 2004 6:57 PM


```
   MKLSKKLLFSAAVLTMVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTV Majority
            10        20        30        40        50
1  MKLSKKLLFSAAVLTMVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTV sag645_2603.pep
1  MKLSKKLLFSAAVLTIVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTV sag645_a909.pep
1  MKLSKKLLFSAAVLTIVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTV sag645_cjb111.pep
1  MKLSKKLLFSAAVLTMVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTV sag645_coh1.pep
1  MKLSKKLLFSAAVLTMVAGSTVEPVAQFATGMSIVRAAEVSQERPAKTTV sag645_nem316.pep

   NIYKLQADSYKSEITSNGGIENKDGEVISNYAKLGDNVKGLQGVQFKRYK Majority
            60        70        80        90       100
51 NIYKLQADSYKSEITSNGGIENKDGEVISNYAKLGDNVKGLQGVQFKRYK sag645_2603.pep
51 NIYKLQADSYKSEITSNGGIENKDGEVISNYAKLGDNVKGLQGVQFKRYK sag645_a909.pep
51 NIYKLQADSYKSEITSNGGIENKDGEVISNYAKLGDNVKGLQGVQFKRYK sag645_cjb111.pep
51 NIYKLQADSYKSEITSNGGIENKDGEVISNYAKLGDNVKGLQGVQFKRYK sag645_coh1.pep
51 NIYKLQADSYKSEITSNGGIENKDGEVISNYAKLGDNVKGLQGVQFKRYK sag645_nem316.pep

   VKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK Majority
           110       120       130       140       150
01 VKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK sag645_2603.pep
01 VKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK sag645_a909.pep
01 VKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK sag645_cjb111.pep
01 VKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK sag645_coh1.pep
01 VKTDISVDELKKLTTVEAADAKVGTILEEGVSLPQKTNAQGLVVDALDSK sag645_nem316.pep

   SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKN Majority
           160       170       180       190       200
51 SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKN sag645_2603.pep
51 SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKN sag645_a909.pep
51 SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKN sag645_cjb111.pep
51 SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKN sag645_coh1.pep
51 SNVRYLYVEDLKNSPSNITKAYAVPFVLELPVANSTGTGFLSEINIYPKN sag645_nem316.pep

   VVTDEPKTDKDVKKLGQDDAGYTIGEEFKWFLKSTIPANLGDYEKFEITD Majority
           210       220       230       240       250
01 VVTDEPKTDKDVKKLGQDDAGYTIGEEFKWFLKSTIPANLGDYEKFEITD sag645_2603.pep
01 VVTDEPKTDKDVKKLGQDDAGYTIGEEFKWFLKSTIPANLGDYEKFEITD sag645_a909.pep
01 VVTDEPKTDKDVKKLGQDDAGYTIGEEFKWFLKSTIPANLGDYEKFEITD sag645_cjb111.pep
01 VVTDEPKTDKDVKKLGQDDAGYTIGEEFKWFLKSTIPANLGDYEKFEITD sag645_coh1.pep
01 VVTDEPKTDKDVKKLGQDDAGYTIGEEFKWFLKSTIPANLGDYEKFEITD sag645_nem316.pep

   KFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK Majority
           260       270       280       290       300
51 KFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK sag645_2603.pep
51 KFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK sag645_a909.pep
51 KFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK sag645_cjb111.pep
51 KFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK sag645_coh1.pep
51 KFADGLTYKSVGKIKIGSKTLNRDEHYTIDEPTVDNQNTLKITFKPEKFK sag645_nem316.pep

   EIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIE Majority
           310       320       330       340       350
01 EIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIE sag645_2603.pep
01 EIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIE sag645_a909.pep
01 EIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIE sag645_cjb111.pep
01 EIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIE sag645_coh1.pep
01 EIAELLKGMTLVKNQDALDKATANTDDAAFLEIPVASTINEKAVLGKAIE sag645_nem316.pep

   NTFELQYDHTPDKADNPKPSNPPRKPEVHTGGKRFVKKDSTETQTLGGAE Majority
           360       370       380       390       400
51 NTFELQYDHTPDKADNPKPSNPPRKPEVHTGGKRFVKKDSTETQTLGGAE sag645_2603.pep
51 NTFELQYDHTPDKADNPKPSNPPRKPEVHTGGKRFVKKDSTETQTLGGAE sag645_a909.pep
51 NTFELQYDHTPDKADNPKPSNPPRKPEVHTGGKRFVKKDSTETQTLGGAE sag645_cjb111.pep
51 NTFELQYDHTPDKADNPKPSNPPRKPEVHTGGKRFVKKDSTETQTLGGAE sag645_coh1.pep
51 NTFELQYDHTPDKADNPKPSNPPRKPEVHTGGKRFVKKDSTETQTLGGAE sag645_nem316.pep
```

Figure 22

```
   F D L L A S D G T A V K W T D A L I K A N T N K N Y I A G E A V T G Q P I K L K S H T D G T F E I K  Majority
                410       420       430       440       450
101 F D L L A S D G T A V K W T D A L I K A N T N K N Y I A G E A V T G Q P I K L K S H T D G T F E I K
101 F D L L A S D G T A V K W T D A L I K A N T N K N Y I A G E A V T G Q P I K L K S H T D G T F E I K  sag645_2603.pep
101 F D L L A S D G T A V K W T D A L I K A N T N K N Y I A G E A V T G Q P I K L K S H T D G T F E I K  sag645_a909.pep
101 F D L L A S D G T A V K W T D A L I K A N T N K N Y I A G E A V T G Q P I K L K S H T D G T F E I K  sag645_cjb111.pep
101 F D L L A S D G T A V K W T D A L I K A N T N K N Y I A G E A V T G Q P I K L K S H T D G T F E I K  sag645_coh1.pep
101 F D L L A S D G T A V K W T D A L I K A N T N K N Y I A G E A V T G Q P I K L K S H T D G T F E I K  sag645_nem316.pep

   G L A Y A V D A N A E G T A V T Y K L K E T K A P E G Y V I P D K E I E F T V S Q T S Y N T K P T D  Majority
                460       470       480       490       500
151 G L A Y A V D A N A E G T A V T Y K L K E T K A P E G Y V I P D K E I E F T V S Q T S Y N T K P T D
151 G L A Y A V D A N A E G T A V T Y K L K E T K A P E G Y V I P D K E I E F T V S Q T S Y N T K P T D  sag645_2603.pep
151 G L A Y A V D A N A E G T A V T Y K L K E T K A P E G Y V I P D K E I E F T V S Q T S Y N T K P T D  sag645_a909.pep
151 G L A Y A V D A N A E G T A V T Y K L K E T K A P E G Y V I P D K E I E F T V S Q T S Y N T K P T D  sag645_cjb111.pep
151 G L A Y A V D A N A E G T A V T Y K L K E T K A P E G Y V I P D K E I E F T V S Q T S Y N T K P T D  sag645_coh1.pep
151 G L A Y A V D A N A E G T A V T Y K L K E T K A P E G Y V I P D K E I E F T V S Q T S Y N T K P T D  sag645_nem316.pep

   I T V D S A D A T P D T I K N N K R P S I P N T G G I G T A I F V A I G A A V M A F A V K G M K R R  Majority
                510       520       530       540       550
201 I T V D S A D A T P D T I K N N K R P S I P N T G G I G T A I F V A I G A A V M A F A V K G M K R R
201 I T V D S A D A T P D T I K N N K R P S I P N T G G I G T A I F V A I G A A V M A F A V K G M K R R  sag645_2603.pep
201 I T V D S A D A T P D T I K N N K R P S I P N T G G I G T A I F V A I G A A V M A F A V K G M K R R  sag645_a909.pep
201 I T V D S A D A T P D T I K N N K R P S I P N T G G I G T A I F V A I G A A V M A F A V K G M K R R  sag645_cjb111.pep
201 I T V D S A D A T P D T I K N N K R P S I P N T G G I G T A I F V A I G A A V M A F A V K G M K R R  sag645_coh1.pep
201 I T V D S A D A T P D T I K N N K R P S I P N T G G I G T A I F V A I G A A V M A F A V K G M K R R  sag645_nem316.pep

   T K D N                                                                                          Majority

251 T K D N
251 T K D N                                                                                         sag645_2603.pep
251 T K D N                                                                                         sag645_a909.pep
251 T K D N                                                                                         sag645_cjb111.pep
251 T K D N                                                                                         sag645_coh1.pep
                                                                                                    sag645_nem316.pep
```

ecoration 'Decoration #1': Shade (with solid black) residues that differ from the Consensus.

FIGURE 22A

```
MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNA Majority
         10        20        30        40        50
1  MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNA sag649_2603.pep
1  MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNA sag649_coh1.pep
1  MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNA sag649_cjb111.pep
1  MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNA sag649_18rs21.pep
1  MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNA sag649_nem316.pep
1  MKKRQKIWRGLSVTLLILSQIPFGILVQGETQDTNQALGKVIVKKTGDNA sag649_cjb111.pep

TPLGKATFVLKNDNDKSETSHETVEGSGEATFENIKPGDYTLREETAPIG Majority
         60        70        80        90       100
51 TPLGKATFVLKNDNDKSETSHETVEGSGEATFENIKPGDYTLREETAPIG sag649_2603.pep
51 TPLGKATFVLKNDNDKSETSHETVEGSGEATFENIKPGDYTLREETAPIG sag649_coh1.pep
51 TPLGKATFVLKNDNDKSETSHETVEGSGEATFENIKPGDYTLREETAPIG sag649_cjb111.pep
51 TPLGKATFVLKNDNDKSETSHETVEGSGEATFENIKPGDYTLREETAPIG sag649_18rs21.pep
51 TPLGKATFVLKNDNDKSETSHETVEGSGEATFENIKPGDYTLREETAPIG sag649_nem316.pep
51 TPLGKATFVLKNDNDKSETSHETVEGSGEATFENIKPGDYTLREETAPIG sag649_cjb111.pep

YKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKE Majority
        110       120       130       140       150
101 YKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKE sag649_2603.pep
101 YKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKE sag649_coh1.pep
101 YKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKE sag649_cjb111.pep
101 YKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKE sag649_18rs21.pep
101 YKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKE sag649_nem316.pep
101 YKKTDKTWKVKVADNGATIIEGMDADKAEKRKEVLNAQYPKSAIYEDTKE sag649_cjb111.pep

NYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKKITGVNDLDKNK Majority
        160       170       180       190       200
151 NYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKKITGVNDLDKNK sag649_2603.pep
151 NYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKK[ ]TGVNDLDKNK sag649_coh1.pep
151 NYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKKITGVNDLDKNK sag649_cjb111.pep
151 NYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKKITGVNDLDKNK sag649_18rs21.pep
151 NYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKKITGVNDLDKNK sag649_nem316.pep
151 NYPLVNVEGSKVGEQYKALNPINGKDGRREIAEGWLSKKITGVNDLDKNK sag649_cjb111.pep

YKIELTVEGKTTVETKELNQPLDVVVLLDNSNSMNNERANNSQRALKAGE Majority
        210       220       230       240       250
201 YKIELTVEGKTTVETKELNQPLDVVVLLDNSNSMNNERANNSQRALKAGE sag649_2603.pep
201 YKIELTVEGKTTVETKELNQPLDVVVLLDNSNSMNNERANNSQRALKAGE sag649_coh1.pep
201 YKIELTVEGKTTVETKELNQPLDVVVLLDNSNSMNNERANNSQRALKAGE sag649_cjb111.pep
201 YKIELTVEGKTTVETKELNQPLDVVVLLDNSNSMNNERANNSQRALKAGE sag649_18rs21.pep
201 YKIELTVEGKTTVETKELNQPLDVVVLLDNSNSMNNERANNSQRALKAGE sag649_nem316.pep
201 YKIELTVEGKTTVETKELNQPLDVVVLLDNSNSMNNERANNSQRALKAGE sag649_cjb111.pep

AVEKLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSV Majority
        260       270       280       290       300
251 AVEKLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSV sag649_2603.pep
251 AVEKLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSV sag649_coh1.pep
251 AVEKLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSV sag649_cjb111.pep
251 AVEKLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSV sag649_18rs21.pep
251 AVEKLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSV sag649_nem316.pep
251 AVEKLIDKITSNKDNRVALVTYASTIFDGTEATVSKGVADQNGKALNDSV sag649_cjb111.pep

SWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQ Majority
        310       320       330       340       350
301 SWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQ sag649_2603.pep
301 SWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQ sag649_coh1.pep
301 SWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQ sag649_cjb111.pep
301 SWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQ sag649_18rs21.pep
301 SWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQ sag649_nem316.pep
301 SWDYHKTTFTATTHNYSYLNLTNDANEVNILKSRIPKEAEHINGDRTLYQ sag649_cjb111.pep

FGATFTQKALMKANEILETQSSNARKKLIFHVTDGVPTMSYAINFNPYIS Majority
        360       370       380       390       400
351 FGATFTQKALMKANEILETQSSNARKKLIFHVTDGVPTMSYAINFNPYIS sag649_2603.pep
351 FGATFTQKALMKANEILETQSSNARKKLIFHVTDGVPTMSYAINFNPYIS sag649_coh1.pep
351 FGATFTQKALMKANEILETQSSNARKKLIFHVTDGVPTMSYAINFNPYIS sag649_cjb111.pep
351 FGATFTQKALMKANEILETQSSNARKKLIFHVTDGVPTMSYAINFNPYIS sag649_18rs21.pep
351 FGATFTQKALMKANEILETQSSNARKKLIFHVTDGVPTMSYAINFNPYIS sag649_nem316.pep
351 FGATFTQKALMKANEILETQSSNARKKLIFHVTDGVPTMSYAINFNPYIS sag649_cjb111.pep
```

Figure 23

```
                                                
```

```
                                                
```

```
                                                                                  
                M R K Y Q K F S K I L T L S L F C L S Q I P L N T N V L G E S T V P E N G A K G K L V V K K T D D Q Majority
                                                                                  
                         10        20        30        40        50               
1               M R K Y Q K F S K I L T L S L F C L S Q I P L N T N V L G E S T V P E N G A K G K L V V K K T D D Q sag1408_2603.pep
1               M R K Y Q K F S K I L T L S L F C L S Q I P L N T N V L G E S T V P E N G A K G K L V V K K T D D Q sag1408_515.pep
1               M R K Y Q K F S K I L T L S L F C L S Q I P L N T N V L G E S T V P E N G A K G K L V V K K T D D Q sag1408_18rs21.seq
1               M R K Y Q K F S K I L T L S L F C L S Q I P L N T N V L G E S T V P E N G A K G K L V V K K T D D Q sag1408_cjb111.pep
1               M R K Y Q K F S K I L T L S L F C L S Q I P L N T N V L G E S T V P E N G A K G K L V V K K T D D Q sag1408_h36b.pep
1               M R K Y Q K F S K I L T L S L F C L S Q I P L N T N V L G E S T V P E N G A K G K L V V K K T D D Q sag1408_nem316.pep

                N K P L S K A T F V L K T T A H P E S K I E K V T A E L T G E A T F D N L I P G D Y T L S E E T A P Majority
                                                                                  
                         60        70        80        90        100              
51              N K P L S K A T F V L K T T A H P E S K I E K V T A E L T G E A T F D N L I P G D Y T L S E E T A P sag1408_2603.pep
51              N K P L S K A T F V L K T T A H P E S K I E K V T A E L T G E A T F D N L I P G D Y T L S E E T A P sag1408_515.pep
51              N K P L S K A T F V L K T T A H P E S K I E K V T A E L T G E A T F D N L I P G D Y T L S E E T A P sag1408_18rs21.seq
51              N K P L S K A T F V L K T T A H P E S K I E K V T A E L T G E A T F D N L I P G D Y T L S E E T A P sag1408_cjb111.pep
51              N K P L S K A T F V L K T T A H P E S K I E K V T A E L T G E A T F D N L I P G D Y T L S E E T A P sag1408_h36b.pep
51              N K P L S K A T F V L K T T A H P E S K I E K V T A E L T G E A T F D N L I P G D Y T L S E E T A P sag1408_nem316.pep

                E G Y K K T N Q T W Q V K V E S N G K T T I Q N S G D K N S T I G Q N Q E E L D K Q Y P P T G I Y E Majority
                                                                                  
                        110       120       130       140       150              
101             E G Y K K T N Q T W Q V K V E S N G K T T I Q N S G D K N S T I G Q N Q E E L D K Q Y P P T G I Y E sag1408_2603.pep
101             E G Y K K T N Q T W Q V K V E S N G K T T I Q N S G D K N S T I G Q N Q E E L D K Q Y P P T G I Y E sag1408_515.pep
101             E G Y K K T N Q T W Q V K V E S N G K T T I Q N S G D K N S T I G Q N Q E E L D K Q Y P P T G I Y E sag1408_18rs21.seq
101             E G Y K K T N Q T W Q V K V E S N G K T T I Q N S G D K N S T I G Q N Q E E L D K Q Y P P T G I Y E sag1408_cjb111.pep
101             E G Y K K T N Q T W Q V K V E S N G K T T I Q N S G D K N S T I G Q N Q E E L D K Q Y P P T G I Y E sag1408_h36b.pep
101             E G Y K K T N Q T W Q V K V E S N G K T T I Q N S G D K N S T I G Q N Q E E L D K Q Y P P T G I Y E sag1408_nem316.pep

                D T K E S Y K L E H V K G S V P N G K S E A K A V N P Y S S E G E H I R E I P E G T L S K R I S E V Majority
                                                                                  
                        160       170       180       190       200              
151             D T K E S Y K L E H V K G S V P N G K S E A K A V N P Y S S E G E H I R E I P E G T L S K R I S E V sag1408_2603.pep
151             D T K E S Y K L E H V K G S V P N G K S E A K A V N P Y S S E G E H I R E I P E G T L S K R I S E V sag1408_515.pep
151             D T K E S Y K L E H V K G S V P N G K S E A K A V N P Y S S E G E H I R E I P E G T L S K R I S E V sag1408_18rs21.seq
151             D T K E S Y K L E H V K G S V P N G K S E A K A V N P Y S S E G E H I R E I P E G T L S K R I S E V sag1408_cjb111.pep
151             D T K E S Y K L E H V K G S V P N G K S E A K A V N P Y S S E G E H I R E I P E G T L S K R I S E V sag1408_h36b.pep
151             D T K E S Y K L E H V K G S V P N G K S E A K A V N P Y S S E G E H I R E I P E G T L S K R I S E V sag1408_nem316.pep

                G D L A H N K Y K I E L T V S G K T I V K P V D K Q K P L D V V F V L D N S N S M N N D G P N F Q R Majority
                                                                                  
                        210       220       230       240       250              
201             G D L A H N K Y K I E L T V S G K T I V K P V D K Q K P L D V V F V L D N S N S M N N D G P N F Q R sag1408_2603.pep
201             G D L A H N K Y K I E L T V S G K T I V K P V D K Q K P L D V V F V L D N S N S M N N D G P N F Q R sag1408_515.pep
201             G D L A H N K Y K I E L T V S G K T I V K P V D K Q K P L D V V F V L D N S N S M N N D G P N F Q R sag1408_18rs21.seq
201             G D L A H N K Y K I E L T V S G K T I V K P V D K Q K P L D V V F V L D N S N S M N N D G P N F Q R sag1408_cjb111.pep
201             G D L A H N K Y K I E L T V S G K T I V K P V D K Q K P L D V V F V L D N S N S M N N D G P N F Q R sag1408_h36b.pep
201             G D L A H N K Y K I E L T V S G K T I V K P V D K Q K P L D V V F V L D N S N S M N N D G P N F Q R sag1408_nem316.pep

                H N K A K K A A E A L G T A V K D I L G A N S D N R V A L V T Y G S D I F D G R S V D V V K G F K E Majority
                                                                                  
                        260       270       280       290       300              
251             H N K A K K A A E A L G T A V K D I L G A N S D N R V A L V T Y G S D I F D G R S V D V V K G F K E sag1408_2603.pep
251             H N K A K K A A E A L G T A V K D I L G A N S D N R V A L V T Y G S D I F D G R S V D V V K G F K E sag1408_515.pep
251             H N K A K K A A E A L G T A V K D I L G A N S D N R V A L V T Y G S D I F D G R S V D V V K G F K E sag1408_18rs21.seq
247             N N K A K K A G E A V H T I I K D V L G A N V I N R A A L V T Y G S D I F D G R I V K V I K G F K E sag1408_cjb111.pep
251             H N K A K K A A E A L G T A V K D I L G A N S D N R V A L V T Y G S D I F D G R S V D V V K G F K E sag1408_h36b.pep
251             H N K A K K A A E A L G T A V K D I L G A N S D N R V A L V T Y G S D I F D G R S V D V V K G F K E sag1408_nem316.pep

                D D K Y Y G L Q T K F T I Q T E N Y S H K Q L T N N A E E I I K R I P T E A P K A K W G S T T N G L Majority
                                                                                  
                        310       320       330       340       350              
301             D D K Y Y G L Q T K F T I Q T E N Y S H K Q L T N N A E E I I K R I P T E A P K A K W G S T T N G L sag1408_2603.pep
301             D D K Y Y G L Q T K F T I Q T E N Y S H K Q L T N N A E E I I K R I P T E A P K A K W G S T T N G L sag1408_515.pep
301             D D K Y Y G L Q T K F T I Q T E N Y S H K Q L T N N A E E I I K R I P T E A P K A K W G S T T Y G L sag1408_18rs21.seq
301             D D K Y Y G L Q T K F T I Q T E N Y S H K Q L T N N A E E I I K R I P T E A P K A K W G S T T N G L sag1408_cjb111.pep
297             D P . Y Y G L F T S F T V Q T A D Y S Y K K F T N I A A D I I K N I P K E A P I A K W G G T S I G L sag1408_h36b.pep
301             D D K Y Y G L Q T K F T I Q T E N Y S H K Q L T N N A E E I I K R I P T E A P R A K W G S T T N G L sag1408_nem316.pep

                T P E Q Q K E Y Y L S K V G E T F T M K A F M E A D D I L S Q V N R N S Q K I I V H V T D G V P T R Majority
                                                                                  
                        360       370       380       390       400              
351             T P E Q Q K E Y Y L S K V G E T F T M K A F M E A D D I L S Q V N R N S Q K I I V H V T D G V P T R sag1408_2603.pep
351             T P E Q Q K E Y Y L S K V G E T F T M K A F M E A D D I L S Q V N R N S Q K I I V H V T D G V P T R sag1408_515.pep
351             T P E Q Q K E Y Y L S K V G E T F T M K A F M E A D D I L S Q V N R N S Q K I I V H V T D G V P T R sag1408_18rs21.seq
351             T P E Q Q K E Y Y L S K V G E T F T M K A F M E A D D I L S Q V N R N S Q K I I V H V T D G V P T R sag1408_cjb111.pep
346             T P E K K R E Y D L S K V G E T F T M K A F M E A D T L L S S T Q R K S R K I I V H L T D G V P T R sag1408_h36b.pep
351             T P E Q Q K Q Y Y L S K V G E T F T M K A F M E A D D I L S Q V D R N S Q K I I V H I T D G V P T R sag1408_nem316.pep
```

Figure 24

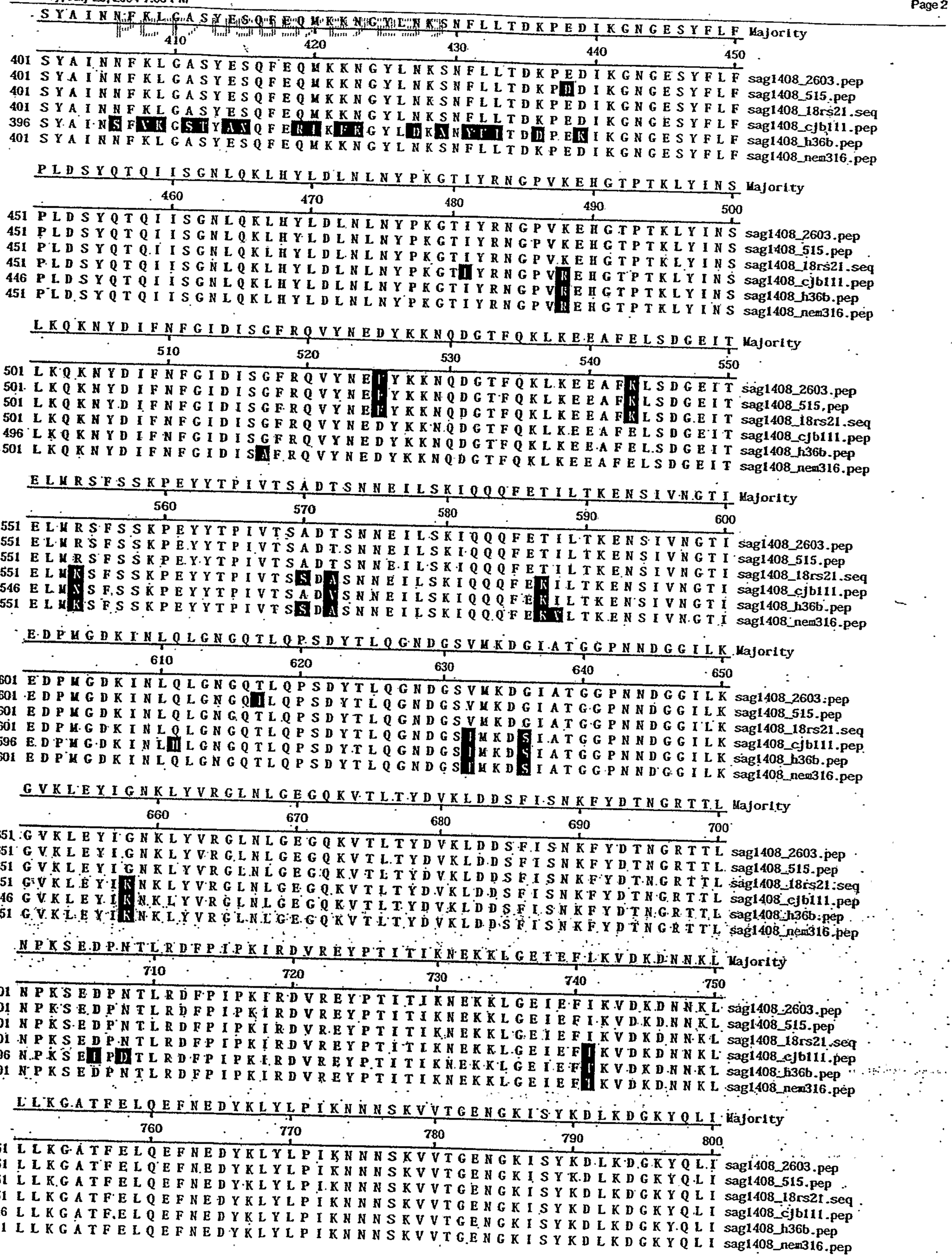

SYAINNFKLGASYESQFEQMKKNGYLNKSNFLLTDKPEDIKGNGESYFLF Majority
410 420 430 440 450
401 SYAINNFKLGASYESQFEQMKKNGYLNKSNFLLTDKPEDIKGNGESYFLF sag1408_2603.pep
401 SYAINNFKLGASYESQFEQMKKNGYLNKSNFLLTDKPDDIKGNGESYFLF sag1408_515.pep
401 SYAINNFKLGASYESQFEQMKKNGYLNKSNFLLTDKPEDIKGNGESYFLF sag1408_18rs21.seq
396 SYAINSFVKGSTYANQFERIKFRGYLDKANYFITDDPEKIKGNGESYFLF sag1408_cjb111.pep
401 SYAINNFKLGASYESQFEQMKKNGYLNKSNFLLTDKPEDIKGNGESYFLF sag1408_h36b.pep
401 SYAINNFKLGASYESQFEQMKKNGYLNKSNFLLTDKPEDIKGNGESYFLF sag1408_nem316.pep
PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVKEHGTPTKLYINS Majority
460 470 480 490 500
451 PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVKEHGTPTKLYINS sag1408_2603.pep
451 PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVKEHGTPTKLYINS sag1408_515.pep
451 PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVKEHGTPTKLYINS sag1408_18rs21.seq
446 PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVNEHGTPTKLYINS sag1408_cjb111.pep
451 PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVNEHGTPTKLYINS sag1408_h36b.pep
451 PLDSYQTQIISGNLQKLHYLDLNLNYPKGTIYRNGPVNEHGTPTKLYINS sag1408_nem316.pep
LKQKNYDIFNFGIDISGFRQVYNEDYKKNQDGTFQKLKEEAFELSDGEIT Majority
510 520 530 540 550
501 LKQKNYDIFNFGIDISGFRQVYNEFYKKNQDGTFQKLKEEAFKLSDGEIT sag1408_2603.pep
501 LKQKNYDIFNFGIDISGFRQVYNEFYKKNQDGTFQKLKEEAFKLSDGEIT sag1408_515.pep
501 LKQKNYDIFNFGIDISGFRQVYNEFYKKNQDGTFQKLKEEAFKLSDGEIT sag1408_18rs21.seq
496 LKQKNYDIFNFGIDISGFRQVYNEDYKKNQDGTFQKLKEEAFELSDGEIT sag1408_cjb111.pep
501 LKQKNYDIFNFGIDISAFRQVYNEDYKKNQDGTFQKLKEEAFELSDGEIT sag1408_h36b.pep
501 LKQKNYDIFNFGIDISGFRQVYNEDYKKNQDGTFQKLKEEAFELSDGEIT sag1408_nem316.pep
ELMRSFSSKPEYYTPIVTSADTSNNEILSKIQQQFETILTKENSIVNGTI Majority
560 570 580 590 600
551 ELMRSFSSKPEYYTPIVTSADTSNNEILSKIQQQFETILTKENSIVNGTI sag1408_2603.pep
551 ELMRSFSSKPEYYTPIVTSADTSNNEILSKIQQQFETILTKENSIVNGTI sag1408_515.pep
551 ELMKSFSSKPEYYTPIVTSSDASNNEILSKIQQQFETILTKENSIVNGTI sag1408_18rs21.seq
546 ELMNSFSSKPEYYTPIVTSADVSNNEILSKIQQQFEKILTKENSIVNGTI sag1408_cjb111.pep
551 ELMKSFSSKPEYYTPIVTSSDASNNEILSKIQQQFEKYLTKENSIVNGTI sag1408_h36b.pep
551 ELMKSFSSKPEYYTPIVTSSDASNNEILSKIQQQFEKYLTKENSIVNGTI sag1408_nem316.pep
EDPMGDKINLQLGNGQTLQPSDYTLQGNDGSVMKDGIATGGPNNDGGILK Majority
610 620 630 640 650
601 EDPMGDKINLQLGNGQILQPSDYTLQGNDGSVMKDGIATGGPNNDGGILK sag1408_2603.pep
601 EDPMGDKINLQLGNGQTLQPSDYTLQGNDGSVMKDGIATGGPNNDGGILK sag1408_515.pep
601 EDPMGDKINLQLGNGQTLQPSDYTLQGNDGSVMKDGIATGGPNNDGGILK sag1408_18rs21.seq
596 EDPMGDKINLHLGNGQTLQPSDYTLQGNDGSIMKDSIATGGPNNDGGILK sag1408_cjb111.pep
601 EDPMGDKINLQLGNGQTLQPSDYTLQGNDGSIMKDSIATGGPNNDGGILK sag1408_h36b.pep
601 EDPMGDKINLQLGNGQTLQPSDYTLQGNDGSIMKDSIATGGPNNDGGILK sag1408_nem316.pep
GVKLEYIGNKLYVRGLNLGEGQKVTLTYDVKLDDSFISNKFYDTNGRTTL Majority
660 670 680 690 700
651 GVKLEYIGNKLYVRGLNLGEGQKVTLTYDVKLDDSFISNKFYDTNGRTTL sag1408_2603.pep
651 GVKLEYIGNKLYVRGLNLGEGQKVTLTYDVKLDDSFISNKFYDTNGRTTL sag1408_515.pep
651 GVKLEYIKNKLYVRGLNLGEGQKVTLTYDVKLDDSFISNKFYDTNGRTTL sag1408_18rs21.seq
646 GVKLEYIKNKLYVRGLNLGEGQKVTLTYDVKLDDSFISNKFYDTNGRTTL sag1408_cjb111.pep
651 GVKLEYIKNKLYVRGLNLGEGQKVTLTYDVKLDDSFISNKFYDTNGRTTL sag1408_h36b.pep
651 GVKLEYIKNKLYVRGLNLGEGQKVTLTYDVKLDDSFISNKFYDTNGRTTL sag1408_nem316.pep
NPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL Majority
710 720 730 740 750
701 NPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL sag1408_2603.pep
701 NPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL sag1408_515.pep
701 NPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL sag1408_18rs21.seq
696 NPKSEIPDTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL sag1408_cjb111.pep
701 NPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL sag1408_h36b.pep
701 NPKSEDPNTLRDFPIPKIRDVREYPTITIKNEKKLGEIEFIKVDKDNNKL sag1408_nem316.pep
LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLI Majority
760 770 780 790 800
751 LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLI sag1408_2603.pep
751 LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLI sag1408_515.pep
751 LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLI sag1408_18rs21.seq
746 LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLI sag1408_cjb111.pep
751 LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLI sag1408_h36b.pep
751 LLKGATFELQEFNEDYKLYLPIKNNNSKVVTGENGKISYKDLKDGKYQLI sag1408_nem316.pep

FIGURE 24A

```
    E A V S P E D Y Q K I T N K P I L T F E V V K G S I Q N I I A V N K Q I S E Y H E E G D K H L I T N Majority
                                  810       820       830       840       850
801 E A V S P E D Y Q K I T N K P I L T F E V V K G S I K N I I A V N K Q I S E Y H E E G D K H L I T N sag1408_2603.pep
801 E A V S P E D Y Q K I T N K P I L T F E V V K G S I K N I I A V N K Q I S E Y H E E G D K H L I T N sag1408_515.pep
801 E A V S P K D Y Q K I T N K P I L T F E V V K G S I K N I I A V N K Q I S E Y H E E G D K H L I T N sag1408_18rs21.seq
796 E A V S P K D Y Q K I T N K P I L T F E V V K G S I Q N I I A V N K Q I S E Y H E E G D K H L I T N sag1408_cjb111.pep
801 E A V S P K D Y Q K I T N K P I L T F E V V K G S I Q N I I A V N K Q I S E Y H E E G D K H L I T N sag1408_h36b.pep
801 E A V S P K D Y Q K I T N K P I L T F E V V K G S I Q N I I A V N K Q I S E Y H E E G D K H L I T N sag1408_nem316.pep

    T H I P P K G I I P M T G G K G I L S F I L I G G A M M S I A G G I Y I W K R Y K K S S D M S I E K Majority
                                  860       870       880       890       900
851 T H I P P K G I I P M T G G K G I L S F I L I G G A M M S I A G G I Y I W K R Y K K S S D M S I K K sag1408_2603.pep
851 T H I P P K G I I P K T G G K G I L S F I L I G G A M M S I A G G I Y I W K R Y K K S S D M S I K K sag1408_515.pep
851 T H I P P K G I I P M T G G K G I L S F I L I G G A M M S I A G G I Y I W K R Y K K S S D M S I K K sag1408_18rs21.seq
846 T H I P P K G I I P M T G G K G I L S F I L I G G S M M S I A G G I Y I W K R Y K K S S D I S R E K sag1408_cjb111.pep
851 T H I P P K G I I P M T G G K G I L S F I L I G G A M M S I A G G I Y I W K R H K K S S D A S I E K sag1408_h36b.pep
851 T H I P P K G I I P M T G G K G I L S F I L I G G S M M S I A G G I Y I W K R Y K K S S D I S R E K sag1408_nem316.pep

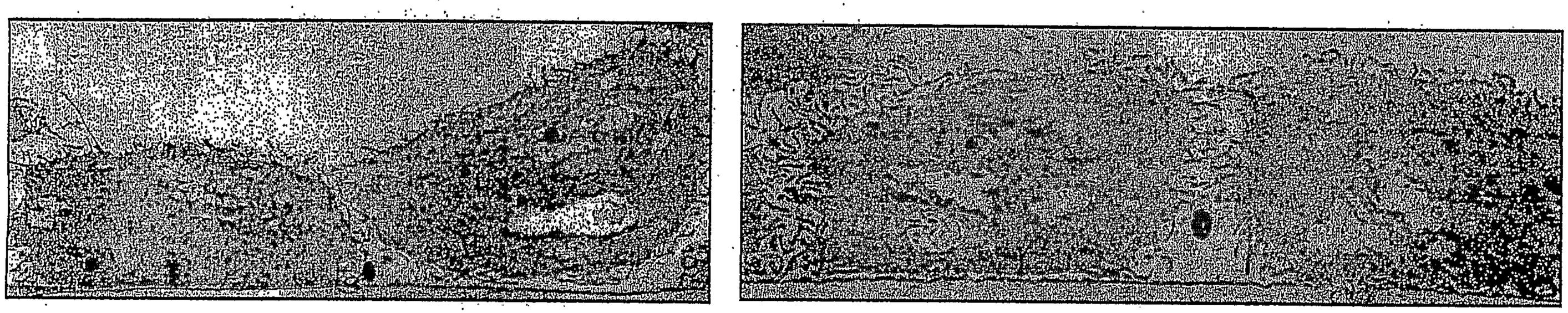
    D                                                                                 Majority

901 D
901 D                                                                                 sag1408_2603.pep
901 D                                                                                 sag1408_515.pep
901 D                                                                                 sag1408_18rs21.seq
896 D                                                                                 sag1408_cjb111.pep
901 D                                                                                 sag1408_h36b.pep
                                                                                      sag1408_nem316.pep
```

Decoration 'Decoration #1': Shade (with solid black) residues that differ from the Consensus.

FIGURE 24B

Figure 25: GBS closely associate with tight junctions and cross the monolayer by a paracellular route



Transmission Electron Microscopy images of GBS infection of ME180 cervical epithelial cells.

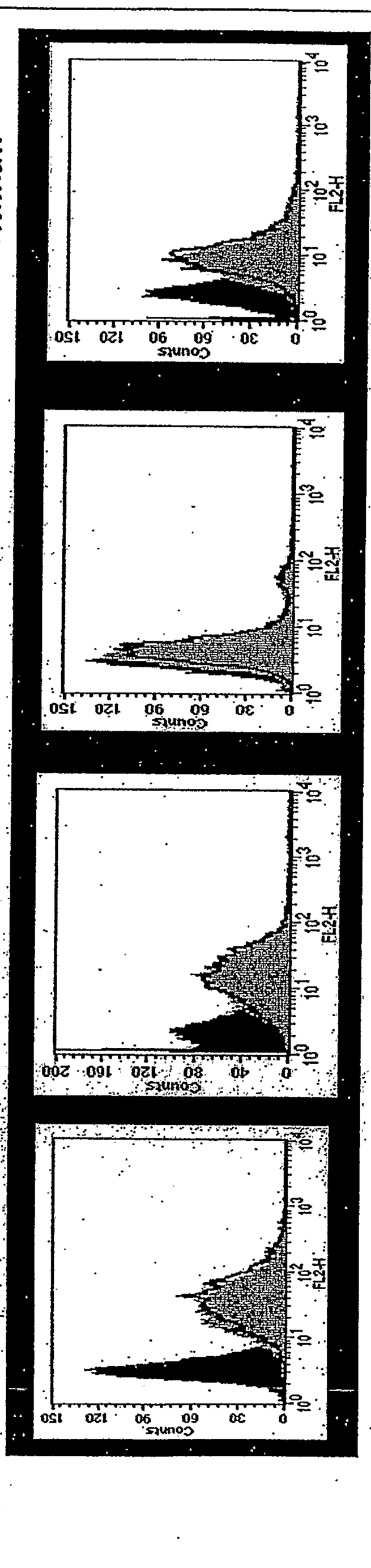
Figure 26: GBS infection of ME180 cells

GBS infection of ME180 cells



Figure 27

Name	Origin	Host	
ME180	Cervical	Human	
IB4HBE	Bronchial	Human	
A549	Lung	Human	
Caco2	Intestinal	Human	



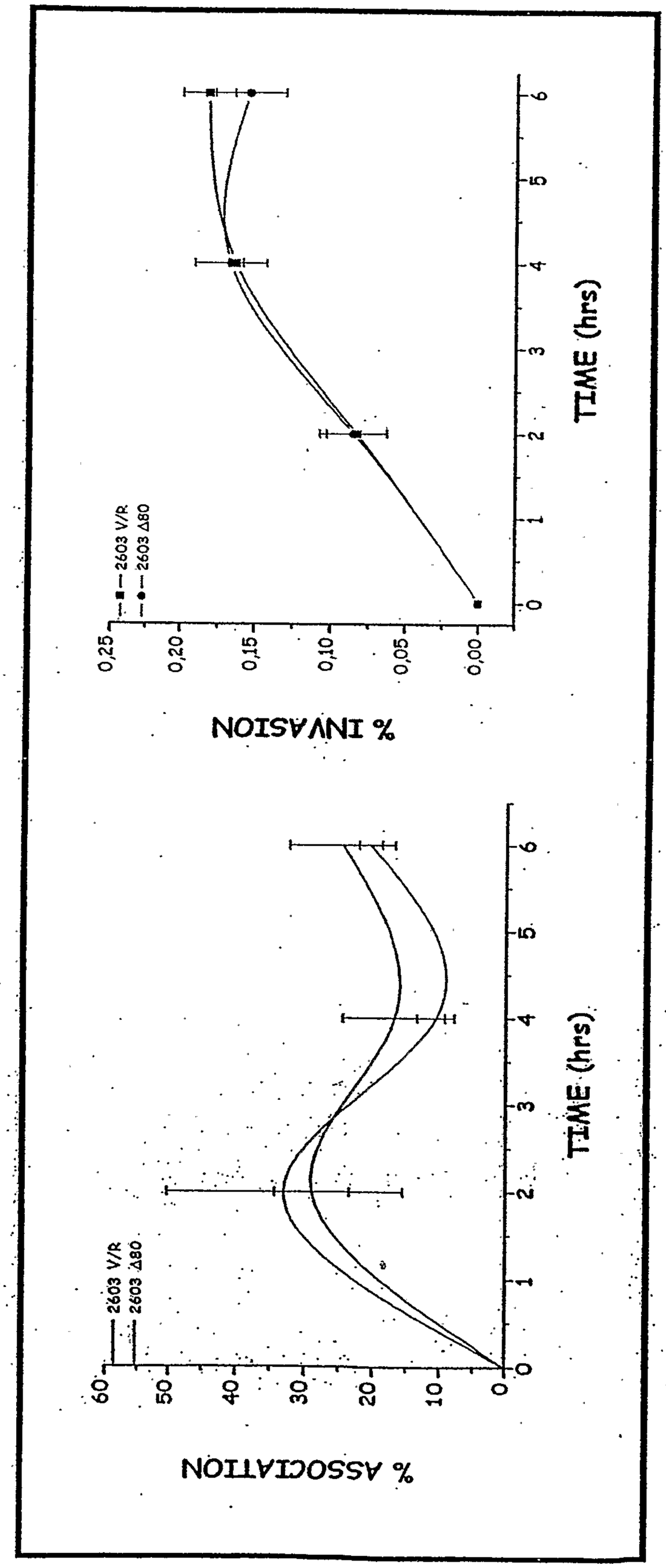
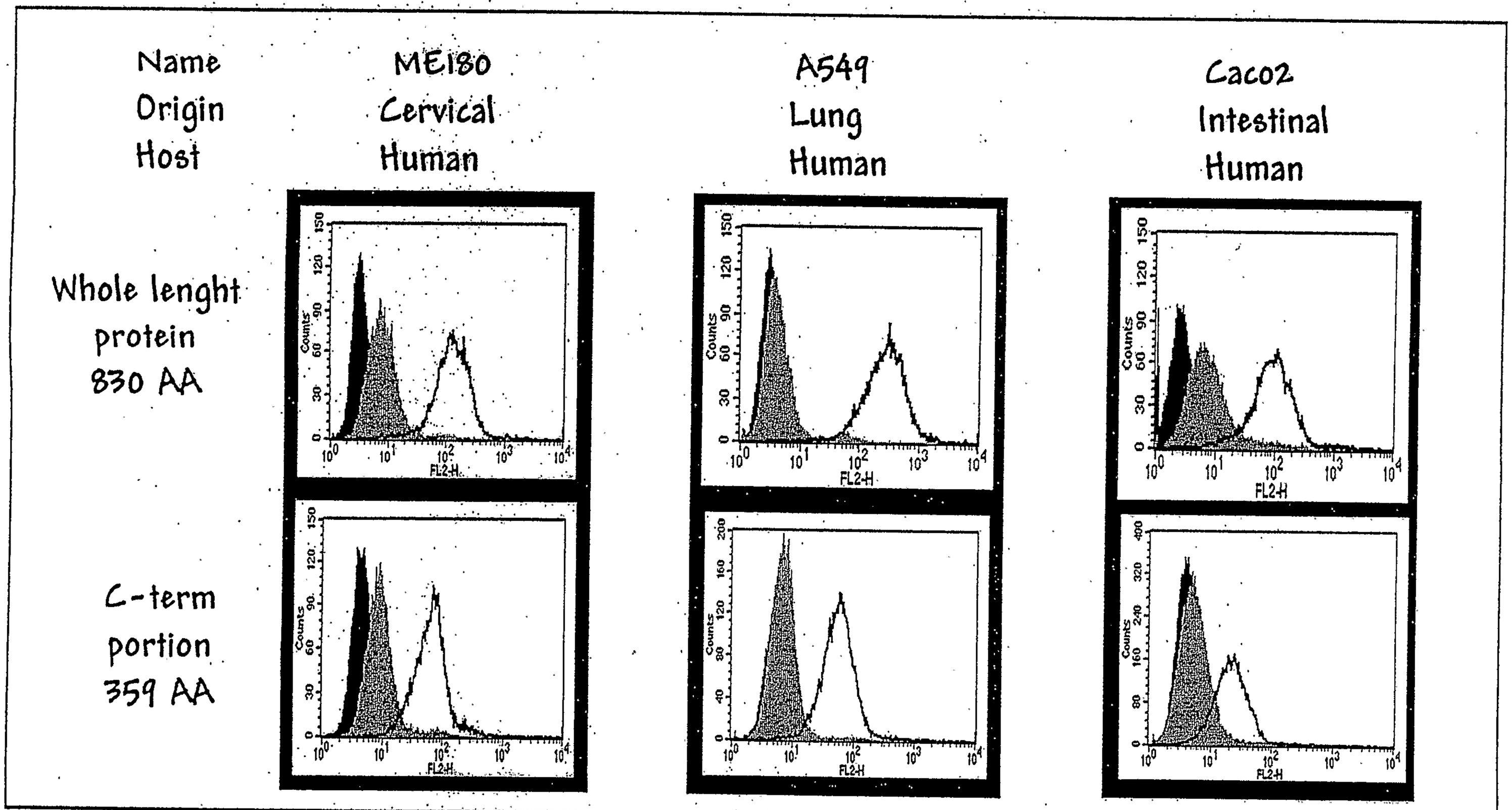

2603 V/R
2603 Δ80
0,25
0,20
0,15
0,10
0,05
0,00
% INVASION
0
1
2
3
4
5
6
TIME (hrs)
2603 V/R
2603 Δ80
60
50
40
30
20
10
0
% ASSOCIATION
0
1
2
3
4
5
6
TIME (hrs)

Figure 28

Figure 29


Name
Origin
Host
ME180
Cervical
Human
A549
Lung
Human
Caco2
Intestinal
Human
Whole lenght protein 830 AA
C-term portion 359 AA
Counts
FL2-H

PCT/US05/27239

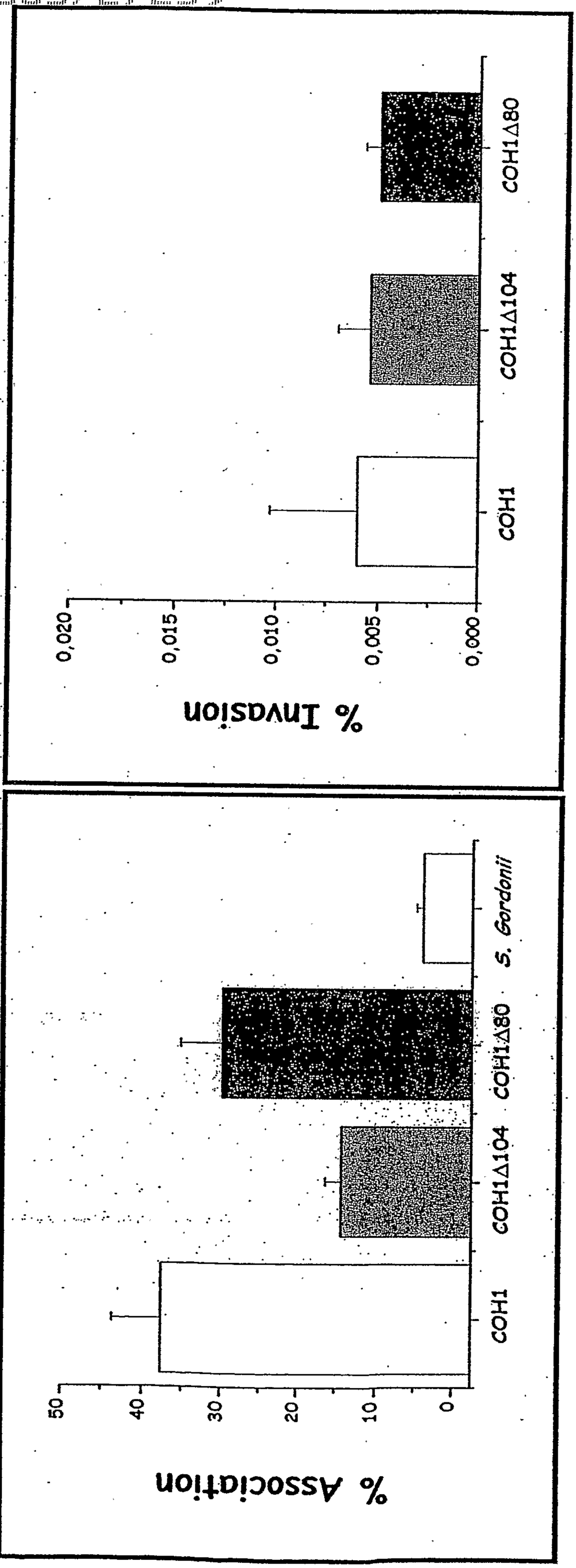


Figure 30

PCT/US2005/027239

Figure 31

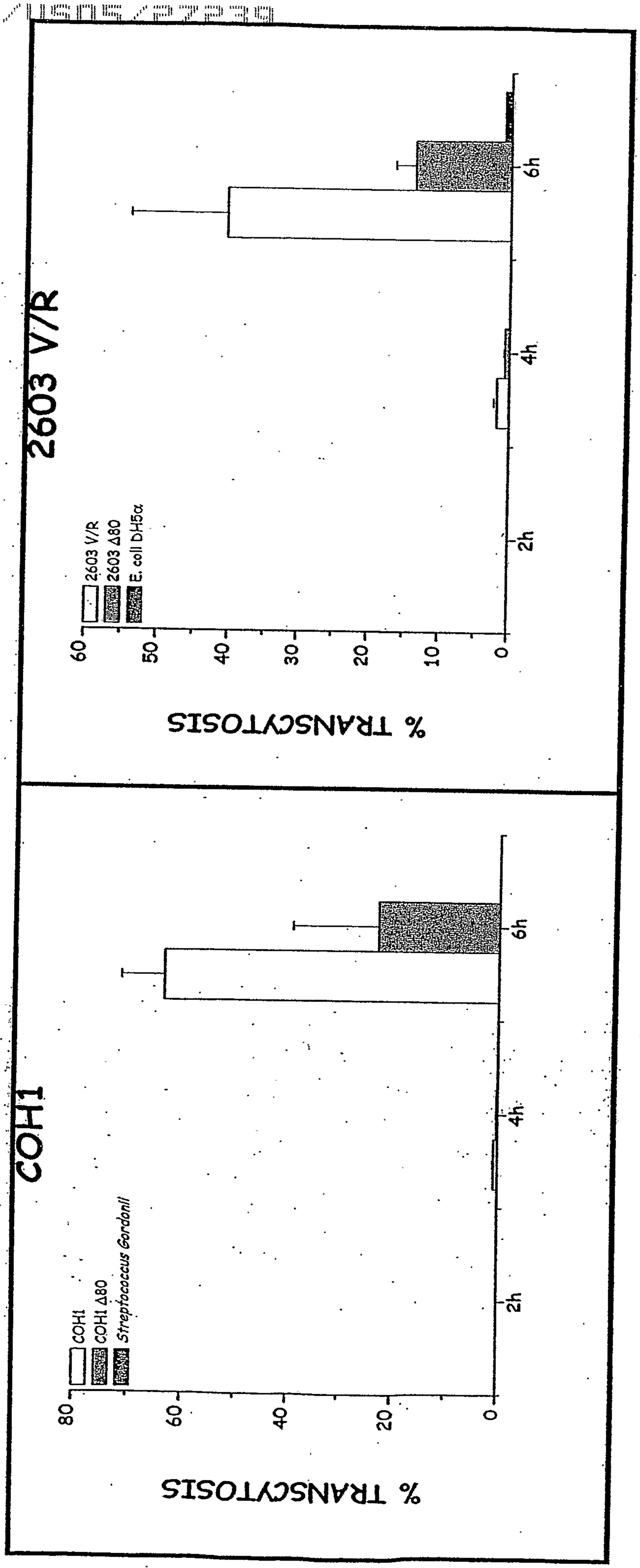


Figure 32

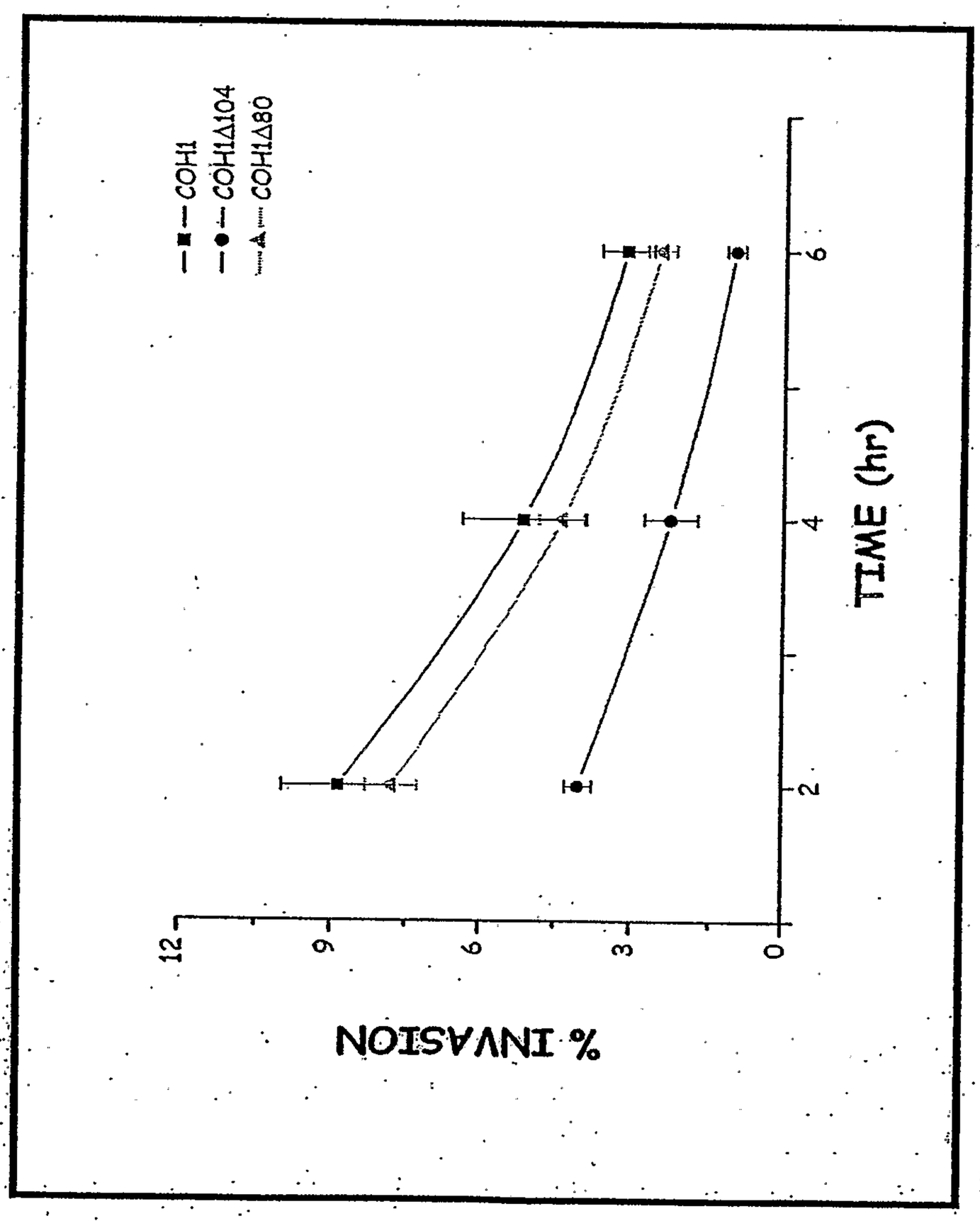
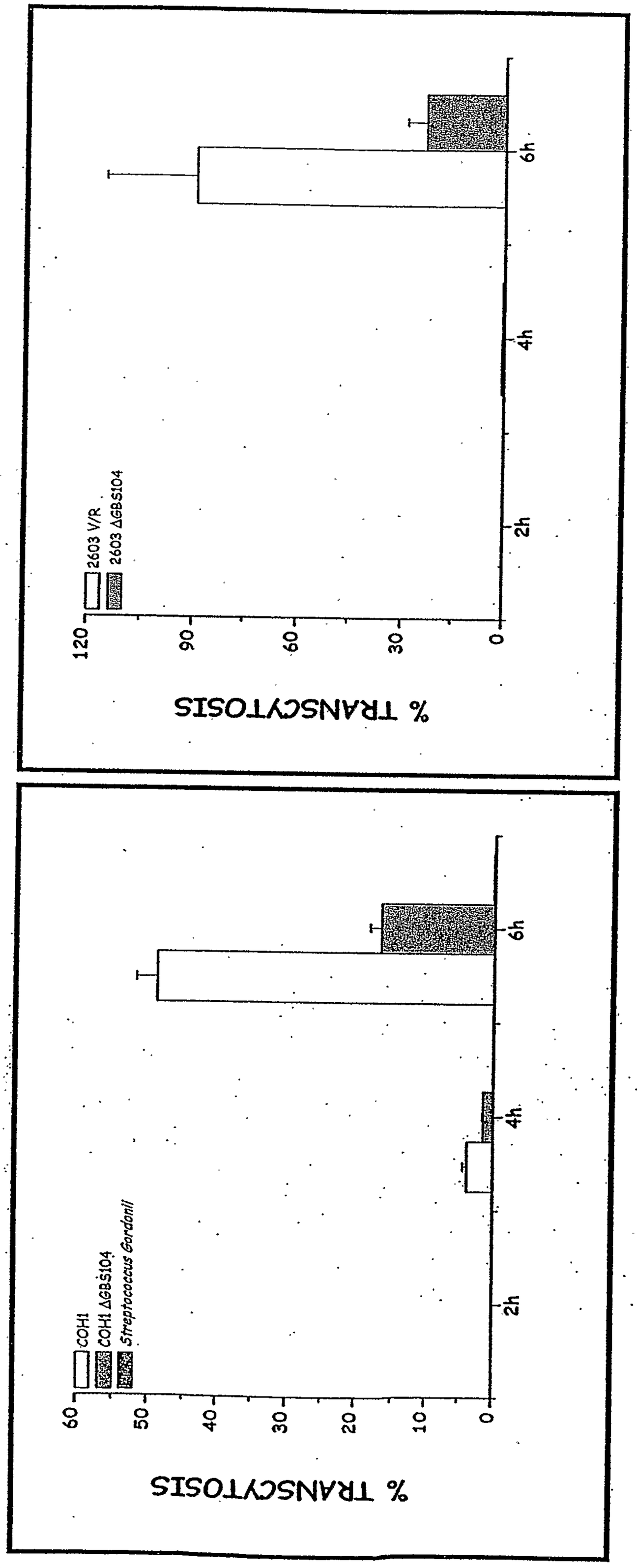


Figure 33

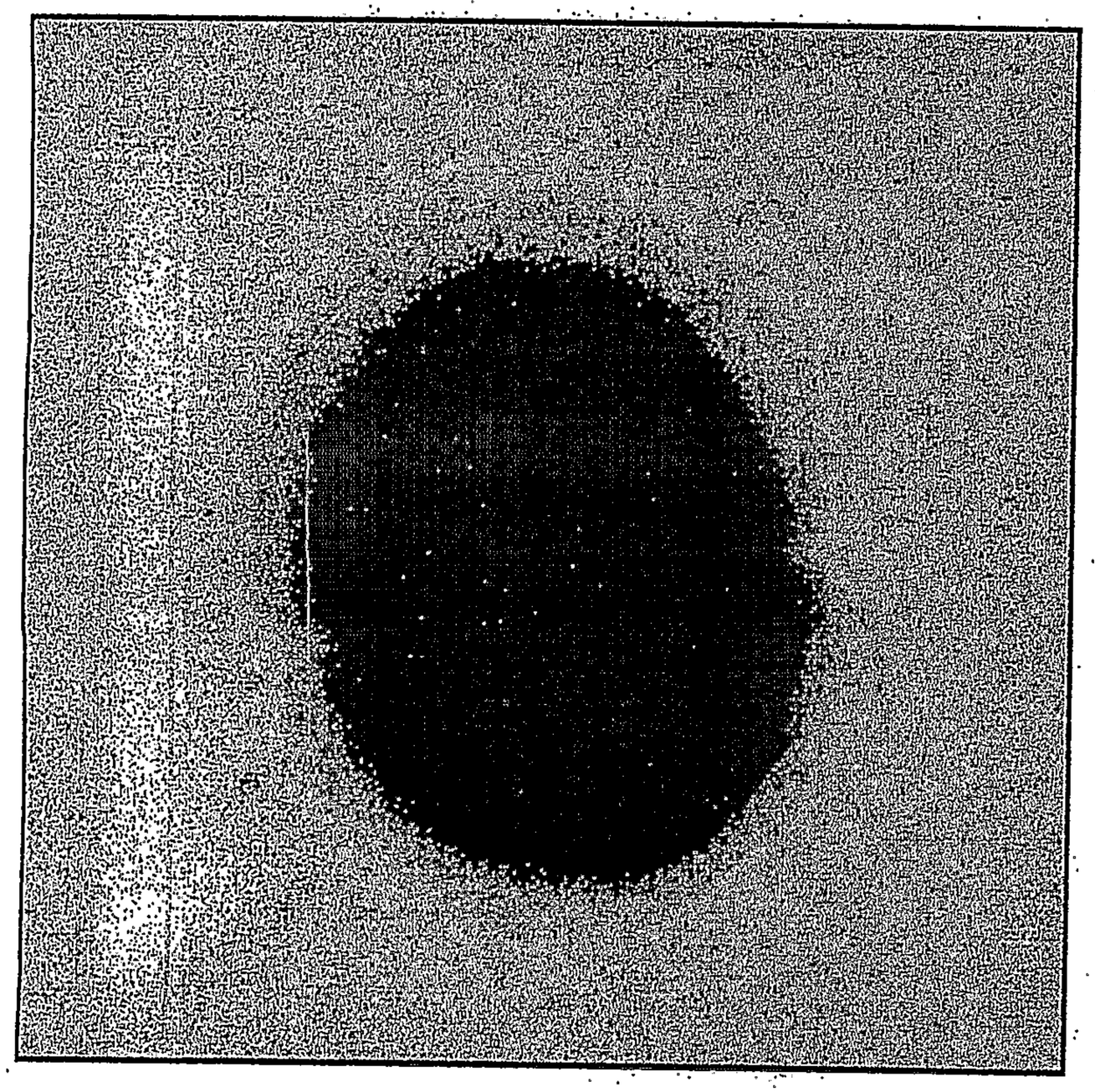

2603 V/R
2603 V/R
2603 ΔGBS104
120
90
60
30
0
% TRANSCYTOSIS
2h
4h
6h
COH1
COH1
COH1 ΔGBS104
Streptococcus Gordonii
60
50
40
30
20
10
0
% TRANSCYTOSIS
2h
4h
6h

WO 2006/078318 PCT/US2005/027239 CA 02575548 2007-01-29

Figure 34

GBS STRAIN COH1 over GBS80

Negative staining EM



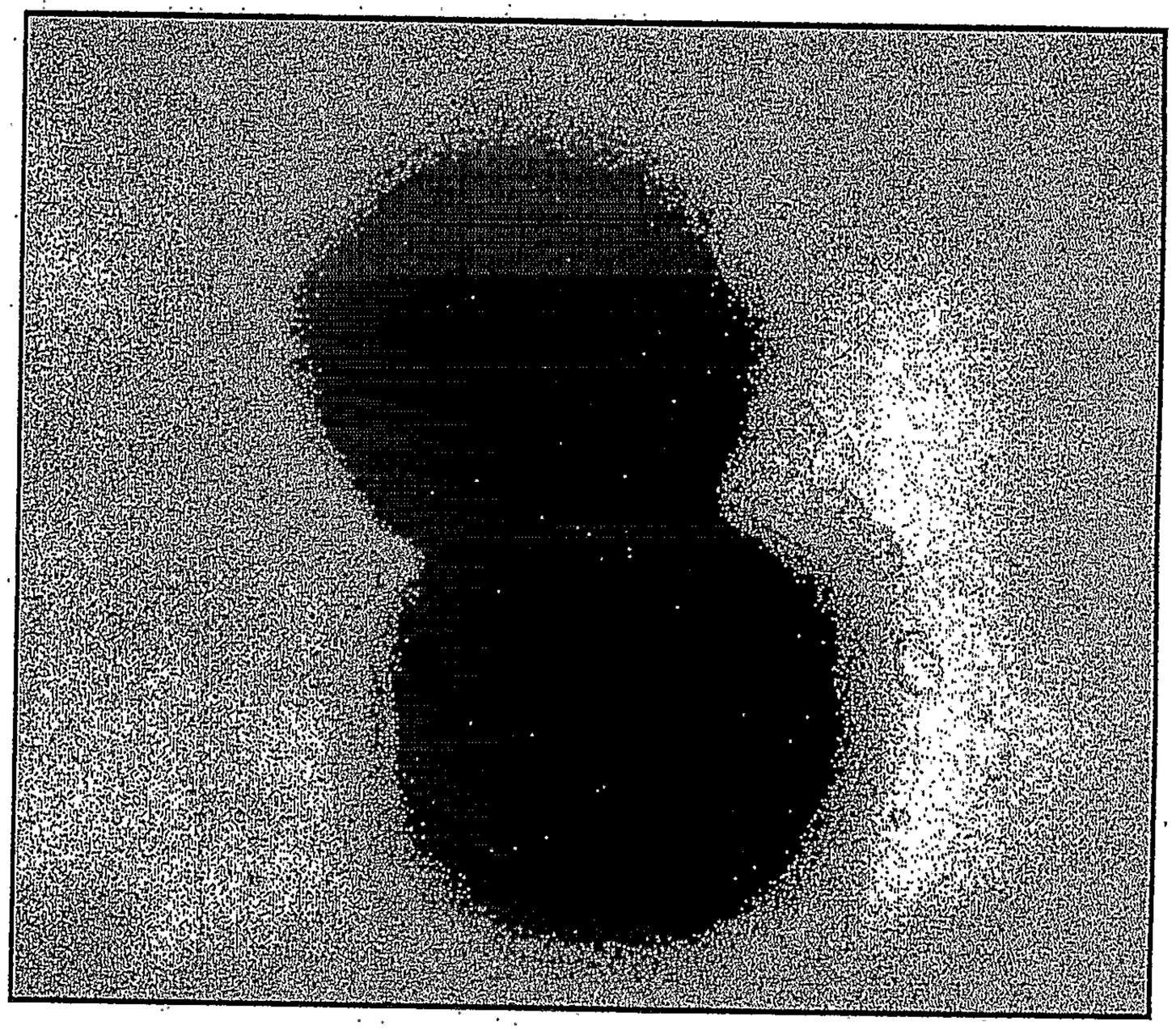
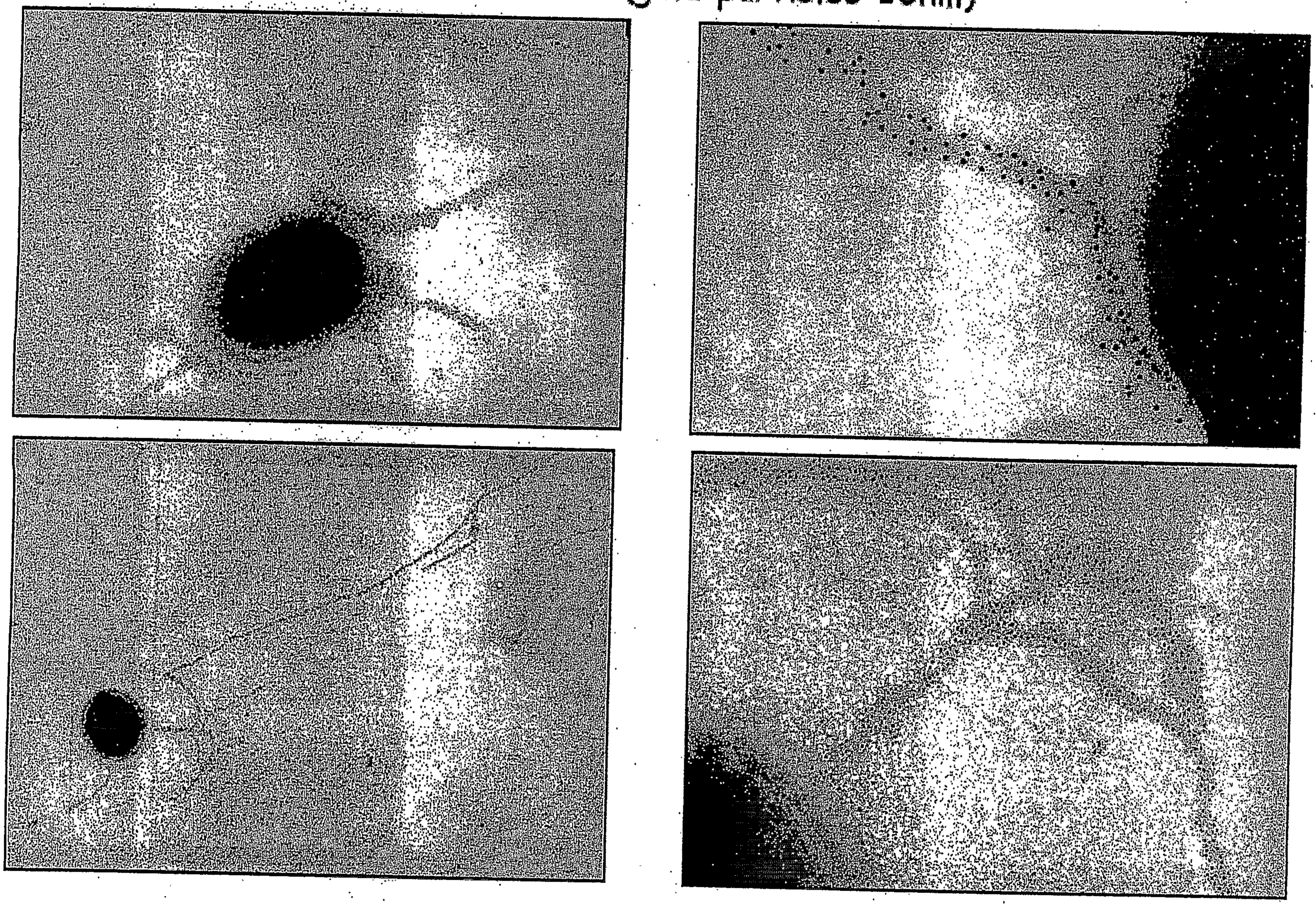


Figure 35

GBS STRAIN COH1 over GBS80

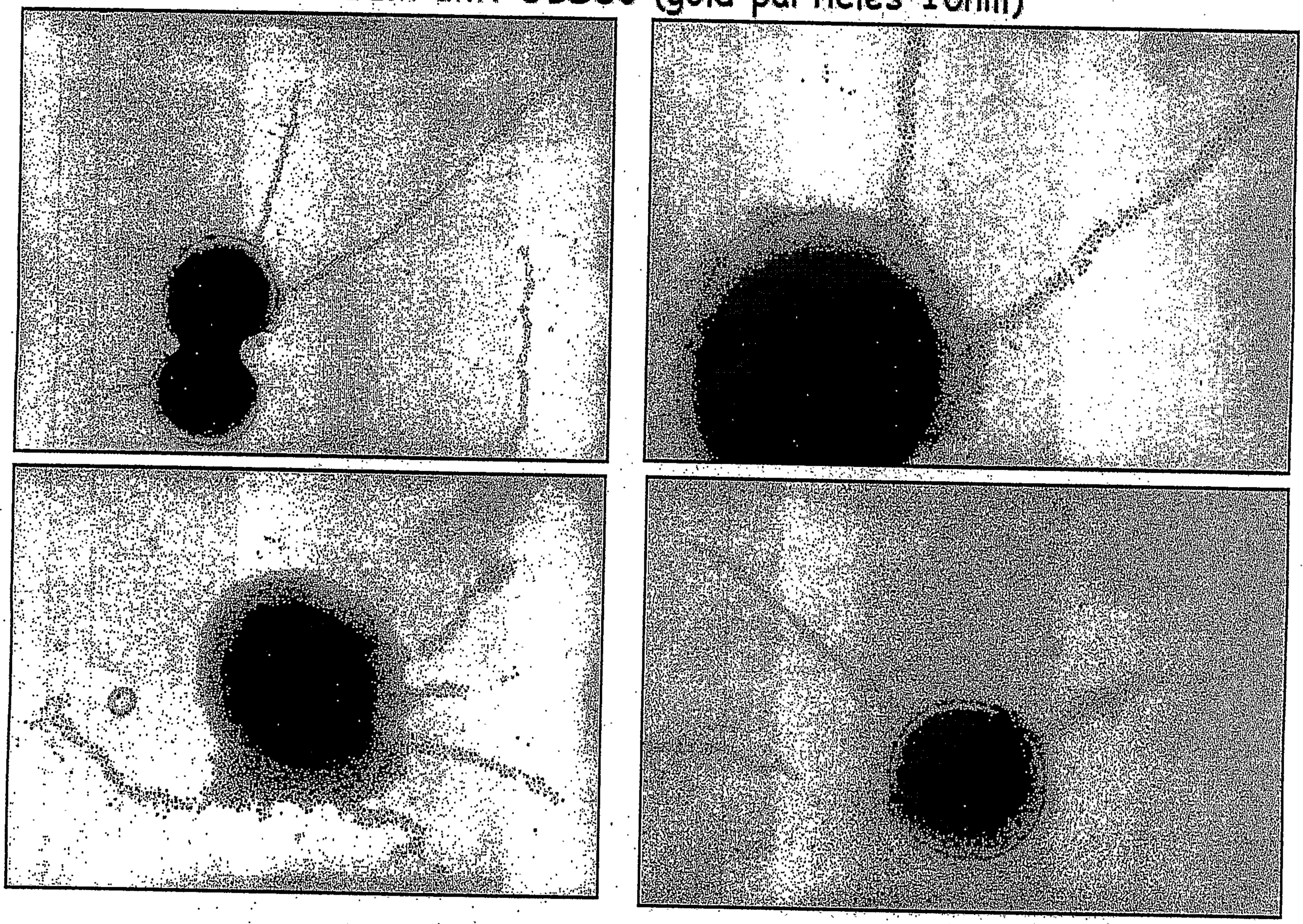
IEM anti-GBS80 (gold particles 10nm)



WO 2006/078318 PCT/US2005/045487 CA 02575548 2007-01-29 PCT/US2005/045487

GBS STRAIN COH1 over GBS80

Figure 36

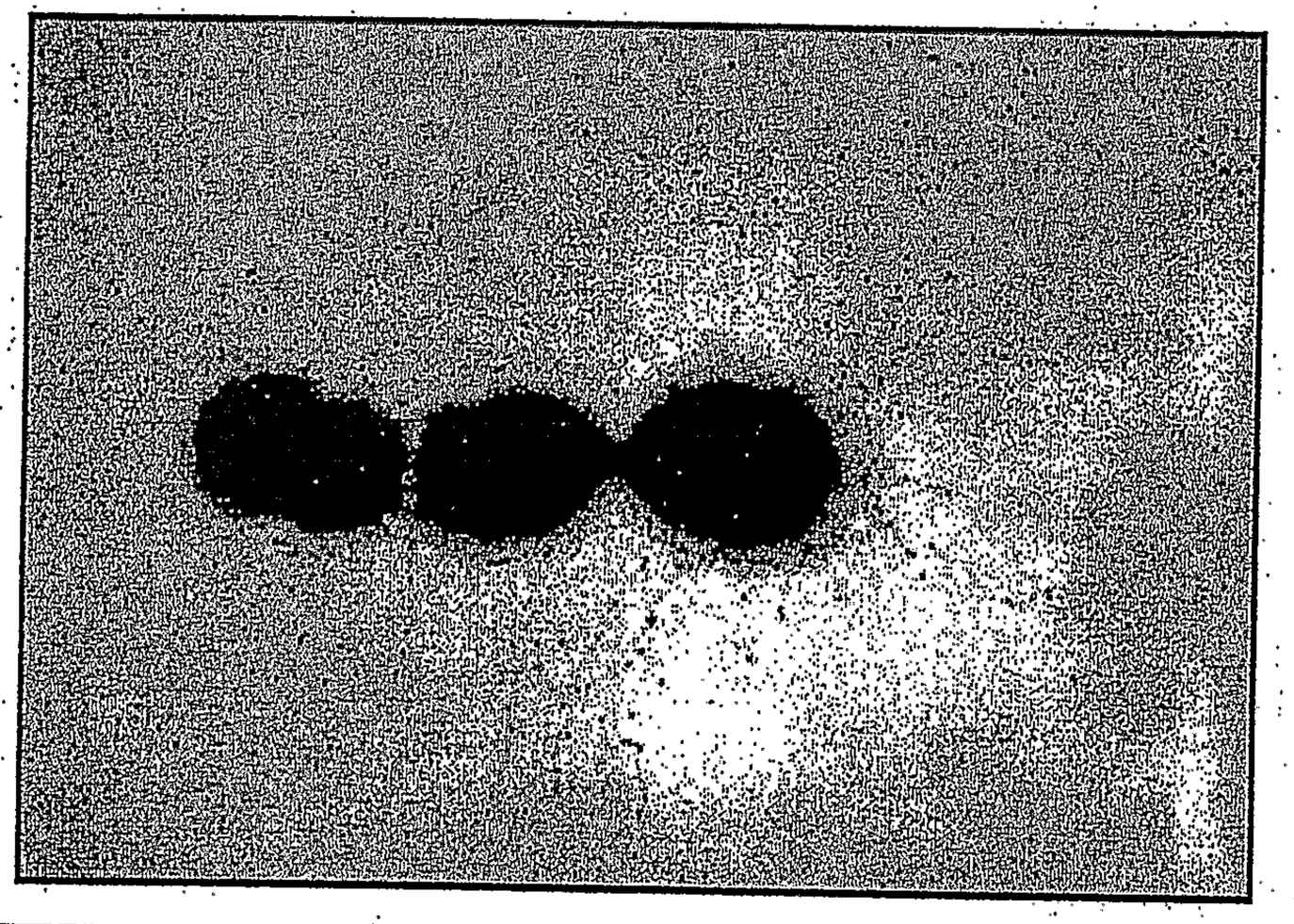


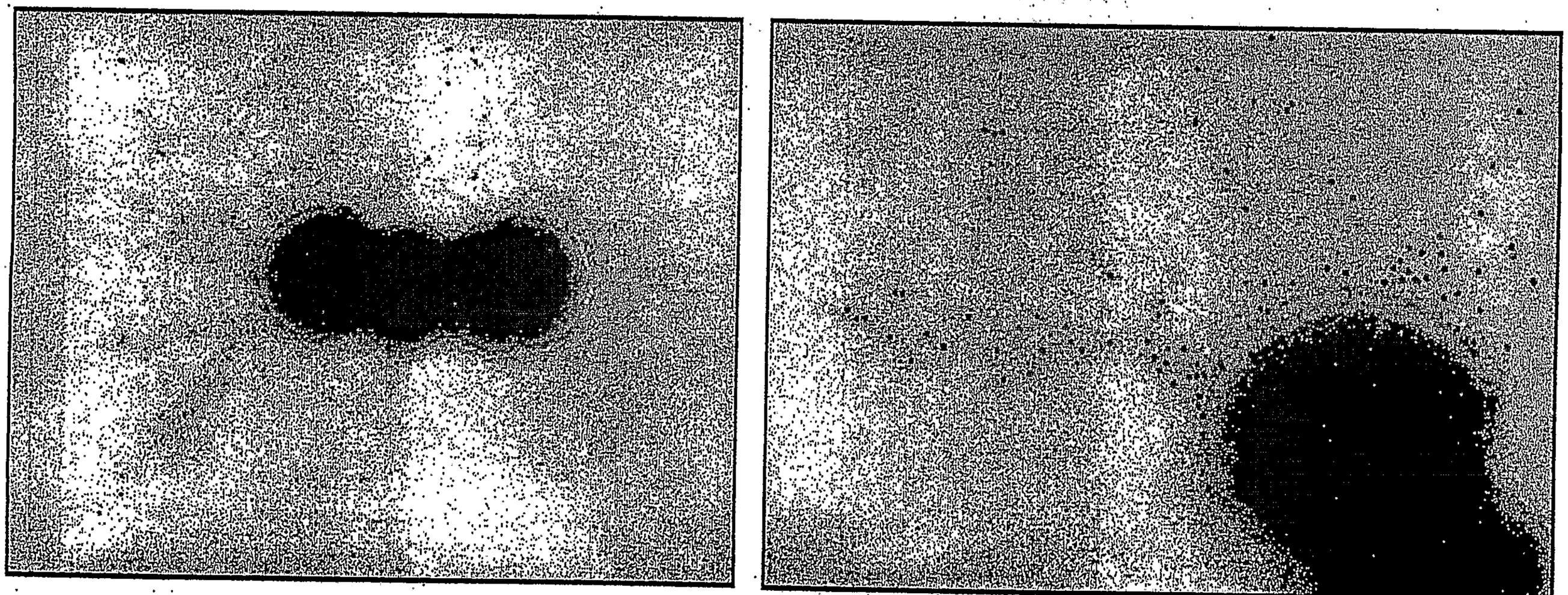
WO 2006/078318 PCT/US2005/027239 151/487 CA 02575548 2007-01-29 PCT/US2005/027239

Figure 37

GBS STRAIN COH1 over GBS80

IEM anti-GBS80 (gold particles 20nm)





DEMANDE OU BREVET VOLUMINEUX

LA PRÉSENTE PARTIE DE CETTE DEMANDE OU CE BREVET COMPREND PLUS D'UN TOME.

CECI EST LE TOME 1 DE 4
CONTENANT LES PAGES 1 À 152

NOTE : Pour les tomes additionels, veuillez contacter le Bureau canadien des brevets

JUMBO APPLICATIONS/PATENTS

THIS SECTION OF THE APPLICATION/PATENT CONTAINS MORE THAN ONE VOLUME

THIS IS VOLUME 1 OF 4
CONTAINING PAGES 1 TO 152

NOTE: For additional volumes, please contact the Canadian Patent Office

NOM DU FICHIER / FILE NAME :

NOTE POUR LE TOME / VOLUME NOTE: