

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
17 January 2008 (17.01.2008)

PCT

(10) International Publication Number
WO 2008/008553 A2

(51) International Patent Classification:

A01K 67/00 (2006.01) **C12N 15/00** (2006.01)
A61K 31/70 (2006.01) **C12N 5/00** (2006.01)
A61P 25/28 (2006.01) **C12Q 1/68** (2006.01)
C07H 21/04 (2006.01) **G01N 33/50** (2006.01)
C07K 16/00 (2006.01) **G01N 33/53** (2006.01)

(21) International Application Number:

PCT/US2007/016254

(22) International Filing Date: 16 July 2007 (16.07.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/831,153 14 July 2006 (14.07.2006) US
11/827,992 13 July 2007 (13.07.2007) US

(71) Applicant (for all designated States except US): **AP-PLERA** [US/US]; c/o Celera, An Applera Business Unit, 1401 Harbor Bay Parkway, Alameda, CA 94502 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **LI, Yonghong** [US/US]; 696 Towle Way, #25, Palo Alto, CA 94306 (US).

GRUPE, Andrew [US/US]; 27 Knickerbocker Lane, Orinda, California 94563 (US).

(74) Agent: **LEE, Victor**; Vice President And Chief Group Counsel, Celera, An Applera Business Unit, 1401 Harbor Bay Parkway, Alameda, CA 94502 (US).

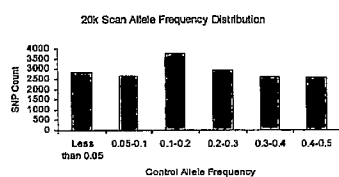
(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN,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

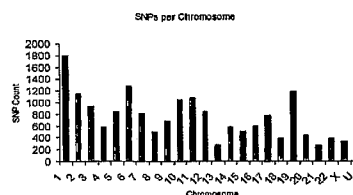
[Continued on next page]

(54) Title: GENETIC POLYMORPHISMS ASSOCIATED WITH ALZHEIMER'S DISEASE, METHODS OF DETECTION AND USES THEREOF

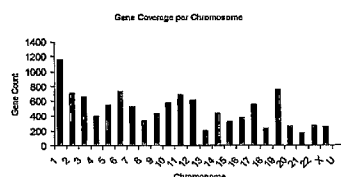
A.



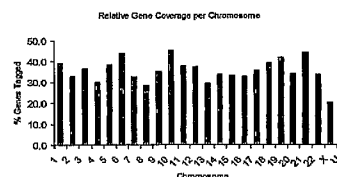
B.



C.



D.



(57) Abstract: The present invention is based on the discovery of genetic polymorphisms that are associated with Alzheimer's Disease. In particular, the present invention relates to nucleic acid molecules containing the polymorphisms, variant proteins encoded by such nucleic acid molecules, reagents for detecting the polymorphic nucleic acid molecules and proteins, and methods of using the nucleic acid and proteins as well as methods of using reagents for their detection.

WO 2008/008553 A2



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.

**GENETIC POLYMORPHISMS ASSOCIATED WITH ALZHEIMER'S DISEASE,
METHODS OF DETECTION AND USES THEREOF**

CROSS-REFERENCE TO RELATED APPLICATIONS

5 This application claims the benefit of U.S. provisional application Serial No.: 60/831,153, filed on July 14, 2006, and U.S. nonprovisional application, having the same Title and Inventors, with attorney docket number: CD000009ORD, filed on July 13, 2007, the contents of which are hereby incorporated by reference in its entirety into this application.

10 **FIELD OF THE INVENTION**

 The present invention is in the field of Alzheimer's Disease (AD) diagnosis and therapy. In particular, the present invention relates to specific single nucleotide polymorphisms (SNPs) in the human genome, and their association with AD and related pathologies. Based on differences in allele frequencies in the AD patient population relative to normal individuals, the naturally-
15 occurring SNPs disclosed herein can be used as targets for the design of diagnostic reagents and the development of therapeutic agents, as well as for disease association and linkage analysis. In particular, the SNPs of the present invention are useful for identifying an individual who is at an increased or decreased risk of developing AD and for early detection of the disease, for providing clinically important information for the prevention and/or treatment of AD, and for screening and
20 selecting therapeutic agents. The SNPs disclosed herein are also useful for human identification applications. Methods, assays, kits, and reagents for detecting the presence of these polymorphisms and their encoded products are provided.

BACKGROUND OF THE INVENTION

25 **NEURODEGENERATIVE DISEASES**

 A varied assortment of central nervous system disorders (neurodegenerative diseases) are associated with aging. Neurodegenerative diseases are characterized by a gradual and progressive loss of neural tissue or nerve cells. These diseases, directly or indirectly, affect millions of people worldwide. The number of individuals affected by neurodegenerative diseases
30 is anticipated to grow attendant with the increase in human life expectancy.

 Specific diseases exemplifying this class of disorders include: age-related dementia, such as AD, leukodystrophies, such as adrenoleukodystrophy, metachromatic leukodystrophy, Krabbe Disease (globoid cell leukodystrophy), Canavan Disease, Alexander Disease, Pelizaeus-Merzbacher Disease, and the like; and others such as neuronal ceroid lipofuscinoses,

amyotrophic lateral sclerosis (ALS, or Lou Gehrig's Disease), Huntington's Disease (HD), dentatorubral-pallidoluysian atrophy (DRPLA), stroke and the like.

Parkinson's Disease affects 1 to 2 percent of people over the age of 50 and 10 to 15% of those over 80. Huntington's Disease and ALS each afflict approximately 30,000 in the United States. Stroke is the leading cause of neurological impairment with half a million new stroke victims surviving each year with some degree of permanent neurological damage.

AD alone affects 20 million people worldwide. AD is the fourth leading cause of death in industrialized societies, afflicting 5-11% of the population over the age of 65 and 30% of those over the age of 85. AD is fast becoming the paramount healthcare problem as the world's geriatric population continues to grow.

AD is the most significant and common cause of dementia in developed countries, accounting for 60% or more of all cases of dementia. AD is a progressive neurodegenerative disorder characterized clinically by memory loss of subtle onset, followed by a slowly progressive dementia that has a course of several years. Brain pathology of AD is characterized by gross, diffuse atrophy of the cerebral cortex with secondary enlargement of the ventricular system. Microscopically, there are neuritic plaques containing A β amyloid, silver-staining neurofibrillary tangles in neuronal cytoplasm, and accumulation of A β amyloid in arterial walls of cerebral blood vessels. A definite diagnosis of AD can only occur at autopsy, where the presence of amyloid plaques and neurofibrillary tangles is confirmed.

The frequency of AD increases with each decade of adult life, reaching 20 to 40 percent of the population over the age of 85. Because more and more people will live into their 80's and 90's, the number of patients is expected to triple over the next 20 years. More than 4 million people suffer from AD in the USA, where 800,000 deaths per year are associated with AD. It is estimated that the cost of AD in the USA is \$80 billion to \$100 billion a year in medical care, personal caretaking and lost productivity. AD also puts a heavy emotional toll on family members and caregivers: about 2.7 million people care for AD patients in the USA. AD patients live for 7 to 10 years after diagnosis and spend an average of 5 years under care either at home or in a nursing home.

In spite of the high prevalence of AD today and its expected prevalence increase in an aging population, there are currently no diagnostic tests available that determine the cause of dementia and adequately differentiate between AD and other types of dementias. A diagnostic test that enables physicians to identify AD early in the disease process, or identify individuals who are at high risk of developing the disease, will provide the option to intervene at an early

stage in the disease process. Early intervention in disease processes does generally result in better treatment results by delaying disease onset or progression compared to later intervention.

AD is presumed to have a genetic component, as evidenced by an increased risk for AD among first degree relatives of affected individuals. So far, three genes have been identified in patients with early onset AD that lead to the less common, dominantly inherited form of dementia. Mutations in the three genes, beta-amyloid precursor protein (Goate *et al.*, *Nature* 1991, 349:704-706), presenilin 1 (Sherrington *et al.*, *Nature* 1995, 375:754-760), and presenilin 2 (Levy-Lahad *et al.*, *Science* 1996, 269:973-977) lead to an increase in the production of long amyloid beta (A β 42), the main component in amyloid plaques. Although early onset AD makes up less than 5% of all AD cases, the identification of these genes has contributed substantially to the understanding of the disease process.

Late onset Alzheimer's Disease (LOAD), the much more common form of this dementia, is inherited in a non-Mendelian pattern and involves genetic susceptibility factors and environmental factors. Early genetic studies of AD demonstrated association and linkage to the same region on chromosome 19 containing the ApoE gene (Schellenberg *et al.*, *J. Neurogenet.* 1987, 4:97-108, Pericak-Vance *et al.*, *Am. J. Hum. Gen.* 1991, 48:1034-1050). Three common alleles were identified for the ApoE gene, ϵ 2, ϵ 3, ϵ 4. The ϵ 4 allele frequency is increased to 50% in affected individuals vs. 14% in controls (Corder *et al.*, *Science* 1993, 281:921-923). Although there is strong association with the ApoE- ϵ 4 allele, which has been replicated in many studies, most investigators consider the ApoE- ϵ 4 allele to be neither necessary nor sufficient for the development of AD. ApoE is considered a major risk factor, but ApoE testing does not provide enough sensitivity and specificity for use as an independent diagnostic test and therefore is not recommended as a diagnostic marker for the prediction of AD (National Institute on Aging/Alzheimer's Association Working Group, 1996).

Genome-wide linkage screens in LOAD patients, duplicated in at least 2 studies, identified regions on four chromosomes, chromosomes 6, 9, 10, and 12 (reviewed by: Myers and Goate, *Curr. Op. Neurol.* 2001, 14:433-440, Lendon and Craddock, *TINS* 2001, 24:557-559), implying that other genetic risk factors besides ApoE must exist. Co-localization of a quantitative trait for A β 42 and a susceptibility locus for LOAD on chromosome 10, for example, suggests the locus influences LOAD risk through increased levels of the A β 42 peptide (Ertekin-Taner, *Science* 2000, 290:2303-2304).

The majority of the putative LOAD susceptibility loci were identified through linkage studies of affected sib pairs (ASPs) by looking for regions with increased allele sharing. In order to identify the genes and mutations for LOAD, it would be beneficial to conduct association

studies, which have relatively better power than linkage studies to detect genes of modest or small effect. Association studies compare unrelated cases to controls and analyze allele frequency differences between affected and unaffected individuals.

Obviously, there is a definite need for novel diagnostic markers that enable the detection of AD at an early stage of the disease. The availability of a genetic test will also provide a non-invasive method to assess an individual's risk for developing AD. Furthermore, there is an urgent need for new and improved treatments for AD to prevent or significantly delay the onset of the disease, or to reverse or slow down disease progression after onset.

SNPs

The genomes of all organisms undergo spontaneous mutation in the course of their continuing evolution, generating variant forms of progenitor genetic sequences (Gusella, *Ann. Rev. Biochem.* 55, 831-854 [1986]). A variant form may confer an evolutionary advantage or disadvantage relative to a progenitor form or may be neutral. In some instances, a variant form confers an evolutionary advantage to the species and is eventually incorporated into the DNA of many or most members of the species and effectively becomes the progenitor form. Additionally, the effects of a variant form may be both beneficial and detrimental, depending on the circumstances. For example, a heterozygous sickle cell mutation confers resistance to malaria, but a homozygous sickle cell mutation is usually lethal. In many cases, both progenitor and variant forms survive and co-exist in a species population. The coexistence of multiple forms of a genetic sequence gives rise to genetic polymorphisms, including SNPs.

Approximately 90% of all polymorphisms in the human genome are SNPs. SNPs are single base positions in DNA at which different alleles, or alternative nucleotides, exist in a population. The SNP position (interchangeably referred to herein as SNP, SNP site, SNP locus, SNP marker, or marker) is usually preceded by and followed by highly conserved sequences of the allele (e.g., sequences that vary in less than 1/100 or 1/1000 members of the populations). An individual may be homozygous or heterozygous for an allele at each SNP position. A SNP can, in some instances, be referred to as a "cSNP" to denote that the nucleotide sequence containing the SNP is an amino acid coding sequence.

A SNP may arise from a substitution of one nucleotide for another at the polymorphic site. Substitutions can be transitions or transversions. A transition is the replacement of one purine nucleotide by another purine nucleotide, or one pyrimidine by another pyrimidine. A transversion is the replacement of a purine by a pyrimidine, or vice versa. A SNP may also be a

single base insertion or deletion variant referred to as an “indel” (Weber *et al.*, “Human diallelic insertion/deletion polymorphisms,” *Am. J. Hum. Genet.* 71[4]:854-62 [Oct. 2002]).

A synonymous codon change, or silent mutation/SNP (terms such as “SNP,” “polymorphism,” “mutation,” “mutant,” “variation,” and “variant” are used herein

5 interchangeably), is one that does not result in a change of amino acid due to the degeneracy of the genetic code. A substitution that changes a codon coding for one amino acid to a codon coding for a different amino acid (i.e., a non-synonymous codon change) is referred to as a missense mutation. A nonsense mutation results in a type of non-synonymous codon change in which a stop codon is formed, thereby leading to premature termination of a polypeptide chain
10 and a truncated protein. A read-through mutation is another type of non-synonymous codon change that causes the destruction of a stop codon, thereby resulting in an extended polypeptide product. While SNPs can be bi-, tri-, or tetra-allelic, the vast majority of SNPs are bi-allelic, and are thus often referred to as “bi-allelic markers,” or “di-allelic markers.”

As used herein, references to SNPs and SNP genotypes include individual SNPs and/or
15 haplotypes, which are groups of SNPs that are generally inherited together. Haplotypes can have stronger correlations with diseases or other phenotypic effects compared with individual SNPs, and therefore may provide increased diagnostic accuracy in some cases (Stephens *et al.*, *Science* 293, 489-493 [20 July 2001]).

Causative SNPs are those SNPs that produce alterations in gene expression or in the
20 expression, structure, and/or function of a gene product, and therefore are most predictive of a possible clinical phenotype. One such class includes SNPs falling within regions of genes encoding a polypeptide product, i.e. cSNPs. These SNPs may result in an alteration of the amino acid sequence of the polypeptide product (i.e., non-synonymous codon changes) and give rise to the expression of a defective or other variant protein. Furthermore, in the case of nonsense
25 mutations, a SNP may lead to premature termination of a polypeptide product. Such variant products can result in a pathological condition, e.g. genetic disease. Examples of genes in which a SNP within a coding sequence causes a genetic disease include sickle cell anemia and cystic fibrosis.

Causative SNPs do not necessarily occur in coding regions; causative SNPs can occur in,
30 for example, any genetic region that can ultimately affect the expression, structure, and/or activity of the protein encoded by a nucleic acid. Such genetic regions include, for example, those involved in transcription, such as SNPs in transcription factor binding domains, SNPs in promoter regions, in areas involved in transcript processing, such as SNPs at intron-exon boundaries that may cause defective splicing, or SNPs in mRNA processing signal sequences

such as polyadenylation signal regions. Some SNPs that are not causative SNPs nevertheless are in close association with, and therefore segregate with, a disease-causing sequence. In this situation, the presence of a SNP correlates with the presence of, or predisposition to, or an increased risk in developing the disease. These SNPs, although not causative, are nonetheless
5 also useful for diagnostics, disease predisposition screening, and other uses.

An association study of a SNP and a specific disorder involves determining the presence or frequency of the SNP allele in biological samples from individuals with the disorder of interest, such as AD, and comparing the information to that of controls (i.e., individuals who do not have the disorder; controls may be also referred to as "healthy" or "normal" individuals) who
10 are preferably of similar age and race. The appropriate selection of patients and controls is important to the success of SNP association studies. Therefore, a pool of individuals with well-characterized phenotypes is extremely desirable.

A SNP may be screened in diseased tissue samples or any biological sample obtained from a diseased individual, and compared to control samples, and selected for its increased (or
15 decreased) occurrence in a specific pathological condition, such as pathologies related to AD. Once a statistically significant association is established between one or more SNPs and a pathological condition (or other phenotype) of interest, then the regions around the SNPs can optionally be thoroughly screened to identify the causative genetic locus or sequences (e.g., the causative SNP/mutation, gene, regulatory region, etc.) that influences the pathological condition
20 or phenotype. Association studies may be conducted within the general population and are not limited to studies performed on related individuals in affected families (linkage studies).

Clinical trials have shown that patient response to treatment with pharmaceuticals is often heterogeneous. There is a continuing need to improve pharmaceutical agent design and therapy. In that regard, SNPs can be used to identify patients most suited to therapy with particular
25 pharmaceutical agents (this is often termed "pharmacogenomics"). Similarly, SNPs can be used to exclude patients from certain treatments due to the patient's increased likelihood of developing toxic side effects or his [their] likelihood of not responding to the treatment. Pharmacogenomics can also be used in pharmaceutical research to assist the drug development and selection process (Linder *et al.*, *Clinical Chemistry* 43, 254 [1997]; Marshall, *Nature Biotechnology* 15, 1249
30 [1997]; International Patent Application WO 97/40462, Spectra Biomedical; and Schafer *et al.*, *Nature Biotechnology* 16, 3 [1998]).

SUMMARY OF THE INVENTION

The present invention relates to the identification of novel SNPs, unique combinations of such SNPs, and haplotypes of SNPs that are associated with AD and related neurological pathologies. The polymorphisms disclosed herein are directly useful as targets for the design of diagnostic reagents, as druggable targets in the development of therapeutic agents for use in the treatment and diagnosis of AD and also other neurological pathologies.

Based on the identification of SNPs associated with AD, the present invention also provides methods of detecting these variants as well as the design and preparation of detection reagents needed to accomplish this task. The invention specifically provides, for example, novel SNPs in genetic sequences involved in AD, isolated nucleic acid molecules (including, for example, DNA and RNA molecules) containing these SNPs, variant proteins encoded by nucleic acid molecules containing such SNPs, antibodies to the encoded variant proteins, computer-based and data storage systems containing the novel SNP information, methods of detecting these SNPs in a test sample, methods of identifying individuals who have an altered (i.e., increased or decreased) risk of developing AD based on the presence or absence of one or more particular nucleotides (alleles) at one or more SNP sites disclosed herein or the detection of one or more encoded variant products (e.g., variant mRNA transcripts or variant proteins), methods of identifying individuals who are more or less likely to respond to a treatment (or more or less likely to experience undesirable side effects from a treatment, etc.), methods of screening for compounds useful in the treatment of a disorder associated with a variant gene/protein, compounds identified by these methods, methods of treating disorders mediated by a variant gene/protein, methods of using the novel SNPs of the present invention for human identification, etc.

In Tables 1-2, the present invention provides gene information, transcript sequences (SEQ ID NOS:1-48), encoded amino acid sequences (SEQ ID NOS:48-96), genomic sequences (SEQ ID NOS:148-176), transcript-based context sequences (SEQ ID NOS:97-147) and genomic-based context sequences (SEQ ID NOS:177-254) that contain the SNPs of the present invention, and extensive SNP information that includes observed alleles, allele frequencies, populations/ethnic groups in which alleles have been observed, information about the type of SNP and corresponding functional effect, and, for cSNPs, information about the encoded polypeptide product. The transcript sequences (SEQ ID NOS:1-48), encoded amino acid sequences (SEQ ID NOS:49-96), genomic sequences (SEQ ID NOS:148-176), transcript-based context sequences (SEQ ID NOS:97-147) and genomic-based context sequences (SEQ ID NOS:177-254) are also provided in the Sequence Listing.

In a specific embodiment of the present invention, SNPs which occur naturally in the human genome are provided as isolated nucleic acid molecules. These SNPs are associated with AD such that they can have a variety of uses in the diagnosis and/or treatment of AD and related pathologies. One aspect of the present invention relates to an isolated nucleic acid molecule comprising a nucleotide sequence in which at least one nucleotide is a SNP disclosed in Table(s) 3 and/or 4. In an alternative embodiment, a nucleic acid of the invention is an amplified polynucleotide, which is produced by amplification of a SNP-containing nucleic acid template. In another embodiment, the invention provides for a variant protein which is encoded by a nucleic acid molecule containing a SNP disclosed herein.

In yet another embodiment of the invention, a reagent for detecting a SNP in the context of its naturally-occurring flanking nucleotide sequences (which can be, e.g., either DNA or mRNA) is provided. In particular, such a reagent may be in the form of, for example, a hybridization probe or an amplification primer that is useful in the specific detection of a SNP of interest. In an alternative embodiment, a protein detection reagent is used to detect a variant protein which is encoded by a nucleic acid molecule containing a SNP disclosed herein. A preferred embodiment of a protein detection reagent is an antibody or an antigen-reactive antibody fragment.

Various embodiments of the invention also provide kits comprising SNP detection reagents, and methods for detecting the SNPs disclosed herein by employing detection reagents. In a specific embodiment, the present invention provides for a method of identifying an individual having an increased or decreased risk of developing AD by detecting the presence or absence of one or more SNP alleles disclosed herein. In another embodiment, a method for diagnosis of AD by detecting the presence or absence of one or more SNP alleles disclosed herein is provided.

The nucleic acid molecules of the invention can be inserted in an expression vector, such as to produce a variant protein in a host cell. Thus, the present invention also provides for a vector comprising a SNP-containing nucleic acid molecule, genetically-engineered host cells containing the vector, and methods for expressing a recombinant variant protein using such host cells.

In another specific embodiment, the host cells, SNP-containing nucleic acid molecules, and/or variant proteins can be used as targets in a method for screening and identifying therapeutic agents or pharmaceutical compounds useful in the treatment of AD and other neurological pathologies.

One aspect of this invention is a method for treating AD in a human subject wherein said subject harbors a SNP, gene, transcript, and/or encoded protein identified in Tables 1-2, which method comprises administering to said subject a therapeutically or prophylactically effective amount of one or more agents counteracting the effects of the disease, such as by inhibiting (or stimulating) the activity of the gene, transcript, and/or encoded protein identified in Tables 1-2.

Another aspect of this invention is a method for identifying an agent useful in therapeutically or prophylactically treating AD in a human subject wherein said subject harbors a SNP, gene, transcript, and/or encoded protein identified in Tables 1-2, which method comprises contacting the gene, transcript, or encoded protein with a candidate agent under conditions suitable to allow formation of a binding complex between the gene, transcript, or encoded protein and the candidate agent and detecting the formation of the binding complex, wherein the presence of the complex identifies said agent.

Another aspect of this invention is a method for treating AD in a human subject, which method comprises:

(i) determining that said subject harbors a SNP, gene, transcript, and/or encoded protein identified in Tables 1-2, and

(ii) administering to said subject a therapeutically or prophylactically effective amount of one or more agents counteracting the effects of the disease.

Many other uses and advantages of the present invention will be apparent to those skilled in the art upon review of the detailed description of the preferred embodiments herein. Solely for clarity of discussion, the invention is described in the sections below by way of non-limiting examples.

DESCRIPTION OF THE FILES CONTAINED ON THE CD-R NAMED CD000009PCT CDR

The CD-R named CD000009PCT CDR contains the following text (ASCII) file:

File SEQLIST_CD000009PCT.txt provides the Sequence Listing. The Sequence Listing provides the transcript sequences (SEQ ID NOS:1-48) and protein sequences (SEQ ID NOS:49-96) as shown in Table 1, and genomic sequences (SEQ ID NOS:148-176) as shown in Table 2, for each AD-associated gene that contains one or more SNPs of the present invention. Also provided in the Sequence Listing are context sequences flanking each SNP, including both transcript-based context sequences as shown in Table 1 (SEQ ID NOS:97-147) and genomic-based context sequences as shown in Table 2 (SEQ ID NOS:177-254). The context sequences generally provide 100bp upstream (5') and 100bp downstream (3') of each SNP, with the SNP in

the middle of the context sequence, for a total of 200bp of context sequence surrounding each SNP. File SEQLIST_CD000009PCT.txt is 2,507 KB in size, and was created on July 16, 2007.

The material contained on the CD-R labeled CD000009PCT CDR is hereby incorporated by reference pursuant to 37 CFR 1.77(b)(4).

5

DESCRIPTION OF TABLE 1 AND TABLE 2

Table 1 and Table 2 (both provided on the CD-R) disclose the SNP and associated gene/transcript/protein information of the present invention. For each gene, Table 1 and Table 2 each provides a header containing gene/transcript/protein information, followed by a transcript
10 and protein sequence (in Table 1) or genomic sequence (in Table 2), and then SNP information regarding each SNP found in that gene/transcript.

NOTE: SNPs may be included in both Table 1 and Table 2; Table 1 presents the SNPs relative to their transcript sequences and encoded protein sequences, whereas Table 2 presents the SNPs relative to their genomic sequences (in some instances Table 2 may also include, after
15 the last gene sequence, genomic sequences of one or more intergenic regions, as well as SNP context sequences and other SNP information for any SNPs that lie within these intergenic regions). SNPs can readily be cross-referenced between Tables based on their hCV (or, in some instances, hDV) identification numbers.

The gene/transcript/protein information includes:

- 20 - a gene number (1 through n, where n = the total number of genes in the Table);
- a Celera hCG and UID internal identification numbers for the gene;
- a Celera hCT and UID internal identification numbers for the transcript (Table 1 only);
- a public Genbank accession number (e.g., RefSeq NM number) for the transcript (Table
1 only);
- 25 - a Celera hCP and UID internal identification numbers for the protein encoded by the hCT transcript (Table 1 only);
- a public Genbank accession number (e.g., RefSeq NP number) for the protein (Table 1 only);
- an art-known gene symbol;
- 30 - an art-known gene/protein name;
- Celera genomic axis position (indicating start nucleotide position-stop nucleotide position);
- the chromosome number of the chromosome on which the gene is located;

- an OMIM (Online Mendelian Inheritance in Man, Johns Hopkins University/NCBI) public reference number for obtaining further information regarding the medical significance of each gene; and

- the alternative gene/protein name(s) and/or symbol(s) in the OMIM entry.

5 NOTE: Due to the presence of alternative splice forms, multiple transcript/protein entries can be provided for a single gene entry in Table 1; i.e., for a single Gene Number, multiple entries may be provided in series that differ in their transcript/protein information and sequences.

Following the gene/transcript/protein information is a transcript sequence and protein sequence (in Table 1), or a genomic sequence (in Table 2), for each gene, as follows:

10 - transcript sequence (Table 1 only) (corresponding to SEQ ID NOS:1-48 of the Sequence Listing), with SNPs identified by their IUB codes (transcript sequences can include 5' UTR, protein coding, and 3' UTR regions);

NOTE: If there are differences between the nucleotide sequence of the hCT transcript and the corresponding public transcript sequence identified by the Genbank accession number, the hCT transcript sequence (and encoded protein) is provided, unless the public sequence is a RefSeq transcript sequence identified by an NM number, in which case the RefSeq NM transcript sequence (and encoded protein) is provided. However, whether the hCT transcript or RefSeq NM transcript is used as the transcript sequence, the disclosed SNPs are represented by their IUB codes within the transcript.

20 - the encoded protein sequence (Table 1 only) (corresponding to SEQ ID NOS:49-96 of the Sequence Listing); and

- the genomic sequence of the gene (Table 2 only), including 6kb on each side of the gene boundaries (i.e., 6kb on the 5' side of the gene plus 6kb on the 3' side of the gene) (corresponding to SEQ ID NOS:148-176 of the Sequence Listing).

25 After the last gene sequence, Table 2 may include additional genomic sequences of intergenic regions (in such instances, these sequences are identified by "Intergenic region:" followed by a numerical identification number), as well as SNP context sequences and other SNP information for any SNPs that lie within each intergenic region (and such SNPs are identified as "INTERGENIC" for SNP type).

30 NOTE: The transcript, protein, and transcript-based SNP context sequences are provided in both Table 1 and in the Sequence Listing. The genomic and genomic-based SNP context sequences are provided in both Table 2 and in the Sequence Listing. SEQ ID NOS are indicated in Table 1 for each transcript sequences (SEQ ID NOS:1-48), encoded amino acid sequences (SEQ ID NOS:49-96), genomic sequences (SEQ ID NOS:148-176), transcript-based context

sequences (SEQ ID NOS:97-147) and genomic-based context sequences (SEQ ID NOS:177-254).

The SNP information includes:

- context sequence (taken from the transcript sequence in Table 1, and taken from the genomic sequence in Table 2) with the SNP represented by its IUB code, including 100 bp upstream (5') of the SNP position plus 100 bp downstream (3') of the SNP position (the transcript-based SNP context sequences in Table 1 are provided in the Sequence Listing as SEQ ID NOS:97-147; the genomic-based SNP context sequences in Table 2 are provided in the Sequence Listing as SEQ ID NOS:177-254);

- Celera hCV internal identification number for the SNP (in some instances, an "hDV" number is given instead of an "hCV" number);

- SNP position (position of the SNP within the given transcript sequence [Table 1] or within the given genomic sequence [Table 2]);

- SNP source (may include any combination of one or more of the following five codes, depending on which internal sequencing projects and/or public databases the SNP has been observed in: "Applera" = SNP observed during the re-sequencing of genes and regulatory regions of 39 individuals, "Celera" = SNP observed during shotgun sequencing and assembly of the Celera human genome sequence, "Celera Diagnostics" = SNP observed during re-sequencing of nucleic acid samples from individuals who have AD or a related pathology, "dbSNP" = SNP observed in the dbSNP public database, "HGBASE" = SNP observed in the HGBASE public database, "HGMD" = SNP observed in the Human Gene Mutation Database [HGMD] public database, "HapMap" = SNP observed in the International HapMap Project public database, "CSNP" = SNP observed in an internal Applied Biosystems [Foster City, CA] database of coding SNPs [cSNPs]);

NOTE: multiple "Applera" source entries for a single SNP indicate that the same SNP was covered by multiple overlapping amplification products and the re-sequencing results (e.g., observed allele counts) from each of these amplification products is being provided.

- Population/allele/allele count information in the format of [population1(first_allele,count|second_allele,count)population2(first_allele,count|second_allele,count) total (first_allele,total count|second_allele,total count)]. The information in this field includes populations/ethnic groups in which particular SNP alleles have been observed ("cau" = Caucasian, "his" = Hispanic, "chn" = Chinese, and "afri" = African-American, "jpn" = Japanese, "ind" = Indian, "mex" = Mexican, "ain" = "American Indian, "cra" = Celera donor, "no_pop" = no population information available), identified SNP alleles, and observed allele counts (within

each population group and total allele counts), where available [“-“ in the allele field represents a deletion allele of an insertion/deletion (“indel”) polymorphism (in which case the corresponding insertion allele, which may be comprised of one or more nucleotides, is indicated in the allele field on the opposite side of the “|”); “-“ in the count field indicates that allele count information is not available]. For certain SNPs from the public dbSNP database, population/ethnic information is indicated as follows (this population information is publicly available in dbSNP): “HISP1” = human individual DNA (anonymized samples) from 23 individuals of self-described HISPANIC heritage; “PAC1” = human individual DNA (anonymized samples) from 24 individuals of self-described PACIFIC RIM heritage; “CAUC1” = human individual DNA (anonymized samples) from 31 individuals of self-described CAUCASIAN heritage; “AFR1” = human individual DNA (anonymized samples) from 24 individuals of self-described AFRICAN/AFRICAN AMERICAN heritage; “P1” = human individual DNA (anonymized samples) from 102 individuals of self-described heritage; “PA130299515”; “SC_12_A” = SANGER 12 DNAs of Asian origin from Coriell cell repositories, 6 of which are male and 6 female; “SC_12_C” = SANGER 12 DNAs of Caucasian origin from Coriell cell repositories from the CEPH/UTAH library. Six male and 6 female; “SC_12_AA” = SANGER 12 DNAs of African-American origin from Coriell cell repositories 6 of which are male and 6 female; “SC_95_C” = SANGER 95 DNAs of Caucasian origin from Coriell cell repositories from the CEPH/UTAH library; and “SC_12_CA” = Caucasians - 12 DNAs from Coriell cell repositories that are from the CEPH/UTAH library. Six male and 6 female;

NOTE: For SNPs of “Applera” SNP source, genes/regulatory regions of 39 individuals (20 Caucasians and 19 African Americans) were re-sequenced and, since each SNP position is represented by two chromosomes in each individual (with the exception of SNPs on X and Y chromosomes in males, for which each SNP position is represented by a single chromosome), up to 78 chromosomes were genotyped for each SNP position. Thus, the sum of the African-American (“afr”) allele counts is up to 38, the sum of the Caucasian allele counts (“cau”) is up to 40, and the total sum of all allele counts is up to 78.

NOTE: semicolons separate population/allele/count information corresponding to each indicated SNP source; i.e., if four SNP sources are indicated, such as “Celera,” “dbSNP,” “HGBASE,” and “HGMD,” then population/allele/count information is provided in four groups which are separated by semicolons and listed in the same order as the listing of SNP sources, with each population/allele/count information group corresponding to the respective SNP source based on order; thus, in this example, the first population/allele/count information group would correspond to the first listed SNP source (Celera) and the third population/allele/count

information group separated by semicolons would correspond to the third listed SNP source (HGBASE); if population/allele/count information is not available for any particular SNP source, then a pair of semicolons is still inserted as a place-holder in order to maintain correspondence between the list of SNP sources and the corresponding listing of population/allele/count information.

- SNP type (e.g., location within gene/transcript and/or predicted functional effect) (“MIS-SENSE MUTATION” = SNP causes a change in the encoded amino acid [i.e., a non-synonymous coding SNP]; “SILENT MUTATION” = SNP does not cause a change in the encoded amino acid [i.e., a synonymous coding SNP]; “STOP CODON MUTATION” = SNP is located in a stop codon; “NONSENSE MUTATION” = SNP creates or destroys a stop codon; “UTR 5” = SNP is located in a 5’ UTR of a transcript; “UTR 3” = SNP is located in a 3’ UTR of a transcript; “PUTATIVE UTR 5” = SNP is located in a putative 5’ UTR; “PUTATIVE UTR 3” = SNP is located in a putative 3’ UTR; “DONOR SPLICE SITE” = SNP is located in a donor splice site [5’ intron boundary]; “ACCEPTOR SPLICE SITE” = SNP is located in an acceptor splice site [3’ intron boundary]; “CODING REGION” = SNP is located in a protein-coding region of the transcript; “EXON” = SNP is located in an exon; “INTRON” = SNP is located in an intron; “hmCS” = SNP is located in a human-mouse conserved segment; “TFBS” = SNP is located in a transcription factor binding site; “UNKNOWN” = SNP type is not defined; “INTERGENIC” = SNP is intergenic, i.e., outside of any gene boundary); and

- Protein coding information (Table 1 only), where relevant, in the format of (protein SEQ ID NO: #, amino acid position, [amino acid-1, codon1] [amino acid-2, codon2]). The information in this field includes SEQ ID NO of the encoded protein sequence, position of the amino acid residue within the protein identified by the SEQ ID NO that is encoded by the codon containing the SNP, amino acids (represented by one-letter amino acid codes) that are encoded by the alternative SNP alleles (in the case of stop codons, “X” is used for the one-letter amino acid code), and alternative codons containing the alternative SNP nucleotides which encode the amino acid residues (thus, for example, for missense mutation-type SNPs, at least two different amino acids and at least two different codons are generally indicated; for silent mutation-type SNPs, one amino acid and at least two different codons are generally indicated, etc.). In instances where the SNP is located outside of a protein-coding region (e.g., in a UTR region), “None” is indicated following the protein SEQ ID NO.

DESCRIPTION OF TABLE 3 AND TABLE 4

Tables 3 and 4 provide a list of a subset of SNPs from Table 1 (in the case of Table 3) or Table 2 (in the case of Table 4) for which the SNP source falls into one of the following three categories: 1) SNPs for which the SNP source is only "Applera" and none other, 2) SNPs for which the SNP source is only "Celera" and none other, and 3) SNPs for which the SNP source is both "Applera" and "Celera" but none other.

These SNPs have not been observed in any of the public databases (dbSNP, HGBASE, and HGMD), and were also not observed during shotgun sequencing and assembly of the Celera human genome sequence (i.e., "Celera" SNP source). Tables 3 and 4 provide the hCV identification number (or hDV identification number for SNPs having "Celera" SNP source) and the SEQ ID NO of the context sequence for each of these SNPs.

DESCRIPTION OF TABLE 5

Table 5 provides sequences (SEQ ID NOS: 255-311) of primers that have been synthesized and used in the laboratory to carry out allele-specific PCR reactions in order to assay the SNPs disclosed in Tables 1 and 2 during the course of association studies to verify the association of these SNPs with AD.

Table 5 provides the following:

- the column labeled "Marker" provides an hCV identification number for each SNP site
- the column labeled "Alleles" designates the two alternative alleles at the SNP site

identified by the hCV identification number that are targeted by the allele-specific primers (the allele-specific primers are shown as "Sequence A" and "Sequence B") [NOTE: Alleles may be presented in Table 5 based on a different orientation (i.e., the reverse complement) relative to how the same alleles are presented in Tables 1-4, 9 and 13].

- the column labeled "Primer 1 (allele-specific primer)" provides an allele-specific primer that is specific for an allele designated in the "Alleles" column

- the column labeled "Primer 2 (allele-specific primer)" provides an allele-specific primer that is specific for the other allele designated in the "Alleles" column

- the column labeled "Common Primer" provides a common primer that is used in conjunction with each of the allele-specific primers (the "Sequence A" primer and the "Sequence B" primer) and which hybridizes at a site away from the SNP position.

All primer sequences are given in the 5' to 3' direction.

Each of the nucleotides designated in the "Alleles" column matches or is the reverse complement of (depending on the orientation of the primer relative to the designated allele) the

3' nucleotide of the allele-specific primer (either "Sequence A" or "Sequence B") that is specific for that allele.

DESCRIPTION OF TABLES 6-13

5 Tables 6-13, and in particular, Table 9, show the statistical results of the association of the SNPs in Tables 1-2 SNPs with AD. For more description of the tables, see the Example Section.

For example, Table 9 shows the significant markers from a meta-analysis of 5 LOAD case control sample sets and their association with LOAD. Here, "rs numbers" refers to the publicly known reference ID number for the SNP. "Chr" refers to the Chromosome on which the SNP is located. "Position (bp)" refers to the location measured in base pairs on a chromosome where the SNP is found. UK1, UK2, UK3, WU, SD refer to case-control sample sets from Cardiff University (UK1 and UK2), from Washington University (WU) and from University of California at San Diego (SD), as described in the Example Section below. "OR" refers to odds ratio. "Full Sample" refers to the combination of all five LOAD sample sets.

DESCRIPTION OF THE FIGURES

Figure 1 shows the characteristics of the 17,343 discovery SNPs. (A) Allele frequency distribution of 17,343 SNPs in control pools of the UK1 sample set. (B) Number of SNPs that were tested per chromosome. (C) Number of genes per chromosome that were tagged with at least one SNP. (D) Percentage of total genes per chromosome (Celera R27 annotation) that were tagged with at least one SNP. U: Chromosome is unknown

Figure 2 shows the discovery sample results and follow up strategy. (A) Allelic P-values for 17,339 SNPs in DNA pools of the UK1 sample set. SNPs are plotted according to chromosome and position. Four SNPs (rs1331212 [chr. 6], rs157581 [chr. 19], rs405509 [chr. 19], hCV3098650 [chr. 19, 57.7Mbp]) with allelic P-values <0.00001 have been omitted from the graph. U: chromosome or position on a chromosome is unknown; (B) Number of unique SNPs that were tested in the specified sample sets. Allele frequency refers to data derived from DNA pools.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides SNPs associated with AD, nucleic acid molecules containing these SNPs, methods and reagents for the detection of the SNPs disclosed herein, uses of these SNPs for the development of detection reagents, and assays or kits that utilize such reagents. The AD-associated SNPs disclosed herein are useful for diagnosing, screening for, and

evaluating predisposition to AD and other neurological pathologies in humans. Furthermore, such SNPs and their encoded products are useful targets for the development of therapeutic agents in treating AD and other neurological pathologies.

A large number of SNPs have been identified from re-sequencing DNA from 39 individuals, and they are indicated as "Applera" SNP source in Tables 1-2. Their allele frequencies, observed in each of the Caucasian and African-American ethnic groups, are provided. Additional SNPs included herein were previously identified during shotgun sequencing and assembly of the human genome, and they are indicated as "Celera" SNP source in Tables 1-2. Furthermore, the information provided in Table 1-2, particularly the allele frequency information obtained from 39 individuals and the identification of the precise position of each SNP within each gene/transcript, allows haplotypes (i.e., groups of SNPs that are co-inherited) to be readily inferred. The present invention encompasses SNP haplotypes, as well as individual SNPs.

Thus, the present invention provides individual SNPs associated with AD, as well as combinations of SNPs and haplotypes in genetic regions associated with AD, polymorphic/variant transcript sequences (SEQ ID NOS:1-48), encoded amino acid sequences (SEQ ID NOS:49-96), genomic sequences (SEQ ID NOS:148-176), transcript-based context sequences (SEQ ID NOS:97-147) and genomic-based context sequences (SEQ ID NOS:177-254) (transcript sequences, protein sequences, and transcript-based SNP context sequences are provided in Table 1 and the Sequence Listing; genomic sequences and genomic-based SNP context sequences are provided in Table 2 and the Sequence Listing), methods of detecting these polymorphisms in a test sample, methods of determining the risk of an individual of having or developing AD, methods of screening for compounds useful for treating neurological pathologies such as AD associated with a variant gene/protein, compounds identified by these screening methods, methods of using the disclosed SNPs to select a treatment strategy, methods of treating a disorder associated with a variant gene/protein (i.e., therapeutic methods), and methods of using the SNPs of the present invention for human identification.

The present invention provides novel SNPs associated with AD, as well as SNPs that were previously known in the art, but were not previously known to be associated with AD. Accordingly, the present invention provides novel compositions and methods based on the novel SNPs disclosed herein, and also provides novel methods of using the known, but previously unassociated, SNPs in methods relating to AD (e.g., for diagnosing AD). In Tables 1-2, known SNPs are identified based on the public database in which they have been observed, which is indicated as one or more of the following SNP types: "dbSNP" = SNP observed in dbSNP,

“HGBASE” = SNP observed in HGBASE, and “HGMD” = SNP observed in the Human Gene Mutation Database (HGMD). Novel SNPs for which the SNP source is only “Applera” and none other, i.e., those that have not been observed in any public databases and which were also not observed during shotgun sequencing and assembly of the Celera human genome sequence (i.e., “Celera” SNP source), are indicated in Tables 3-4.

Particular SNP alleles of the present invention can be associated with either an increased risk of having or developing AD, or a decreased risk of having or developing AD. SNP alleles that are associated with a decreased risk of having or developing AD may be referred to as “protective” alleles, and SNP alleles that are associated with an increased risk of having or developing AD may be referred to as “susceptibility” alleles, “risk” alleles, or “risk factors.” Thus, whereas certain SNPs (or their encoded products) can be assayed to determine whether an individual possesses a SNP allele that is indicative of an increased risk of having or developing AD (i.e., a susceptibility allele), other SNPs (or their encoded products) can be assayed to determine whether an individual possesses a SNP allele that is indicative of a decreased risk of having or developing AD (i.e., a protective allele). Similarly, particular SNP alleles of the present invention can be associated with either an increased or decreased likelihood of responding to a particular treatment or therapeutic compound, or an increased or decreased likelihood of experiencing toxic effects from a particular treatment or therapeutic compound. The term “altered” may be used herein to encompass either of these two possibilities (e.g., an increased or a decreased risk/likelihood).

Those skilled in the art will readily recognize that nucleic acid molecules may be double-stranded molecules and that reference to a particular site on one strand refers, as well, to the corresponding site on a complementary strand. In defining a SNP position, SNP allele, or nucleotide sequence, reference to an adenine, a thymine (uridine), a cytosine, or a guanine at a particular site on one strand of a nucleic acid molecule also defines the thymine (uridine), adenine, guanine, or cytosine (respectively) at the corresponding site on a complementary strand of the nucleic acid molecule. Thus, reference may be made to either strand in order to refer to a particular SNP position, SNP allele, or nucleotide sequence. Probes and primers, may be designed to hybridize to either strand and SNP genotyping methods disclosed herein may generally target either strand. Throughout the specification, in identifying a SNP position, reference is generally made to the protein-encoding strand, only for the purpose of convenience.

References to variant peptides, polypeptides, or proteins of the present invention include peptides, polypeptides, proteins, or fragments thereof, that contain at least one amino acid residue that differs from the corresponding amino acid sequence of the art-known

peptide/polypeptide/protein (the art-known protein may be interchangeably referred to as the “wild-type,” “reference,” or “normal” protein). Such variant peptides/polypeptides/proteins can result from a codon change caused by a nonsynonymous nucleotide substitution at a protein-coding SNP position (i.e., a missense mutation) disclosed by the present invention. Variant peptides/polypeptides/proteins of the present invention can also result from a nonsense mutation, i.e. a SNP that creates a premature stop codon, a SNP that generates a read-through mutation by abolishing a stop codon, or due to any SNP disclosed by the present invention that otherwise alters the structure, function/activity, or expression of a protein, such as a SNP in a regulatory region (e.g. a promoter or enhancer) or a SNP that leads to alternative or defective splicing, such as a SNP in an intron or a SNP at an exon/intron boundary. As used herein, the terms “polypeptide,” “peptide,” and “protein” are used interchangeably.

ISOLATED NUCLEIC ACID MOLECULES AND SNP DETECTION REAGENTS & KITS

Tables 1 and 2 provide a variety of information about each SNP of the present invention that is associated with AD, including the transcript sequences (SEQ ID NOS:1-48), genomic sequences (SEQ ID NOS:148-176), and protein sequences (SEQ ID NOS:49-96) of the encoded gene products (with the SNPs indicated by IUB codes in the nucleic acid sequences). In addition, Tables 1 and 2 include SNP context sequences, which generally include 100 nucleotide upstream (5') plus 100 nucleotides downstream (3') of each SNP position (SEQ ID NOS:97-147 correspond to transcript-based SNP context sequences disclosed in Table 1, and SEQ ID NOS:177-254 correspond to genomic-based context sequences disclosed in Table 2), the alternative nucleotides (alleles) at each SNP position, and additional information about the variant where relevant, such as SNP type (coding, missense, splice site, UTR, etc.), human populations in which the SNP was observed, observed allele frequencies, information about the encoded protein, etc.

Isolated Nucleic Acid Molecules

The present invention provides isolated nucleic acid molecules that contain one or more SNPs disclosed Table 1 and/or Table 2. Preferred isolated nucleic acid molecules contain one or more SNPs identified in Table 3 and/or Table 4. Isolated nucleic acid molecules containing one or more SNPs disclosed in at least one of Tables 1-4 may be interchangeably referred to throughout the present text as “SNP-containing nucleic acid molecules.” Isolated nucleic acid molecules may optionally encode a full-length variant protein or fragment thereof. The isolated

nucleic acid molecules of the present invention also include probes and primers (which are described in greater detail below in the section entitled "SNP Detection Reagents"), which may be used for assaying the disclosed SNPs, and isolated full-length genes, transcripts, cDNA molecules, and fragments thereof, which may be used for such purposes as expressing an encoded protein.

As used herein, an "isolated nucleic acid molecule" generally is one that contains a SNP of the present invention or one that hybridizes to such molecule such as a nucleic acid with a complementary sequence, and is separated from most other nucleic acids present in the natural source of the nucleic acid molecule. Moreover, an "isolated" nucleic acid molecule, such as a cDNA molecule containing a SNP of the present invention, can be substantially free of other cellular material, or culture medium when produced by recombinant techniques, or chemical precursors or other chemicals when chemically synthesized. A nucleic acid molecule can be fused to other coding or regulatory sequences and still be considered "isolated." Nucleic acid molecules present in non-human transgenic animals, which do not naturally occur in the animal, are also considered "isolated." For example, recombinant DNA molecules contained in a vector are considered "isolated." Further examples of "isolated" DNA molecules include recombinant DNA molecules maintained in heterologous host cells, and purified (partially or substantially) DNA molecules in solution. Isolated RNA molecules include in vivo or in vitro RNA transcripts of the isolated SNP-containing DNA molecules of the present invention. Isolated nucleic acid molecules according to the present invention further include such molecules produced synthetically.

Generally, an isolated SNP-containing nucleic acid molecule comprises one or more SNP positions disclosed by the present invention with flanking nucleotide sequences on either side of the SNP positions. A flanking sequence can include nucleotide residues that are naturally associated with the SNP site and/or heterologous nucleotide sequences. Preferably the flanking sequence is up to about 500, 300, 100, 60, 50, 30, 25, 20, 15, 10, 8, or 4 nucleotides (or any other length in-between) on either side of a SNP position, or as long as the full-length gene or entire protein-coding sequence (or any portion thereof such as an exon), especially if the SNP-containing nucleic acid molecule is to be used to produce a protein or protein fragment.

For full-length genes and entire protein-coding sequences, a SNP flanking sequence can be, for example, up to about 5KB, 4KB, 3KB, 2KB, 1KB on either side of the SNP. Furthermore, in such instances, the isolated nucleic acid molecule comprises exonic sequences (including protein-coding and/or non-coding exonic sequences), but may also include intronic sequences. Thus, any protein coding sequence may be either contiguous or separated by introns. The important point is that the nucleic acid is isolated from remote and unimportant flanking sequences and is of

appropriate length such that it can be subjected to the specific manipulations or uses described herein such as recombinant protein expression, preparation of probes and primers for assaying the SNP position, and other uses specific to the SNP-containing nucleic acid sequences.

An isolated SNP-containing nucleic acid molecule can comprise, for example, a full-length
5 gene or transcript, such as a gene isolated from genomic DNA (e.g., by cloning or polymerase chain reaction [PCR] amplification), a cDNA molecule, or an mRNA transcript molecule. Polymorphic transcript sequences are provided in Table 1 and in the Sequence Listing (SEQ ID NOS: 1-48), and polymorphic genomic sequences are provided in Table 2 and in the Sequence Listing (SEQ ID
10 NOS:148-176). Furthermore, fragments of such full-length genes and transcripts that contain one or more SNPs disclosed herein are also encompassed by the present invention, and such fragments may be used, for example, to express any part of a protein, such as a particular functional domain or an antigenic epitope.

Thus, the present invention also encompasses fragments of the nucleic acid sequences provided in Tables 1-2 (transcript sequences are provided in Table 1 as SEQ ID NOS: 1-48,
15 genomic sequences are provided in Table 2 as SEQ ID NOS:148-176, transcript-based SNP context sequences are provided in Table 1 as SEQ ID NO:97-147, and genomic-based SNP context sequences are provided in Table 2 as SEQ ID NO:177-254) and their complements. A fragment typically comprises a contiguous nucleotide sequence at least about eight or more nucleotides, more preferably at least about twelve or more nucleotides, and even more preferably at least about sixteen
20 or more nucleotides. Further, a fragment could comprise at least about 18, 20, 22, 25, 30, 40, 50, 60, 80, 100, 150, 200, 250 or 500 nucleotides in length, or any other number in between. The length of the fragment will be based on its intended use. For example, the fragment can encode epitope-bearing regions of a variant peptide or regions of a variant peptide that differ from the normal/wild-type protein, or can be useful as a polynucleotide probe or primer. Such fragments can be isolated
25 using the nucleotide sequences provided in Table 1 and/or Table 2 for the synthesis of a polynucleotide probe. A labeled probe can then be used, for example, to screen a cDNA library, genomic DNA library, or mRNA to isolate nucleic acid corresponding to the coding region. Further, primers can be used in amplification reactions, such as for purposes of assaying one or more SNPs sites or for cloning specific regions of a gene.

30 An isolated nucleic acid molecule of the present invention further encompasses a SNP-containing polynucleotide that is the product of any one of a variety of nucleic acid amplification methods, which are used to increase the copy numbers of a polynucleotide of interest in a nucleic acid sample. Such amplification methods are well known in the art, and they include but are not limited to, polymerase chain reaction (PCR) (U.S. Patent Nos. 4,683,195 and 4,683,202; *PCR*

Technology: Principles and Applications for DNA Amplification, ed. H.A. Erlich, Freeman Press, New York, New York [1992]), ligase chain reaction (LCR) (Wu and Wallace, *Genomics* 4:560 [1989]; Landegren *et al.*, *Science* 241:1077 [1988]), strand displacement amplification (SDA) (U.S. Patent Nos. 5,270,184 and 5,422,252), transcription-mediated amplification (TMA) (U.S. Patent No. 5,399,491), linked linear amplification (LLA) (U.S. Patent No. 6,027,923), and the like, and isothermal amplification methods such as nucleic acid sequence based amplification (NASBA), and self-sustained sequence replication (Guatelli *et al.*, *Proc. Natl. Acad. Sci. USA* 87: 1874 [1990]). Based on such methodologies, a person skilled in the art can readily design primers in any suitable regions 5' and 3' to a SNP disclosed herein. Such primers may be used to amplify DNA of any length so long that it contains the SNP of interest in its sequence.

As used herein, an "amplified polynucleotide" of the invention is a SNP-containing nucleic acid molecule whose amount has been increased at least two fold by any nucleic acid amplification method performed *in vitro* as compared to its starting amount in a test sample. In other preferred embodiments, an amplified polynucleotide is the result of at least ten fold, fifty fold, one hundred fold, one thousand fold, or even ten thousand fold increase as compared to its starting amount in a test sample. In a typical PCR amplification, a polynucleotide of interest is often amplified at least fifty thousand fold in amount over the unamplified genomic DNA, but the precise amount of amplification needed for an assay depends on the sensitivity of the subsequent detection method used.

Generally, an amplified polynucleotide is at least about 16 nucleotides in length. More typically, an amplified polynucleotide is at least about 20 nucleotides in length. In a preferred embodiment of the invention, an amplified polynucleotide is at least about 30 nucleotides in length. In a more preferred embodiment of the invention, an amplified polynucleotide is at least about 32, 40, 45, 50, or 60 nucleotides in length. In yet another preferred embodiment of the invention, an amplified polynucleotide is at least about 100, 200, 300, 400, or 500 nucleotides in length. While the total length of an amplified polynucleotide of the invention can be as long as an exon, an intron or the entire gene where the SNP of interest resides, an amplified product is typically up to about 1,000 nucleotides in length (although certain amplification methods may generate amplified products greater than 1000 nucleotides in length). More preferably, an amplified polynucleotide is not greater than about 600-700 nucleotides in length. It is understood that irrespective of the length of an amplified polynucleotide, a SNP of interest may be located anywhere along its sequence.

In a specific embodiment of the invention, the amplified product is at least about 201 nucleotides in length, and comprises one of the transcript-based context sequences or the

genomic-based context sequences shown in Tables 1-2. Such a product may have additional sequences on its 5' end or 3' end or both. In another embodiment, the amplified product is about 101 nucleotides in length, and it contains a SNP disclosed herein. Preferably, the SNP is located at the middle of the amplified product (e.g., at position 101 in an amplified product that is 201
5 nucleotides in length, or at position 51 in an amplified product that is 101 nucleotides in length), or within 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, or 20 nucleotides from the middle of the amplified product (however, as indicated above, the SNP of interest may be located anywhere along the length of the amplified product).

The present invention provides isolated nucleic acid molecules that comprise, consist of, or
10 consist essentially of one or more polynucleotide sequences that contain one or more SNPs disclosed herein, complements thereof, and SNP-containing fragments thereof.

Accordingly, the present invention provides nucleic acid molecules that consist of any of the nucleotide sequences shown in Table 1 and/or Table 2 (transcript sequences are provided in Table 1 as SEQ ID NOS:1-48, genomic sequences are provided in Table 2 as SEQ ID NOS:148-176,
15 transcript-based SNP context sequences are provided in Table 1 as SEQ ID NO:97-147, and genomic-based SNP context sequences are provided in Table 2 as SEQ ID NO:177-254), or any nucleic acid molecule that encodes any of the variant proteins provided in Table 1 (SEQ ID NOS:49-96). A nucleic acid molecule consists of a nucleotide sequence when the nucleotide sequence is the complete nucleotide sequence of the nucleic acid molecule.

20 The present invention further provides nucleic acid molecules that consist essentially of any of the nucleotide sequences shown in Table 1 and/or Table 2 (transcript sequences are provided in Table 1 as SEQ ID NOS:1-48, genomic sequences are provided in Table 2 as SEQ ID NOS:148-176, transcript-based SNP context sequences are provided in Table 1 as SEQ ID NO:97-147, and genomic-based SNP context sequences are provided in Table 2 as SEQ ID NO:177-254), or any
25 nucleic acid molecule that encodes any of the variant proteins provided in Table 1 (SEQ ID NOS:49-96). A nucleic acid molecule consists essentially of a nucleotide sequence when such a nucleotide sequence is present with only a few additional nucleotide residues in the final nucleic acid molecule.

The present invention further provides nucleic acid molecules that comprise any of the
30 nucleotide sequences shown in Table 1 and/or Table 2 or a SNP-containing fragment thereof (transcript sequences are provided in Table 1 as SEQ ID NOS:1-48, genomic sequences are provided in Table 2 as SEQ ID NOS:148-176, transcript-based SNP context sequences are provided in Table 1 as SEQ ID NO:97-147, and genomic-based SNP context sequences are provided in Table 2 as SEQ ID NO:177-254), or any nucleic acid molecule that encodes any of the variant proteins

provided in Table 1 (SEQ ID NOS:49-96). A nucleic acid molecule comprises a nucleotide sequence when the nucleotide sequence is at least part of the final nucleotide sequence of the nucleic acid molecule. In such a fashion, the nucleic acid molecule can be only the nucleotide sequence or have additional nucleotide residues, such as residues that are naturally associated with it or
5 heterologous nucleotide sequences. Such a nucleic acid molecule can have one to a few additional nucleotides or can comprise many more additional nucleotides. A brief description of how various types of these nucleic acid molecules can be readily made and isolated is provided below, and such techniques are well known to those of ordinary skill in the art (*Molecular Cloning: A Laboratory Manual*, Sambrook and Russell, Cold Spring Harbor Press, New York [2000]).

10 The isolated nucleic acid molecules can encode mature proteins plus additional amino or carboxyl-terminal amino acids or both, or amino acids interior to the mature peptide (when the mature form has more than one peptide chain, for instance). Such sequences may play a role in processing of a protein from precursor to a mature form, facilitate protein trafficking, prolong or shorten protein half-life, or facilitate manipulation of a protein for assay or production. As generally
15 is the case *in situ*, the additional amino acids may be processed away from the mature protein by cellular enzymes.

Thus, the isolated nucleic acid molecules include, but are not limited to, nucleic acid molecules having a sequence encoding a peptide alone, a sequence encoding a mature peptide and additional coding sequences such as a leader or secretory sequence (e.g., a pre-pro or pro-protein
20 sequence), a sequence encoding a mature peptide with or without additional coding sequences, plus additional non-coding sequences, for example introns and non-coding 5' and 3' sequences such as transcribed but untranslated sequences that play a role in, for example, transcription, mRNA processing (including splicing and polyadenylation signals), ribosome binding, and/or stability of mRNA. In addition, the nucleic acid molecules may be fused to heterologous marker sequences
25 encoding, for example, a peptide that facilitates purification.

Isolated nucleic acid molecules can be in the form of RNA, such as mRNA, or in the form DNA, including cDNA and genomic DNA, which may be obtained, for example, by molecular cloning or produced by chemical synthetic techniques or by a combination thereof (*Molecular Cloning: A Laboratory Manual*, Sambrook and Russell, Cold Spring Harbor Press, New York
30 [2000]). Furthermore, isolated nucleic acid molecules, particularly SNP detection reagents such as probes and primers, can also be partially or completely in the form of one or more types of nucleic acid analogs, such as peptide nucleic acid (PNA) (U.S. Patent Nos. 5,539,082; 5,527,675; 5,623,049; 5,714,331). The nucleic acid, especially DNA, can be double-stranded or single-stranded. Single-stranded nucleic acid can be the coding strand (sense strand) or the

complementary non-coding strand (anti-sense strand). DNA, RNA, or PNA segments can be assembled, for example, from fragments of the human genome (in the case of DNA or RNA) or single nucleotides, short oligonucleotide linkers, or from a series of oligonucleotides, to provide a synthetic nucleic acid molecule. Nucleic acid molecules can be readily synthesized using the sequences provided herein as a reference; oligonucleotide and PNA oligomer synthesis techniques are well known in the art (see, e.g., Corey, "Peptide nucleic acids: expanding the scope of nucleic acid recognition," *Trends Biotechnol.* 15[6]:224-9 [June 1997], and Hyrup *et al.*, "Peptide nucleic acids [PNA]: synthesis, properties and potential applications," *Bioorg. Med. Chem.* 4[1]:5-23 [Jan. 1996]). Furthermore, large-scale automated oligonucleotide/PNA synthesis (including synthesis on an array or bead surface or other solid support) can readily be accomplished using commercially available nucleic acid synthesizers, such as the Applied Biosystems (Foster City, CA) 3900 High-Throughput DNA Synthesizer or Expedite 8909 Nucleic Acid Synthesis System, and the sequence information provided herein.

The present invention encompasses nucleic acid analogs that contain modified, synthetic, or non-naturally occurring nucleotides or structural elements or other alternative/modified nucleic acid chemistries known in the art. Such nucleic acid analogs are useful, for example, as detection reagents (e.g., primers/probes) for detecting one or more SNPs identified in Table 1 and/or Table 2. Furthermore, kits/systems (such as beads, arrays, etc.) that include these analogs are also encompassed by the present invention. For example, PNA oligomers that are based on the polymorphic sequences of the present invention are specifically contemplated. PNA oligomers are analogs of DNA in which the phosphate backbone is replaced with a peptide-like backbone (Lagriffoul *et al.*, *Bioorganic & Medicinal Chemistry Letters* 4:1081-1082 [1994], Petersen *et al.*, *Bioorganic & Medicinal Chemistry Letters* 6:793-796 [1996], Kumar *et al.*, *Organic Letters* 3[9]:1269-1272 [2001], WO96/04000). PNA hybridizes to complementary RNA or DNA with higher affinity and specificity than conventional oligonucleotides and oligonucleotide analogs. The properties of PNA enable novel molecular biology and biochemistry applications unachievable with traditional oligonucleotides and peptides.

Additional examples of nucleic acid modifications that improve the binding properties and/or stability of a nucleic acid include the use of base analogs such as inosine, intercalators (U.S. Patent No. 4,835,263) and the minor groove binders (U.S. Patent No. 5,801,115). Thus, references herein to nucleic acid molecules, SNP-containing nucleic acid molecules, SNP detection reagents (e.g., probes and primers), oligonucleotides/polynucleotides include PNA oligomers and other nucleic acid analogs. Other examples of nucleic acid analogs and

alternative/modified nucleic acid chemistries known in the art are described in *Current Protocols in Nucleic Acid Chemistry*, John Wiley & Sons, New York (2002).

The present invention further provides nucleic acid molecules that encode fragments of the variant polypeptides disclosed herein as well as nucleic acid molecules that encode obvious
5 variants of such variant polypeptides. Such nucleic acid molecules may be naturally occurring, such as paralogs (different locus) and orthologs (different organism), or may be constructed by recombinant DNA methods or by chemical synthesis. Non-naturally occurring variants may be made by mutagenesis techniques, including those applied to nucleic acid molecules, cells, or organisms. Accordingly, the variants can contain nucleotide substitutions, deletions, inversions
10 and insertions (in addition to the SNPs disclosed in Tables 1-2). Variation can occur in either or both the coding and non-coding regions. The variations can produce conservative and/or non-conservative amino acid substitutions.

Further variants of the nucleic acid molecules disclosed in Tables 1-2, such as naturally occurring allelic variants (as well as orthologs and paralogs) and synthetic variants produced by
15 mutagenesis techniques, can be identified and/or produced using methods well known in the art. Such further variants can comprise a nucleotide sequence that shares at least 70-80%, 80-85%, 85-90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% sequence identity with a nucleic acid sequence disclosed in Table 1 and/or Table 2 (or a fragment thereof) and that includes a novel SNP allele disclosed in Table 1 and/or Table 2. Further, variants can comprise a nucleotide
20 sequence that encodes a polypeptide that shares at least 70-80%, 80-85%, 85-90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% sequence identity with a polypeptide sequence disclosed in Table 1 (or a fragment thereof) and that includes a novel SNP allele disclosed in Table 1 and/or Table 2. Thus, an aspect of the present invention that is specifically contemplated is isolated nucleic acid molecules that have a certain degree of sequence variation compared with
25 the sequences shown in Tables 1-2, but that contain a novel SNP allele disclosed herein. In other words, as long as an isolated nucleic acid molecule contains a novel SNP allele disclosed herein, other portions of the nucleic acid molecule that flank the novel SNP allele can vary to some degree from the specific transcript, genomic, and context sequences shown in Tables 1-2, and can encode a polypeptide that varies to some degree from the specific polypeptide sequences shown
30 in Table 1.

To determine the percent identity of two amino acid sequences or two nucleotide sequences of two molecules that share sequence homology, the sequences are aligned for optimal comparison purposes (e.g., gaps can be introduced in one or both of a first and a second amino acid or nucleic acid sequence for optimal alignment and non-homologous sequences can be

disregarded for comparison purposes). In a preferred embodiment, at least 30%, 40%, 50%, 60%, 70%, 80%, or 90% or more of the length of a reference sequence is aligned for comparison purposes. The amino acid residues or nucleotides at corresponding amino acid positions or nucleotide positions are then compared. When a position in the first sequence is occupied by the same amino acid residue or nucleotide as the corresponding position in the second sequence, then the molecules are identical at that position (as used herein, amino acid or nucleic acid "identity" is equivalent to amino acid or nucleic acid "homology"). The percent identity between the two sequences is a function of the number of identical positions shared by the sequences, taking into account the number of gaps, and the length of each gap, which need to be introduced for optimal alignment of the two sequences.

The comparison of sequences and determination of percent identity between two sequences can be accomplished using a mathematical algorithm (*Computational Molecular Biology*, ed. A.M. Lesk, Oxford University Press, New York [1988]; *Biocomputing: Informatics and Genome Projects*, ed. Smith, D.W., Academic Press, New York [1993]; *Computer Analysis of Sequence Data, Part 1*, ed. A.M. Griffin, and H.G. Griffin, Humana Press, New Jersey [1994]; *Sequence Analysis in Molecular Biology*, G. von Heinje, Academic Press [1987]; and *Sequence Analysis Primer*, eds. M. Gribskov and J.M. Devereux, Stockton Press, New York [1991]). In a preferred embodiment, the percent identity between two amino acid sequences is determined using the Needleman and Wunsch algorithm (*J. Mol. Biol.* 48:444-453 [1970]) which has been incorporated into the GAP program in the GCG software package, using either a Blossom 62 matrix or a PAM250 matrix, and a gap weight of 16, 14, 12, 10, 8, 6, or 4 and a length weight of 1, 2, 3, 4, 5, or 6.

In yet another preferred embodiment, the percent identity between two nucleotide sequences is determined using the GAP program in the GCG software package (J. Devereux *et al.*, *Nucleic Acids Res.* 12[1]:387 [1984]), using a NWSgapdna.CMP matrix and a gap weight of 40, 50, 60, 70, or 80 and a length weight of 1, 2, 3, 4, 5, or 6. In another embodiment, the percent identity between two amino acid or nucleotide sequences is determined using the algorithm of E. Myers and W. Miller (CABIOS, 4:11-17 [1989]) which has been incorporated into the ALIGN program (version 2.0), using a PAM120 weight residue table, a gap length penalty of 12, and a gap penalty of 4.

The nucleotide and amino acid sequences of the present invention can further be used as a "query sequence" to perform a search against sequence databases to, for example, identify other family members or related sequences. Such searches can be performed using the NBLAST and XBLAST programs (version 2.0) of Altschul *et al.* (*J. Mol. Biol.* 215:403-10 [1990]). BLAST

nucleotide searches can be performed with the NBLAST program, score = 100, wordlength = 12 to obtain nucleotide sequences homologous to the nucleic acid molecules of the invention.

BLAST protein searches can be performed with the XBLAST program, score = 50, wordlength = 3 to obtain amino acid sequences homologous to the proteins of the invention. To obtain gapped
5 alignments for comparison purposes, Gapped BLAST can be utilized as described in Altschul *et al.* (*Nucleic Acids Res.* 25[17]:3389-3402 [1997]). When utilizing BLAST and gapped BLAST programs, the default parameters of the respective programs (e.g., XBLAST and NBLAST) can be used. In addition to BLAST, examples of other search and sequence comparison programs used in the art include, but are not limited to, FASTA (Pearson, *Methods Mol. Biol.* 25, 365-389
10 [1994]) and KERR (Dufresne *et al.*, *Nat. Biotechnol.* 20[12]:1269-71 [Dec. 2002]). For further information regarding bioinformatics techniques, see *Current Protocols in Bioinformatics*, John Wiley & Sons, Inc., New York.

The present invention further provides non-coding fragments of the nucleic acid molecules disclosed in Table 1 and/or Table 2. Preferred non-coding fragments include, but are
15 not limited to, promoter sequences, enhancer sequences, intronic sequences, 5' untranslated regions (UTRs), 3' untranslated regions, gene modulating sequences and gene termination sequences. Such fragments are useful, for example, in controlling heterologous gene expression and in developing screens to identify gene-modulating agents.

20 SNP Detection Reagents

In a specific aspect of the present invention, the SNPs disclosed in Table 1 and/or Table 2, and their associated transcript sequences (provided in Table 1 as SEQ ID NOS:1-48), genomic sequences (provided in Table 2 as SEQ ID NOS:148-176), and context sequences (transcript-based context sequences are provided in Table 1 as SEQ ID NOS:97-147; genomic-based context
25 sequences are provided in Table 2 as SEQ ID NOS:177-254), can be used for the design of SNP detection reagents. As used herein, a "SNP detection reagent" is a reagent that specifically detects a specific target SNP position disclosed herein, and that is preferably specific for a particular nucleotide (allele) of the target SNP position (i.e., the detection reagent preferably can differentiate between different alternative nucleotides at a target SNP position, thereby allowing the identity of
30 the nucleotide present at the target SNP position to be determined). Typically, such detection reagent hybridizes to a target SNP-containing nucleic acid molecule by complementary base-pairing in a sequence specific manner, and discriminates the target variant sequence from other nucleic acid sequences such as an art-known form in a test sample. An example of a detection reagent is a probe that hybridizes to a target nucleic acid containing one or more of the SNPs provided in Table 1

and/or Table 2. In a preferred embodiment, such a probe can differentiate between nucleic acids having a particular nucleotide (allele) at a target SNP position from other nucleic acids that have a different nucleotide at the same target SNP position. In addition, a detection reagent may hybridize to a specific region 5' and/or 3' to a SNP position, particularly a region corresponding to the context sequences provided in Table 1 and/or Table 2 (transcript-based context sequences are provided in Table 1 as SEQ ID NOS:97-147; genomic-based context sequences are provided in Table 2 as SEQ ID NOS:177-254). Another example of a detection reagent is a primer which acts as an initiation point of nucleotide extension along a complementary strand of a target polynucleotide. The SNP sequence information provided herein is also useful for designing primers, e.g. allele-specific primers, to amplify (e.g., using PCR) any SNP of the present invention.

In one preferred embodiment of the invention, a SNP detection reagent is an isolated or synthetic DNA or RNA polynucleotide probe or primer or PNA oligomer, or a combination of DNA, RNA and/or PNA, that hybridizes to a segment of a target nucleic acid molecule containing a SNP identified in Table 1 and/or Table 2. A detection reagent in the form of a polynucleotide may optionally contain modified base analogs, intercalators or minor groove binders. Multiple detection reagents such as probes may be, for example, affixed to a solid support (e.g., arrays or beads) or supplied in solution (e.g., probe/primer sets for enzymatic reactions such as PCR, RT-PCR, TaqMan assays, or primer-extension reactions) to form a SNP detection kit.

A probe or primer typically is a substantially purified oligonucleotide or PNA oligomer. Such oligonucleotide typically comprises a region of complementary nucleotide sequence that hybridizes under stringent conditions to at least about 8, 10, 12, 16, 18, 20, 22, 25, 30, 40, 50, 55, 60, 65, 70, 80, 90, 100, 120 (or any other number in-between) or more consecutive nucleotides in a target nucleic acid molecule. Depending on the particular assay, the consecutive nucleotides can either include the target SNP position, or be a specific region in close enough proximity 5' and/or 3' to the SNP position to carry out the desired assay.

Other preferred primer and probe sequences can readily be determined using the transcript sequences (SEQ ID NOS:1-48), genomic sequences (SEQ ID NOS:148-176), and SNP context sequences (transcript-based context sequences are provided in Table 1 as SEQ ID NOS:97-147; genomic-based context sequences are provided in Table 2 as SEQ ID NOS:177-254) disclosed in the Sequence Listing and in Tables 1-2. It will be apparent to one of skill in the art that such primers and probes are directly useful as reagents for genotyping the SNPs of the present invention, and can be incorporated into any kit/system format.

In order to produce a probe or primer specific for a target SNP-containing sequence, the gene/transcript and/or context sequence surrounding the SNP of interest is typically examined using a computer algorithm which starts at the 5' or at the 3' end of the nucleotide sequence. Typical algorithms will then identify oligomers of defined length that are unique to the gene/SNP context sequence, have a GC content within a range suitable for hybridization, lack predicted secondary structure that may interfere with hybridization, and/or possess other desired characteristics or that lack other undesired characteristics.

A primer or probe of the present invention is typically at least about 8 nucleotides in length. In one embodiment of the invention, a primer or a probe is at least about 10 nucleotides in length. In a preferred embodiment, a primer or a probe is at least about 12 nucleotides in length. In a more preferred embodiment, a primer or probe is at least about 16, 17, 18, 19, 20, 21, 22, 23, 24 or 25 nucleotides in length. While the maximal length of a probe can be as long as the target sequence to be detected, depending on the type of assay in which it is employed, it is typically less than about 50, 60, 65, or 70 nucleotides in length. In the case of a primer, it is typically less than about 30 nucleotides in length. In a specific preferred embodiment of the invention, a primer or a probe is within the length of about 18 and about 28 nucleotides. However, in other embodiments, such as nucleic acid arrays and other embodiments in which probes are affixed to a substrate, the probes can be longer, such as on the order of 30-70, 75, 80, 90, 100, or more nucleotides in length (see the section below entitled "SNP Detection Kits and Systems").

For analyzing SNPs, it may be appropriate to use oligonucleotides specific for alternative SNP alleles. Such oligonucleotides which detect single nucleotide variations in target sequences may be referred to by such terms as "allele-specific oligonucleotides," "allele-specific probes," or "allele-specific primers." The design and use of allele-specific probes for analyzing polymorphisms is described in, e.g., *Mutation Detection: A Practical Approach*, ed. Cotton *et al.*, Oxford University Press [1998]; Saiki *et al.*, *Nature* 324, 163-166 [1986]; Dattagupta, EP235,726; and Saiki, WO 89/11548.

While the design of each allele-specific primer or probe depends on variables such as the precise composition of the nucleotide sequences flanking a SNP position in a target nucleic acid molecule, and the length of the primer or probe, another factor in the use of primers and probes is the stringency of the condition under which the hybridization between the probe or primer and the target sequence is performed. Higher stringency conditions utilize buffers with lower ionic strength and/or a higher reaction temperature, and tend to require a more perfect match between probe/primer and a target sequence in order to form a stable duplex. If the stringency is too high,

however, hybridization may not occur at all. In contrast, lower stringency conditions utilize buffers with higher ionic strength and/or a lower reaction temperature, and permit the formation of stable duplexes with more mismatched bases between a probe/primer and a target sequence. By way of example and not limitation, exemplary conditions for high stringency hybridization
5 conditions using an allele-specific probe are as follows: Prehybridization with a solution containing 5X standard saline phosphate EDTA (SSPE), 0.5% NaDodSO₄ (SDS) at 55°C, and incubating probe with target nucleic acid molecules in the same solution at the same temperature, followed by washing with a solution containing 2X SSPE, and 0.1% SDS at 55°C or room temperature.

10 Moderate stringency hybridization conditions may be used for allele-specific primer extension reactions with a solution containing, e.g., about 50mM KCl at about 46°C. Alternatively, the reaction may be carried out at an elevated temperature such as 60°C. In another embodiment, a moderately stringent hybridization condition suitable for oligonucleotide ligation assay (OLA) reactions wherein two probes are ligated if they are completely
15 complementary to the target sequence may utilize a solution of about 100mM KCl at a temperature of 46°C.

In a hybridization-based assay, allele-specific probes can be designed that hybridize to a segment of target DNA from one individual but do not hybridize to the corresponding segment from another individual due to the presence of different polymorphic forms (e.g., alternative SNP
20 alleles/nucleotides) in the respective DNA segments from the two individuals. Hybridization conditions should be sufficiently stringent that there is a significant detectable difference in hybridization intensity between alleles, and preferably an essentially binary response, whereby a probe hybridizes to only one of the alleles or significantly more strongly to one allele. While a probe may be designed to hybridize to a target sequence that contains a SNP site such that the
25 SNP site aligns anywhere along the sequence of the probe, the probe is preferably designed to hybridize to a segment of the target sequence such that the SNP site aligns with a central position of the probe (e.g., a position within the probe that is at least three nucleotides from either end of the probe). This design of probe generally achieves good discrimination in hybridization between different allelic forms.

30 In another embodiment, a probe or primer may be designed to hybridize to a segment of target DNA such that the SNP aligns with either the 5'-most end or the 3'-most end of the probe or primer. In a specific preferred embodiment which is particularly suitable for use in a oligonucleotide ligation assay (U.S. Patent No. 4,988,617), the 3'-most nucleotide of the probe aligns with the SNP position in the target sequence.

Oligonucleotide probes and primers may be prepared by methods well known in the art. Chemical synthetic methods include, but are not limited to, the phosphotriester method described by Narang *et al.*, *Methods in Enzymology* 68:90 [1979]; the phosphodiester method described by Brown *et al.*, *Methods in Enzymology* 68:109 [1979], the diethylphosphoamidate method
5 described by Beaucage *et al.*, *Tetrahedron Letters* 22:1859 [1981]; and the solid support method described in U.S. Patent No. 4,458,066.

Allele-specific probes are often used in pairs (or, less commonly, in sets of 3 or 4, such as if a SNP position is known to have 3 or 4 alleles, respectively, or to assay both strands of a nucleic acid molecule for a target SNP allele), and such pairs may be identical except for a one
10 nucleotide mismatch that represents the allelic variants at the SNP position. Commonly, one member of a pair perfectly matches a reference form of a target sequence that has a more common SNP allele (i.e., the allele that is more frequent in the target population) and the other member of the pair perfectly matches a form of the target sequence that has a less common SNP allele (i.e., the allele that is rarer in the target population). In the case of an array, multiple pairs
15 of probes can be immobilized on the same support for simultaneous analysis of multiple different polymorphisms.

In one type of PCR-based assay, an allele-specific primer hybridizes to a region on a target nucleic acid molecule that overlaps a SNP position and only primes amplification of an allelic form to which the primer exhibits perfect complementarity (Gibbs, *Nucleic Acid Res.*
20 17:2427-2448 [1989]). Typically, the primer's 3'-most nucleotide is aligned with and complementary to the SNP position of the target nucleic acid molecule. This primer is used in conjunction with a second primer that hybridizes at a distal site. Amplification proceeds from the two primers, producing a detectable product that indicates which allelic form is present in the test sample. A control is usually performed with a second pair of primers, one of which shows a
25 single base mismatch at the polymorphic site and the other of which exhibits perfect complementarity to a distal site. The single-base mismatch prevents amplification or substantially reduces amplification efficiency, so that either no detectable product is formed or it is formed in lower amounts or at a slower pace. The method generally works most effectively when the mismatch is at the 3'-most position of the oligonucleotide (i.e., the 3'-most position of
30 the oligonucleotide aligns with the target SNP position) because this position is most destabilizing to elongation from the primer (see, e.g., WO 93/22456). This PCR-based assay can be utilized as part of the TaqMan assay, described below.

In a specific embodiment of the invention, a primer of the invention contains a sequence substantially complementary to a segment of a target SNP-containing nucleic acid molecule except

that the primer has a mismatched nucleotide in one of the three nucleotide positions at the 3'-most end of the primer, such that the mismatched nucleotide does not base pair with a particular allele at the SNP site. In a preferred embodiment, the mismatched nucleotide in the primer is the second from the last nucleotide at the 3'-most position of the primer. In a more preferred embodiment, the mismatched nucleotide in the primer is the last nucleotide at the 3'-most position of the primer.

In another embodiment of the invention, a SNP detection reagent of the invention is labeled with a fluorogenic reporter dye that emits a detectable signal. While the preferred reporter dye is a fluorescent dye, any reporter dye that can be attached to a detection reagent such as an oligonucleotide probe or primer is suitable for use in the invention. Such dyes include, but are not limited to, Acridine, AMCA, BODIPY, Cascade Blue, Cy2, Cy3, Cy5, Cy7, Dabcyl, Edans, Eosin, Erythrosin, Fluorescein, 6-Fam, Tet, Joe, Hex, Oregon Green, Rhodamine, Rhodol Green, Tamra, Rox, and Texas Red.

In yet another embodiment of the invention, the detection reagent may be further labeled with a quencher dye such as Tamra, especially when the reagent is used as a self-quenching probe such as a TaqMan (U.S. Patent Nos. 5,210,015 and 5,538,848) or Molecular Beacon probe (U.S. Patent Nos. 5,118,801 and 5,312,728), or other stemless or linear beacon probe (Livak *et al.*, *PCR Method Appl.* 4:357-362 [1995]; Tyagi *et al.*, *Nature Biotechnology* 14:303-308 [1996]; Nazarenko *et al.*, *Nucl. Acids Res.* 25:2516-2521 [1997]; U.S. Patent Nos. 5,866,336 and 6,117,635).

The detection reagents of the invention may also contain other labels, including but not limited to, biotin for streptavidin binding, hapten for antibody binding, and oligonucleotide for binding to another complementary oligonucleotide such as pairs of zipcodes.

The present invention also contemplates reagents that do not contain (or that are complementary to) a SNP nucleotide identified herein but that are used to assay one or more SNPs disclosed herein. For example, primers that flank, but do not hybridize directly to a target SNP position provided herein are useful in primer extension reactions in which the primers hybridize to a region adjacent to the target SNP position (i.e., within one or more nucleotides from the target SNP site). During the primer extension reaction, a primer is typically not able to extend past a target SNP site if a particular nucleotide (allele) is present at that target SNP site, and the primer extension product can be detected in order to determine which SNP allele is present at the target SNP site. For example, particular ddNTPs are typically used in the primer extension reaction to terminate primer extension once a ddNTP is incorporated into the extension product (a primer extension product which includes a ddNTP at the 3'-most end of the primer extension product, and in which the ddNTP is a nucleotide of a SNP disclosed herein, is a composition that is specifically contemplated by the present invention). Thus, reagents that bind

to a nucleic acid molecule in a region adjacent to a SNP site and that are used for assaying the SNP site, even though the bound sequences do not necessarily include the SNP site itself, are also contemplated by the present invention.

5 **SNP Detection Kits and Systems**

A person skilled in the art will recognize that, based on the SNP and associated sequence information disclosed herein, detection reagents can be developed and used to assay any SNP of the present invention individually or in combination, and such detection reagents can be readily incorporated into one of the established kit or system formats which are well known in the art.

10 The terms “kits” and “systems”, as used herein in the context of SNP detection reagents, are intended to refer to such things as combinations of multiple SNP detection reagents, or one or more SNP detection reagents in combination with one or more other types of elements or components (e.g., other types of biochemical reagents, containers, packages such as packaging intended for commercial sale, substrates to which SNP detection reagents are attached, electronic
15 hardware components, etc.). Accordingly, the present invention further provides SNP detection kits and systems, including but not limited to, packaged probe and primer sets (e.g., TaqMan probe/primer sets), arrays/microarrays of nucleic acid molecules, and beads that contain one or more probes, primers, or other detection reagents for detecting one or more SNPs of the present invention. The kits/systems can optionally include various electronic hardware components; for
20 example, arrays (“DNA chips”) and microfluidic systems (“lab-on-a-chip” systems) provided by various manufacturers typically comprise hardware components. Other kits/systems (e.g., probe/primer sets) may not include electronic hardware components, but may be comprised of, for example, one or more SNP detection reagents (along with, optionally, other biochemical reagents) packaged in one or more containers.

25 In some embodiments, a SNP detection kit typically contains one or more detection reagents and other components (e.g., a buffer, enzymes such as DNA polymerases or ligases, chain extension nucleotides such as deoxynucleotide triphosphates, and in the case of Sanger-type DNA sequencing reactions, chain terminating nucleotides, positive control sequences, negative control sequences, and the like) necessary to carry out an assay or reaction, such as
30 amplification and/or detection of a SNP-containing nucleic acid molecule. A kit may further contain means for determining the amount of a target nucleic acid, and means for comparing the amount with a standard, and can comprise instructions for using the kit to detect the SNP-containing nucleic acid molecule of interest. In one embodiment of the present invention, kits are provided which contain the necessary reagents to carry out one or more assays to detect one

or more SNPs disclosed herein. In a preferred embodiment of the present invention, SNP detection kits/systems are in the form of nucleic acid arrays, or compartmentalized kits, including microfluidic/lab-on-a-chip systems.

SNP detection kits/systems may contain, for example, one or more probes, or pairs of probes, that hybridize to a nucleic acid molecule at or near each target SNP position. Multiple pairs of allele-specific probes may be included in the kit/system to simultaneously assay large numbers of SNPs, at least one of which is a SNP of the present invention. In some kits/systems, the allele-specific probes are immobilized to a substrate such as an array or bead. For example, the same substrate can comprise allele-specific probes for detecting at least 1; 10; 100; 1000; 10,000; 100,000 (or any other number in-between) or substantially all of the SNPs shown in Table 1 and/or Table 2.

The terms "arrays," "microarrays," and "DNA chips" are used herein interchangeably to refer to an array of distinct polynucleotides affixed to a substrate, such as glass, plastic, paper, nylon or other type of membrane, filter, chip, or any other suitable solid support. The polynucleotides can be synthesized directly on the substrate, or synthesized separate from the substrate and then affixed to the substrate. In one embodiment, the microarray is prepared and used according to the methods described in U.S. Patent No. 5,837,832 (Chee *et al.*), PCT application W095/11995 (Chee *et al.*), Lockhart, D. J. *et al.* (*Nat. Biotech.* 14:1675-1680 [1996]) and Schena, M. *et al.* (*Proc. Natl. Acad. Sci.* 93:10614-10619 [1996]), all of which are incorporated herein in their entirety by reference. In other embodiments, such arrays are produced by the methods described by Brown *et al.*, U.S. Patent No. 5,807,522.

Nucleic acid arrays are reviewed in the following references: Zammattéo *et al.*, "New chips for molecular biology and diagnostics," *Biotechnol. Annu. Rev.* 8:85-101 (2002); Sosnowski *et al.*, "Active microelectronic array system for DNA hybridization, genotyping and pharmacogenomic applications," *Psychiatr. Genet.* 12(4):181-92 (Dec. 2002); Heller, "DNA microarray technology: devices, systems, and applications," *Annu. Rev. Biomed. Eng.* 4:129-53 (2002); Epub Mar. 22 2002; Kolchinsky *et al.*, "Analysis of SNPs and other genomic variations using gel-based chips," *Hum. Mutat.* 19(4):343-60 (Apr. 2002); and McGall *et al.*, "High-density genechip oligonucleotide probe arrays," *Adv. Biochem. Eng. Biotechnol.* 77:21-42.(2002).

Any number of probes, such as allele-specific probes, may be implemented in an array, and each probe or pair of probes can hybridize to a different SNP position. In the case of polynucleotide probes, they can be synthesized at designated areas (or synthesized separately and then affixed to designated areas) on a substrate using a light-directed chemical process. Each DNA chip can contain, for example, thousands to millions of individual synthetic polynucleotide probes

arranged in a grid-like pattern and miniaturized (e.g., to the size of a dime). Preferably, probes are attached to a solid support in an ordered, addressable array.

A microarray can be composed of a large number of unique, single-stranded polynucleotides, usually either synthetic antisense polynucleotides or fragments of cDNAs, fixed to a solid support. Typical polynucleotides are preferably about 6-60 nucleotides in length, more preferably about 15-30 nucleotides in length, and most preferably about 18-25 nucleotides in length. For certain types of microarrays or other detection kits/systems, it may be preferable to use oligonucleotides that are only about 7-20 nucleotides in length. In other types of arrays, such as arrays used in conjunction with chemiluminescent detection technology, preferred probe lengths can be, for example, about 15-80 nucleotides in length, preferably about 50-70 nucleotides in length, more preferably about 55-65 nucleotides in length, and most preferably about 60 nucleotides in length. The microarray or detection kit can contain polynucleotides that cover the known 5' or 3' sequence of a gene/transcript or target SNP site, sequential polynucleotides that cover the full-length sequence of a gene/transcript; or unique polynucleotides selected from particular areas along the length of a target gene/transcript sequence, particularly areas corresponding to one or more SNPs disclosed in Table 1 and/or Table 2. Polynucleotides used in the microarray or detection kit can be specific to a SNP or SNPs of interest (e.g., specific to a particular SNP allele at a target SNP site, or specific to particular SNP alleles at multiple different SNP sites), or specific to a polymorphic gene/transcript or genes/transcripts of interest.

Hybridization assays based on polynucleotide arrays rely on the differences in hybridization stability of the probes to perfectly matched and mismatched target sequence variants. For SNP genotyping, it is generally preferable that stringency conditions used in hybridization assays are high enough such that nucleic acid molecules that differ from one another at as little as a single SNP position can be differentiated (e.g., typical SNP hybridization assays are designed so that hybridization will occur only if one particular nucleotide is present at a SNP position, but will not occur if an alternative nucleotide is present at that SNP position). Such high stringency conditions may be preferable when using, for example, nucleic acid arrays of allele-specific probes for SNP detection. Such high stringency conditions are described in the preceding section, and are well known to those skilled in the art and can be found in, for example, *Current Protocols in Molecular Biology* 6.3.1-6.3.6, John Wiley & Sons, New York (1989).

In other embodiments, the arrays are used in conjunction with chemiluminescent detection technology. The following patents and patent applications, which are all hereby incorporated by reference, provide additional information pertaining to chemiluminescent

detection: U.S. patent applications 10/620332 and 10/620333 describe chemiluminescent approaches for microarray detection; U.S. Patent Nos. 6124478, 6107024, 5994073, 5981768, 5871938, 5843681, 5800999, and 5773628 describe methods and compositions of dioxetane for performing chemiluminescent detection; and U.S. published application US2002/0110828
5 discloses methods and compositions for microarray controls.

In one embodiment of the invention, a nucleic acid array can comprise an array of probes of about 15-25 nucleotides in length. In further embodiments, a nucleic acid array can comprise any number of probes, in which at least one probe is capable of detecting one or more SNPs disclosed in Table 1 and/or Table 2, and/or at least one probe comprises a fragment of one of the
10 sequences selected from the group consisting of those disclosed in Table 1, Table 2, the Sequence Listing, and sequences complementary thereto, said fragment comprising at least about 8 consecutive nucleotides, preferably 10, 12, 15, 16, 18, 20, more preferably 22, 25, 30, 40, 47, 50, 55, 60, 65, 70, 80, 90, 100, or more consecutive nucleotides (or any other number in-between) and containing (or being complementary to) a novel SNP allele disclosed in Table 1
15 and/or Table 2. In some embodiments, the nucleotide complementary to the SNP site is within 5, 4, 3, 2, or 1 nucleotide from the center of the probe, more preferably at the center of said probe.

A polynucleotide probe can be synthesized on the surface of the substrate by using a chemical coupling procedure and an ink jet application apparatus, as described in PCT application W095/251116 (Baldeschweiler *et al.*) which is incorporated herein in its entirety by reference. In
20 another aspect, a "gridded" array analogous to a dot (or slot) blot may be used to arrange and link cDNA fragments or oligonucleotides to the surface of a substrate using a vacuum system, thermal, UV, mechanical or chemical bonding procedures. An array, such as those described above, may be produced by hand or by using available devices (slot blot or dot blot apparatus), materials (any suitable solid support), and machines (including robotic instruments), and may contain 8, 24, 96,
25 384, 1536, 6144 or more polynucleotides, or any other number which lends itself to the efficient use of commercially available instrumentation.

Using such arrays or other kits/systems, the present invention provides methods of identifying the SNPs disclosed herein in a test sample. Such methods typically involve incubating a test sample of nucleic acids with an array comprising one or more probes corresponding to at least
30 one SNP position of the present invention, and assaying for binding of a nucleic acid from the test sample with one or more of the probes. Conditions for incubating a SNP detection reagent (or a kit/system that employs one or more such SNP detection reagents) with a test sample vary. Incubation conditions depend on such factors as the format employed in the assay, the detection methods employed, and the type and nature of the detection reagents used in the assay. One skilled

in the art will recognize that any one of the commonly available hybridization, amplification and array assay formats can readily be adapted to detect the SNPs disclosed herein.

A SNP detection kit/system of the present invention may include components that are used to prepare nucleic acids from a test sample for the subsequent amplification and/or detection of a SNP-containing nucleic acid molecule. Such sample preparation components can be used to produce nucleic acid extracts (including DNA and/or RNA), proteins or membrane extracts from any bodily fluids (such as blood, serum, plasma, urine, saliva, phlegm, gastric juices, semen, tears, sweat, etc.), skin, hair, cells (especially nucleated cells), biopsies, buccal swabs or tissue specimens. The test samples used in the above-described methods will vary based on such factors as the assay format, nature of the detection method, and the specific tissues, cells or extracts used as the test sample to be assayed. Methods of preparing nucleic acids, proteins, and cell extracts are well known in the art and can be readily adapted to obtain a sample that is compatible with the system utilized. Automated sample preparation systems for extracting nucleic acids from a test sample are commercially available, and examples are Qiagen's BioRobot 9600, Applied Biosystems' PRISM 6700, and Roche Molecular Systems' COBAS AmpliPrep System.

Another form of kit contemplated by the present invention is a compartmentalized kit. A compartmentalized kit includes any kit in which reagents are contained in separate containers. Such containers include, for example, small glass containers, plastic containers, strips of plastic, glass or paper, or arraying material such as silica. Such containers allow one to efficiently transfer reagents from one compartment to another compartment such that the test samples and reagents are not cross-contaminated, or from one container to another vessel not included in the kit, and the agents or solutions of each container can be added in a quantitative fashion from one compartment to another or to another vessel. Such containers may include, for example, one or more containers which will accept the test sample, one or more containers which contain at least one probe or other SNP detection reagent for detecting one or more SNPs of the present invention, one or more containers which contain wash reagents (such as phosphate buffered saline, Tris-buffers, etc.), and one or more containers which contain the reagents used to reveal the presence of the bound probe or other SNP detection reagents. The kit can optionally further comprise compartments and/or reagents for, for example, nucleic acid amplification or other enzymatic reactions such as primer extension reactions, hybridization, ligation, electrophoresis (preferably capillary electrophoresis), mass spectrometry, and/or laser-induced fluorescent detection. The kit may also include instructions for using the kit. Exemplary compartmentalized kits include microfluidic devices known in the art (see, e.g., Weigl *et al.*, "Lab-on-a-chip for drug development,"

Adv. Drug Deliv. Rev. 24, 55[3]:349-77 [Feb. 2003]). In such microfluidic devices, the containers may be referred to as, for example, microfluidic “compartments,” “chambers,” or “channels.”

Microfluidic devices, which may also be referred to as “lab-on-a-chip” systems, biomedical micro-electro-mechanical systems (bioMEMs), or multicomponent integrated systems, are exemplary kits/systems of the present invention for analyzing SNPs. Such systems miniaturize and compartmentalize processes such as probe/target hybridization, nucleic acid amplification, and capillary electrophoresis reactions in a single functional device. Such microfluidic devices typically utilize detection reagents in at least one aspect of the system, and such detection reagents may be used to detect one or more SNPs of the present invention. One example of a microfluidic system is disclosed in U.S. Patent No. 5,589,136, which describes the integration of PCR amplification and capillary electrophoresis in chips. Exemplary microfluidic systems comprise a pattern of microchannels designed onto a glass, silicon, quartz, or plastic wafer included on a microchip. The movements of the samples may be controlled by electric, electroosmotic or hydrostatic forces applied across different areas of the microchip to create functional microscopic valves and pumps with no moving parts. Varying the voltage can be used as a means to control the liquid flow at intersections between the micro-machined channels and to change the liquid flow rate for pumping across different sections of the microchip. See, for example, U.S. Patent Nos. 6,153,073, Dubrow *et al.*, and 6,156,181, Parce *et al.*

For genotyping SNPs, an exemplary microfluidic system may integrate, for example, nucleic acid amplification, primer extension, capillary electrophoresis, and a detection method such as laser induced fluorescence detection. In a first step of an exemplary process for using such an exemplary system, nucleic acid samples are amplified, preferably by PCR. Then, the amplification products are subjected to automated primer extension reactions using ddNTPs (specific fluorescence for each ddNTP) and the appropriate oligonucleotide primers to carry out primer extension reactions which hybridize just upstream of the targeted SNP. Once the extension at the 3' end is completed, the primers are separated from the unincorporated fluorescent ddNTPs by capillary electrophoresis. The separation medium used in capillary electrophoresis can be, for example, polyacrylamide, polyethyleneglycol or dextran. The incorporated ddNTPs in the single nucleotide primer extension products are identified by laser-induced fluorescence detection. Such an exemplary microchip can be used to process, for example, at least 96 to 384 samples, or more, in parallel.

USES OF NUCLEIC ACID MOLECULES

The nucleic acid molecules of the present invention have a variety of uses, especially in the diagnosis and treatment of AD. For example, the nucleic acid molecules are useful as hybridization

probes, such as for genotyping SNPs in messenger RNA, transcript, cDNA, genomic DNA, amplified DNA or other nucleic acid molecules, and for isolating full-length cDNA and genomic clones encoding the variant peptides disclosed in Table 1 as well as their orthologs.

5 A probe can hybridize to any nucleotide sequence along the entire length of a nucleic acid molecule provided in Table 1 and/or Table 2. Preferably, a probe of the present invention hybridizes to a region of a target sequence that encompasses a SNP position indicated in Table 1 and/or Table 2. More preferably, a probe hybridizes to a SNP-containing target sequence in a sequence-specific manner such that it distinguishes the target sequence from other nucleotide sequences which vary from the target sequence only by which nucleotide is present at the SNP site. Such a probe is particularly useful for detecting the presence of a SNP-containing nucleic acid in a test sample, or for determining which nucleotide (allele) is present at a particular SNP site (i.e., genotyping the SNP site).

A nucleic acid hybridization probe may be used for determining the presence, level, form, and/or distribution of nucleic acid expression. The nucleic acid whose level is determined can be DNA or RNA. Accordingly, probes specific for the SNPs described herein can be used to assess the presence, expression and/or gene copy number in a given cell, tissue, or organism. These uses are relevant for diagnosis of disorders involving an increase or decrease in gene expression relative to normal levels. *In vitro* techniques for detection of mRNA include, for example, Northern blot hybridizations and *in situ* hybridizations. *In vitro* techniques for detecting DNA include Southern blot hybridizations and *in situ* hybridizations (Sambrook and Russell, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Press, Cold Spring Harbor, New York [2000]).

25 Probes can be used as part of a diagnostic test kit for identifying cells or tissues in which a variant protein is expressed, such as by measuring the level of a variant protein-encoding nucleic acid (e.g., mRNA) in a sample of cells from a subject or determining if a polynucleotide contains a SNP of interest.

Thus, the nucleic acid molecules of the invention can be used as hybridization probes to detect the SNPs disclosed herein, thereby determining whether an individual with the polymorphisms is at risk for AD or has developed early stage AD. Detection of a SNP associated with a disease phenotype provides a diagnostic tool for an active disease and/or genetic predisposition to the disease.

Furthermore, the nucleic acid molecules of the invention are therefore useful for detecting a gene (gene information is disclosed in Table 2, for example) which contains a SNP disclosed herein and/or products of such genes, such as expressed mRNA transcript molecules (transcript

information is disclosed in Table 1, for example), and are thus useful for detecting gene expression. The nucleic acid molecules can optionally be implemented in, for example, an array or kit format for use in detecting gene expression.

5 The nucleic acid molecules of the invention are also useful as primers to amplify any given region of a nucleic acid molecule, particularly a region containing a SNP identified in Table 1 and/or Table 2.

10 The nucleic acid molecules of the invention are also useful for constructing recombinant vectors (described in greater detail below). Such vectors include expression vectors that express a portion of, or all of, any of the variant peptide sequences provided in Table 1. Vectors also include insertion vectors, used to integrate into another nucleic acid molecule sequence, such as into the cellular genome, to alter *in situ* expression of a gene and/or gene product. For example, an endogenous coding sequence can be replaced via homologous recombination with all or part of the coding region containing one or more specifically introduced SNPs.

15 The nucleic acid molecules of the invention are also useful for expressing antigenic portions of the variant proteins, particularly antigenic portions that contain a variant amino acid sequence (e.g., an amino acid substitution) caused by a SNP disclosed in Table 1 and/or Table 2.

The nucleic acid molecules of the invention are also useful for constructing vectors containing a gene regulatory region of the nucleic acid molecules of the present invention.

20 The nucleic acid molecules of the invention are also useful for designing ribozymes corresponding to all, or a part, of an mRNA molecule expressed from a SNP-containing nucleic acid molecule described herein.

The nucleic acid molecules of the invention are also useful for constructing host cells expressing a part, or all, of the nucleic acid molecules and variant peptides.

25 The nucleic acid molecules of the invention are also useful for constructing transgenic animals expressing all, or a part, of the nucleic acid molecules and variant peptides. The production of recombinant cells and transgenic animals having nucleic acid molecules which contain the SNPs disclosed in Table 1 and/or Table 2 allow, for example, effective clinical design of treatment compounds and dosage regimens.

30 The nucleic acid molecules of the invention are also useful in assays for drug screening to identify compounds that, for example, modulate nucleic acid expression.

The nucleic acid molecules of the invention are also useful in gene therapy in patients whose cells have aberrant gene expression. Thus, recombinant cells, which include a patient's cells that have been engineered *ex vivo* and returned to the patient, can be introduced into an individual where the recombinant cells produce the desired protein to treat the individual.

SNP Genotyping Methods

The process of determining which specific nucleotide (i.e., allele) is present at each of one or more SNP positions, such as a SNP position in a nucleic acid molecule disclosed in Table 1 and/or Table 2, is referred to as SNP genotyping. The present invention provides methods of SNP genotyping, such as for use in screening for AD or related pathologies, or determining predisposition thereto, or determining responsiveness to a form of treatment, or in genome mapping or SNP association analysis, etc.

Nucleic acid samples can be genotyped to determine which allele(s) is/are present at any given genetic region (e.g., SNP position) of interest by methods well known in the art. The neighboring sequence can be used to design SNP detection reagents such as oligonucleotide probes, which may optionally be implemented in a kit format. Exemplary SNP genotyping methods are described in Chen *et al.*, "Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput," *Pharmacogenomics J.* 3(2):77-96 (2003); Kwok *et al.*, "Detection of single nucleotide polymorphisms," *Curr. Issues Mol. Biol.* 5(2):43-60 (Apr. 2003); Shi, "Technologies for individual genotyping: detection of genetic polymorphisms in drug targets and disease genes," *Am. J. Pharmacogenomics* 2(3):197-205 (2002); and Kwok, "Methods for genotyping single nucleotide polymorphisms," *Annu. Rev. Genomics Hum. Genet.* 2:235-58 (2001). Exemplary techniques for high-throughput SNP genotyping are described in Marnellos, "High-throughput SNP analysis for genetic association studies," *Curr. Opin. Drug Discov. Devel.* 6(3):317-21 (May 2003). Common SNP genotyping methods include, but are not limited to, TaqMan assays, molecular beacon assays, nucleic acid arrays, allele-specific primer extension, allele-specific PCR, arrayed primer extension, homogeneous primer extension assays, primer extension with detection by mass spectrometry, pyrosequencing, multiplex primer extension sorted on genetic arrays, ligation with rolling circle amplification, homogeneous ligation, OLA (U.S. Patent No. 4,988,167), multiplex ligation reaction sorted on genetic arrays, restriction-fragment length polymorphism, single base extension-tag assays, and the Invader assay. Such methods may be used in combination with detection mechanisms such as, for example, luminescence or chemiluminescence detection, fluorescence detection, time-resolved fluorescence detection, fluorescence resonance energy transfer, fluorescence polarization, mass spectrometry, and electrical detection.

Various methods for detecting polymorphisms include, but are not limited to, methods in which protection from cleavage agents is used to detect mismatched bases in RNA/RNA or RNA/DNA duplexes (Myers *et al.*, *Science* 230:1242 [1985]; Cotton *et al.*, *PNAS* 85:4397 [1988]; and Saleeba *et al.*, *Meth. Enzymol.* 217:286-295 [1992]), comparison of the electrophoretic mobility

of variant and wild type nucleic acid molecules (Orita *et al.*, *PNAS* 86:2766 [1989]; Cotton *et al.*, *Mutat. Res.* 285:125-144 [1993]; and Hayashi *et al.*, *Genet. Anal. Tech. Appl.* 9:73-79 [1992]), and assaying the movement of polymorphic or wild-type fragments in polyacrylamide gels containing a gradient of denaturant using denaturing gradient gel electrophoresis (DGGE) (Myers *et al.*, *Nature* 313:495 [1985]). Sequence variations at specific locations can also be assessed by nuclease protection assays such as RNase and S1 protection or chemical cleavage methods.

In a preferred embodiment, SNP genotyping is performed using the TaqMan assay, which is also known as the 5' nuclease assay (U.S. Patent Nos. 5,210,015 and 5,538,848). The TaqMan assay detects the accumulation of a specific amplified product during PCR. The TaqMan assay utilizes an oligonucleotide probe labeled with a fluorescent reporter dye and a quencher dye. The reporter dye is excited by irradiation at an appropriate wavelength, it transfers energy to the quencher dye in the same probe via a process called fluorescence resonance energy transfer (FRET). When attached to the probe, the excited reporter dye does not emit a signal. The proximity of the quencher dye to the reporter dye in the intact probe maintains a reduced fluorescence for the reporter. The reporter dye and quencher dye may be at the 5'-most and the 3'-most ends, respectively, or vice versa. Alternatively, the reporter dye may be at the 5'- or 3'-most end while the quencher dye is attached to an internal nucleotide, or vice versa. In yet another embodiment, both the reporter and the quencher may be attached to internal nucleotides at a distance from each other such that fluorescence of the reporter is reduced.

During PCR, the 5' nuclease activity of DNA polymerase cleaves the probe, thereby separating the reporter dye and the quencher dye and resulting in increased fluorescence of the reporter. Accumulation of PCR product is detected directly by monitoring the increase in fluorescence of the reporter dye. The DNA polymerase cleaves the probe between the reporter dye and the quencher dye only if the probe hybridizes to the target SNP-containing template which is amplified during PCR, and the probe is designed to hybridize to the target SNP site only if a particular SNP allele is present.

Preferred TaqMan primer and probe sequences can readily be determined using the SNP and associated nucleic acid sequence information provided herein. A number of computer programs, such as Primer Express (Applied Biosystems, Foster City, CA), can be used to rapidly obtain optimal primer/probe sets. It will be apparent to one of skill in the art that such primers and probes for detecting the SNPs of the present invention are useful in diagnostic assays for AD and related pathologies, and can be readily incorporated into a kit format. The present invention also includes modifications of the Taqman assay well known in the art such as the use of

Molecular Beacon probes (U.S. Patent Nos. 5,118,801 and 5,312,728) and other variant formats (U.S. Patent Nos. 5,866,336 and 6,117,635).

Another preferred method for genotyping the SNPs of the present invention is the use of two oligonucleotide probes in an OLA (see, e.g., U.S. Patent No. 4,988,617). In this method, one probe hybridizes to a segment of a target nucleic acid with its 3'-most end aligned with the SNP site. A second probe hybridizes to an adjacent segment of the target nucleic acid molecule directly 3' to the first probe. The two juxtaposed probes hybridize to the target nucleic acid molecule, and are ligated in the presence of a linking agent such as a ligase if there is perfect complementarity between the 3' most nucleotide of the first probe with the SNP site. If there is a mismatch, ligation would not occur. After the reaction, the ligated probes are separated from the target nucleic acid molecule, and detected as indicators of the presence of a SNP.

The following patents, patent applications, and published international patent applications, which are all hereby incorporated by reference, provide additional information pertaining to techniques for carrying out various types of OLA: U.S. Patent Nos. 6027889, 6268148, 5494810, 5830711, and 6054564 describe OLA strategies for performing SNP detection; WO 97/31256 and WO 00/56927 describe OLA strategies for performing SNP detection using universal arrays, wherein a zipcode sequence can be introduced into one of the hybridization probes, and the resulting product, or amplified product, hybridized to a universal zip code array; U.S. application US01/17329 (and 09/584,905) describes OLA (or LDR) followed by PCR, wherein zipcodes are incorporated into OLA probes, and amplified PCR products are determined by electrophoretic or universal zipcode array readout; U.S. applications 60/427818, 60/445636, and 60/445494 describe SNPllex methods and software for multiplexed SNP detection using OLA followed by PCR, wherein zipcodes are incorporated into OLA probes, and amplified PCR products are hybridized with a zipchute reagent, and the identity of the SNP determined from electrophoretic readout of the zipchute. In some embodiments, OLA is carried out prior to PCR (or another method of nucleic acid amplification). In other embodiments, PCR (or another method of nucleic acid amplification) is carried out prior to OLA.

Another method for SNP genotyping is based on mass spectrometry. Mass spectrometry takes advantage of the unique mass of each of the four nucleotides of DNA. SNPs can be unambiguously genotyped by mass spectrometry by measuring the differences in the mass of nucleic acids having alternative SNP alleles. MALDI-TOF (Matrix Assisted Laser Desorption Ionization – Time of Flight) mass spectrometry technology is preferred for extremely precise determinations of molecular mass, such as SNPs. Numerous approaches to SNP analysis have been developed based on mass spectrometry. Preferred mass spectrometry-based methods of

SNP genotyping include primer extension assays, which can also be utilized in combination with other approaches, such as traditional gel-based formats and microarrays.

Typically, the primer extension assay involves designing and annealing a primer to a template PCR amplicon upstream (5') from a target SNP position. A mix of dideoxynucleotide triphosphates (ddNTPs) and/or deoxynucleotide triphosphates (dNTPs) are added to a reaction mixture containing template (e.g., a SNP-containing nucleic acid molecule which has typically been amplified, such as by PCR), primer, and DNA polymerase. Extension of the primer terminates at the first position in the template where a nucleotide complementary to one of the ddNTPs in the mix occurs. The primer can be either immediately adjacent (i.e., the nucleotide at the 3' end of the primer hybridizes to the nucleotide next to the target SNP site) or two or more nucleotides removed from the SNP position. If the primer is several nucleotides removed from the target SNP position, the only limitation is that the template sequence between the 3' end of the primer and the SNP position cannot contain a nucleotide of the same type as the one to be detected, or this will cause premature termination of the extension primer. Alternatively, if all four ddNTPs alone, with no dNTPs, are added to the reaction mixture, the primer will always be extended by only one nucleotide, corresponding to the target SNP position. In this instance, primers are designed to bind one nucleotide upstream from the SNP position (i.e., the nucleotide at the 3' end of the primer hybridizes to the nucleotide that is immediately adjacent to the target SNP site on the 5' side of the target SNP site). Extension by only one nucleotide is preferable, as it minimizes the overall mass of the extended primer, thereby increasing the resolution of mass differences between alternative SNP nucleotides. Furthermore, mass-tagged ddNTPs can be employed in the primer extension reactions in place of unmodified ddNTPs. This increases the mass difference between primers extended with these ddNTPs, thereby providing increased sensitivity and accuracy, and is particularly useful for typing heterozygous base positions. Mass-tagging also alleviates the need for intensive sample-preparation procedures and decreases the necessary resolving power of the mass spectrometer.

The extended primers can then be purified and analyzed by MALDI-TOF mass spectrometry to determine the identity of the nucleotide present at the target SNP position. In one method of analysis, the products from the primer extension reaction are combined with light absorbing crystals that form a matrix. The matrix is then hit with an energy source such as a laser to ionize and desorb the nucleic acid molecules into the gas-phase. The ionized molecules are then ejected into a flight tube and accelerated down the tube towards a detector. The time between the ionization event, such as a laser pulse, and collision of the molecule with the detector is the time of flight of that molecule. The time of flight is precisely correlated with the.

mass-to-charge ratio (m/z) of the ionized molecule. Ions with smaller m/z travel down the tube faster than ions with larger m/z and therefore the lighter ions reach the detector before the heavier ions. The time-of-flight is then converted into a corresponding, and highly precise, m/z . In this manner, SNPs can be identified based on the slight differences in mass, and the corresponding time of flight differences, inherent in nucleic acid molecules having different nucleotides at a single base position. For further information regarding the use of primer extension assays in conjunction with MALDI-TOF mass spectrometry for SNP genotyping, see, e.g., Wise *et al.*, "A standard protocol for single nucleotide primer extension in the human genome using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry," *Rapid Commun. Mass Spectrom.* 17(11):1195-202 (2003).

The following references provide further information describing mass spectrometry-based methods for SNP genotyping: Bocker, "SNP and mutation discovery using base-specific cleavage and MALDI-TOF mass spectrometry," *Bioinformatics* 19 Suppl 1:I44-I53 (Jul. 2003); Storm *et al.*, "MALDI-TOF mass spectrometry-based SNP genotyping," *Methods Mol. Biol.* 212:241-62 (2003); Jurinke *et al.*, "The use of MassARRAY technology for high throughput genotyping," *Adv. Biochem. Eng. Biotechnol.* 77:57-74 (2002); and Jurinke *et al.*, "Automated genotyping using the DNA MassArray technology," *Methods Mol. Biol.* 187:179-92 (2002).

SNPs can also be scored by direct DNA sequencing. A variety of automated sequencing procedures can be utilized (*Biotechniques* 19:448 [1995]), including sequencing by mass spectrometry (see, e.g., PCT International Publication No. WO94/16101; Cohen *et al.*, *Adv. Chromatogr.* 36:127-162 [1996]; and Griffin *et al.*, *Appl. Biochem. Biotechnol.* 38:147-159 [1993]). The nucleic acid sequences of the present invention enable one of ordinary skill in the art to readily design sequencing primers for such automated sequencing procedures. Commercial instrumentation, such as the Applied Biosystems 377, 3100, 3700, 3730, and 3730xl DNA Analyzers (Foster City, CA), is commonly used in the art for automated sequencing.

Other methods that can be used to genotype the SNPs of the present invention include single-strand conformational polymorphism (SSCP), and denaturing gradient gel electrophoresis (DGGE) (Myers *et al.*, *Nature* 313:495 [1985]). SSCP identifies base differences by alteration in electrophoretic migration of single stranded PCR products, as described in Orita *et al.*, *Proc. Nat. Acad.* Single-stranded PCR products can be generated by heating or otherwise denaturing double stranded PCR products. Single-stranded nucleic acids may refold or form secondary structures that are partially dependent on the base sequence. The different electrophoretic mobilities of single-stranded amplification products are related to base-sequence differences at SNP positions. DGGE differentiates SNP alleles based on the different sequence-dependent stabilities and

melting properties inherent in polymorphic DNA and the corresponding differences in electrophoretic migration patterns in a denaturing gradient gel ("PCR Technology," *Principles and Applications for DNA Amplification* Chapter 7, ed. Erlich, W.H. Freeman and Co., New York, [1992]).

5 Sequence-specific ribozymes (U.S. Patent No. 5,498,531) can also be used to score SNPs based on the development or loss of a ribozyme cleavage site. Perfectly matched sequences can be distinguished from mismatched sequences by nuclease cleavage digestion assays or by differences in melting temperature. If the SNP affects a restriction enzyme cleavage site, the SNP can be identified by alterations in restriction enzyme digestion patterns, and the
10 corresponding changes in nucleic acid fragment lengths determined by gel electrophoresis.

 SNP genotyping can include the steps of, for example, collecting a biological sample from a human subject (e.g., sample of tissues, cells, fluids, secretions, etc.), isolating nucleic acids (e.g., genomic DNA, mRNA or both) from the cells of the sample, contacting the nucleic acids with one or more primers which specifically hybridize to a region of the isolated nucleic
15 acid containing a target SNP under conditions such that hybridization and amplification of the target nucleic acid region occurs, and determining the nucleotide present at the SNP position of interest, or, in some assays, detecting the presence or absence of an amplification product (assays can be designed so that hybridization and/or amplification will only occur if a particular SNP allele is present or absent). In some assays, the size of the amplification product is detected and
20 compared to the length of a control sample; for example, deletions and insertions can be detected by a change in size of the amplified product compared to a normal genotype.

 SNP genotyping is useful for numerous practical applications, as described below. Examples of such applications include, but are not limited to, SNP-disease association analysis, disease predisposition screening, disease diagnosis, disease prognosis, disease progression
25 monitoring, determining therapeutic strategies based on an individual's genotype ("pharmacogenomics"), developing therapeutic agents based on SNP genotypes associated with a disease or likelihood of responding to a drug, stratifying a patient population for clinical trial for a treatment regimen, predicting the likelihood that an individual will experience toxic side effects from a therapeutic agent, and human identification applications such as forensics.

Analysis of Genetic Association Between SNPs and Phenotypic Traits

 SNP genotyping for disease diagnosis, disease predisposition screening, disease prognosis, determining drug responsiveness (pharmacogenomics), drug toxicity screening, and

other uses described herein, typically relies on initially establishing a genetic association between one or more specific SNPs and the particular phenotypic traits of interest.

Different study designs may be used for genetic association studies (*Modern Epidemiology* 609-622, Lippincott Williams & Wilkins [1998]). Observational studies are most frequently carried out in which the response of the patients is not interfered with. The first type of observational study identifies a sample of persons in whom the suspected cause of the disease is present and another sample of persons in whom the suspected cause is absent, and then the frequency of development of disease in the two samples is compared. These sampled populations are called cohorts, and the study is a prospective study. The other type of observational study is case-control or a retrospective study. In typical case-control studies, samples are collected from individuals with the phenotype of interest (cases) such as certain manifestations of a disease, and from individuals without the phenotype (controls) in a population (target population) that conclusions are to be drawn from. Then the possible causes of the disease are investigated retrospectively. As the time and costs of collecting samples in case-control studies are considerably less than those for prospective studies, case-control studies are the more commonly used study design in genetic association studies, at least during the exploration and discovery stage.

In both types of observational studies, there may be potential confounding factors that should be taken into consideration. Confounding factors are those that are associated with both the real cause(s) of the disease and the disease itself, and they include demographic information such as age, gender, ethnicity as well as environmental factors. When confounding factors are not matched in cases and controls in a study, and are not controlled properly, spurious association results can arise. If potential confounding factors are identified, they should be controlled for by analysis methods explained below.

In a genetic association study, the cause of interest to be tested is a certain allele or a SNP or a combination of alleles or a haplotype from several SNPs. Thus, tissue specimens (e.g., whole blood) from the sampled individuals may be collected and genomic DNA genotyped for the SNP(s) of interest. In addition to the phenotypic trait of interest, other information such as demographic (e.g., age, gender, ethnicity, etc.), clinical, and environmental information that may influence the outcome of the trait can be collected to further characterize and define the sample set. In many cases, these factors are known to be associated with diseases and/or SNP allele frequencies. There are likely gene-environment and/or gene-gene interactions as well. Analysis methods to address gene-environment and gene-gene interactions (for example, the effects of the

presence of both susceptibility alleles at two different genes can be greater than the effects of the individual alleles at two genes combined) are discussed below.

After all the relevant phenotypic and genotypic information has been obtained, statistical analyses are carried out to determine if there is any significant correlation between the presence of an allele or a genotype with the phenotypic characteristics of an individual. Preferably, data inspection and cleaning are first performed before carrying out statistical tests for genetic association. Epidemiological and clinical data of the samples can be summarized by descriptive statistics with tables and graphs. Data validation is preferably performed to check for data completion, inconsistent entries, and outliers. Chi-squared tests and t-tests (Wilcoxon rank-sum tests if distributions are not normal) may then be used to check for significant differences between cases and controls for discrete and continuous variables, respectively. To ensure genotyping quality, Hardy-Weinberg disequilibrium tests can be performed on cases and controls separately. Significant deviation from Hardy-Weinberg equilibrium (HWE) in both cases and controls for individual markers can be indicative of genotyping errors. If HWE is violated in a majority of markers, it is indicative of population substructure that should be further investigated. Moreover, Hardy-Weinberg disequilibrium in cases only can indicate genetic association of the markers with the disease (*Genetic Data Analysis*, Weir B., Sinauer [1990]).

To test whether an allele of a single SNP is associated with the case or control status of a phenotypic trait, one skilled in the art can compare allele frequencies in cases and controls. Standard chi-squared tests and Fisher exact tests can be carried out on a 2x2 table (2 SNP alleles x 2 outcomes in the categorical trait of interest). To test whether genotypes of a SNP are associated, chi-squared tests can be carried out on a 3x2 table (3 genotypes x 2 outcomes). Score tests are also carried out for genotypic association to contrast the three genotypic frequencies (major homozygotes, heterozygotes and minor homozygotes) in cases and controls, and to look for trends using 3 different modes of inheritance, namely dominant (with contrast coefficients 2, -1, -1), additive (with contrast coefficients 1, 0, -1) and recessive (with contrast coefficients 1, 1, -2). Odds ratios for minor versus major alleles, and odds ratios for heterozygote and homozygote variants versus the wild type genotypes are calculated with the desired confidence limits, usually 95%.

In order to control for confounders and to test for interaction and effect modifiers, stratified analyses may be performed using stratified factors that are likely to be confounding, including demographic information such as age, ethnicity, and gender, or an interacting element or effect modifier, such as a known major gene (e.g., APOE for AD or HLA genes for autoimmune diseases), or environmental factors such as smoking in lung cancer. Stratified

association tests may be carried out using Cochran-Mantel-Haenszel tests that take into account the ordinal nature of genotypes with 0, 1, and 2 variant alleles. Exact tests by StatXact may also be performed when computationally possible. Another way to adjust for confounding effects and test for interactions is to perform stepwise multiple logistic regression analysis using statistical packages such as SAS or R. Logistic regression is a model-building technique in which the best fitting and most parsimonious model is built to describe the relation between the dichotomous outcome (for instance, getting a certain disease or not) and a set of independent variables (for instance, genotypes of different associated genes, and the associated demographic and environmental factors). The most common model is one in which the logit transformation of the odds ratios is expressed as a linear combination of the variables (main effects) and their cross-product terms (interactions) (Applied Logistic Regression, Hosmer and Lemeshow, Wiley [2000]). To test whether a certain variable or interaction is significantly associated with the outcome, coefficients in the model are first estimated and then tested for statistical significance of their departure from zero.

In addition to performing association tests one marker at a time, haplotype association analysis may also be performed to study a number of markers that are closely linked together. Haplotype association tests can have better power than genotypic or allelic association tests when the tested markers are not the disease-causing mutations themselves but are in linkage disequilibrium with such mutations. The test will even be more powerful if the disease is indeed caused by a combination of alleles on a haplotype (e.g., APOE is a haplotype formed by 2 SNPs that are very close to each other). In order to perform haplotype association effectively, marker-marker linkage disequilibrium measures, both D' and R^2 , are typically calculated for the markers within a gene to elucidate the haplotype structure. Recent studies (Daly *et al.*, *Nature Genetics* 29, 232-235 [2001]) in linkage disequilibrium indicate that SNPs within a gene are organized in block pattern, and a high degree of linkage disequilibrium exists within blocks and very little linkage disequilibrium exists between blocks. Haplotype association with the disease status can be performed using such blocks once they have been elucidated.

Haplotype association tests can be carried out in a similar fashion as the allelic and genotypic association tests. Each haplotype in a gene is analogous to an allele in a multi-allelic marker. One skilled in the art can either compare the haplotype frequencies in cases and controls or test genetic association with different pairs of haplotypes. It has been proposed (Schaid *et al.*, *Am. J. Hum. Genet.* 70, 425-434 [2002]) that score tests can be done on haplotypes using the program "haplo.score." In that method, haplotypes are first inferred by EM algorithm and score

tests are carried out with a generalized linear model (GLM) framework that allows the adjustment of other factors.

An important decision in the performance of genetic association tests is the determination of the significance level at which significant association can be declared when the P value of the tests reaches that level. In an exploratory analysis where positive hits will be followed up in subsequent confirmatory testing, an unadjusted P value < 0.1 (a significance level on the lenient side) may be used for generating hypotheses for significant association of a SNP with certain phenotypic characteristics of a disease. It is preferred that a P value < 0.05 (a significance level traditionally used in the art) is achieved in order for a SNP to be considered to have an association with a disease. It is more preferred that a P value < 0.01 (a significance level on the stringent side) is achieved for an association to be declared. When hits are followed up in confirmatory analyses in more samples of the same source or in different samples from different sources, adjustment for multiple testing will be performed as to avoid excess number of hits while maintaining the experiment-wise error rates at 0.05. While there are different methods to adjust for multiple testing to control for different kinds of error rates, a commonly used but rather conservative method is Bonferroni correction to control the experiment-wise or family-wise error rate (Westfall *et al.*, *Multiple comparisons and multiple tests*, SAS Institute [1999]). Permutation tests to control for the false discovery rates, FDR, can be more powerful (Benjamini and Hochberg, *Journal of the Royal Statistical Society Series B* 57, 1289-1300 [1995], *Resampling-based Multiple Testing*, Westfall and Young, Wiley [1993]). Such methods to control for multiplicity would be preferred when the tests are dependent and controlling for false discovery rates is sufficient as opposed to controlling for the experiment-wise error rates.

In replication studies using samples from different populations after statistically significant markers have been identified in the exploratory stage, meta-analyses can then be performed by combining evidence of different studies (*Modern Epidemiology* 643-673, Lippincott Williams & Wilkins [1998]). If available, association results known in the art for the same SNPs can be included in the meta-analyses.

Since both genotyping and disease status classification can involve errors, sensitivity analyses may be performed to see how odds ratios and P values would change upon various estimates on genotyping and disease classification error rates.

It has been well known that subpopulation-based sampling bias between cases and controls can lead to spurious results in case-control association studies (Ewens and Spielman, *Am. J. Hum. Genet.* 62, 450-458 [1995]) when prevalence of the disease is associated with different subpopulation groups. Such bias can also lead to a loss of statistical power in genetic

association studies. To detect population stratification, Pritchard and Rosenberg (Pritchard *et al.*, *Am. J. Hum. Gen.* 65:220-228 [1999]) suggested typing markers that are unlinked to the disease and using results of association tests on those markers to determine whether there is any population stratification. When stratification is detected, the genomic control (GC) method as
 5 proposed by Devlin and Roeder (Devlin *et al.*, *Biometrics* 55:997-1004 [1999]) can be used to adjust for the inflation of test statistics due to population stratification. GC method is robust to changes in population structure levels as well as being applicable to DNA pooling designs (Devlin *et al.*, *Genet. Epidemiol.* 21:273-284 [2001]).

While Pritchard's method recommended using 15-20 unlinked microsatellite markers, it
 10 suggested using more than 30 biallelic markers to get enough power to detect population stratification. For the GC method, it has been shown (Bacanu *et al.*, *Am. J. Hum. Genet.* 66:1933-1944 [2000]) that about 60-70 biallelic markers are sufficient to estimate the inflation factor for the test statistics due to population stratification. Hence, 70 intergenic SNPs can be chosen in unlinked regions as indicated in a genome scan (Kehoe *et al.*, *Hum. Mol. Genet.* 8:237-
 15 245 [1999]).

Once individual risk factors, genetic or non-genetic, have been found for the predisposition to disease, the next step is to set up a classification/prediction scheme to predict the category (for instance, disease or no-disease) that an individual will be in depending on his genotypes of associated SNPs and other non-genetic risk factors. Logistic regression for discrete
 20 trait and linear regression for continuous trait are standard techniques for such tasks (*Applied Regression Analysis*, Draper and Smith, Wiley [1998]). Moreover, other techniques can also be used for setting up classification. Such techniques include, but are not limited to, MART, CART, neural network, and discriminant analyses that are suitable for use in comparing the performance of different methods (*The Elements of Statistical Learning*, Hastie, Tibshirani & Friedman,
 25 Springer [2002]).

Disease Diagnosis and Predisposition Screening

Information on association/correlation between genotypes and disease-related phenotypes can be exploited in several ways. For example, in the case of a highly statistically significant
 30 association between one or more SNPs with predisposition to a disease for which treatment is available, detection of such a genotype pattern in an individual may justify immediate administration of treatment, or at least the institution of regular monitoring of the individual. Detection of the susceptibility alleles associated with serious disease in a couple contemplating having children may also be valuable to the couple in their reproductive decisions. In the case of

a weaker but still statistically significant association between a SNP and a human disease, immediate therapeutic intervention or monitoring may not be justified after detecting the susceptibility allele or SNP. Nevertheless, the subject can be motivated to begin simple life-style changes (e.g., diet, exercise) that can be accomplished at little or no cost to the individual but would confer potential benefits in reducing the risk of developing conditions for which that individual may have an increased risk by virtue of having the susceptibility allele(s).

The SNPs of the invention may contribute to AD in an individual in different ways. Some polymorphisms occur within a protein coding sequence and contribute to disease phenotype by affecting protein structure. Other polymorphisms occur in noncoding regions but may exert phenotypic effects indirectly via influence on, for example, replication, transcription, and/or translation. A single SNP may affect more than one phenotypic trait. Likewise, a single phenotypic trait may be affected by multiple SNPs in different genes.

As used herein, the terms "diagnose," "diagnosis," and "diagnostics" include, but are not limited to any of the following: detection of AD that an individual may presently have, predisposition/susceptibility screening (i.e., determining the increased risk of an individual in developing AD in the future, or determining whether an individual has a decreased risk of developing AD in the future), determining a particular type or subclass of AD in an individual known to have AD, confirming or reinforcing a previously made diagnosis of AD, pharmacogenomic evaluation of an individual to determine which therapeutic strategy that individual is most likely to positively respond to or to predict whether a patient is likely to respond to a particular treatment, predicting whether a patient is likely to experience toxic effects from a particular treatment or therapeutic compound, and evaluating the future prognosis of an individual having AD. Such diagnostic uses are based on the SNPs individually or in a unique combination or SNP haplotypes of the present invention.

Haplotypes are particularly useful in that, for example, fewer SNPs can be genotyped to determine if a particular genomic region harbors a locus that influences a particular phenotype, such as in linkage disequilibrium-based SNP association analysis.

Linkage disequilibrium (LD) refers to the co-inheritance of alleles (e.g., alternative nucleotides) at two or more different SNP sites at frequencies greater than would be expected from the separate frequencies of random occurrence of each allele in a given population. The expected frequency of co-occurrence of two alleles that are inherited independently is the frequency of the first allele multiplied by the frequency of the second allele. Alleles that co-occur at expected frequencies are said to be in "linkage equilibrium." In contrast, LD refers to any non-random genetic association between allele(s) at two or more different SNP sites, which

is generally due to the physical proximity of the two loci along a chromosome. LD can occur when two or more SNPs sites are in close physical proximity to each other on a given chromosome and therefore alleles at these SNP sites will tend to remain unseparated for multiple generations, with the consequence that a particular nucleotide (allele) at one SNP site will show a non-random association with a particular nucleotide (allele) at another SNP site located nearby. Hence, genotyping one of the SNP sites will give almost the same information as genotyping the other SNP site that is in LD. The physical area of the chromosome that contains SNPs in LD with each other is referred to as an LD block.

Various degrees of LD can be encountered between two or more SNPs with the result being that some SNPs are more closely associated (i.e., in stronger LD) than others. Furthermore, the physical distance over which LD extends along a chromosome differs between different regions of the genome, and therefore the degree of physical separation between two or more SNP sites necessary for LD to occur can differ between different regions of the genome.

For diagnostic purposes and similar uses, if a particular SNP site is found to be useful for diagnosing AD (e.g., has a significant statistical association with the condition and/or is recognized as a causative polymorphism for the condition), then the skilled artisan would recognize that other SNP sites which are in LD with this SNP site would also be useful for diagnosing the condition. Thus, polymorphisms (e.g., SNPs and/or haplotypes) that are not the actual disease-causing (causative) polymorphisms, but are in LD with such causative polymorphisms, are also useful. In such instances, the genotype of the polymorphism(s) that is/are in LD with the causative polymorphism is predictive of the genotype of the causative polymorphism and, consequently, predictive of the phenotype (e.g., AD) that is influenced by the causative SNP(s). Therefore, polymorphic markers that are in LD with causative polymorphisms are useful as diagnostic markers, and are particularly useful when the actual causative polymorphism(s) is/are unknown.

Examples of polymorphisms that can be in LD with one or more causative polymorphisms (and/or in LD with one or more polymorphisms that have a significant statistical association with a condition) and therefore useful for diagnosing the same condition that the causative/associated SNP(s) is used to diagnose, include, for example, other SNPs in the same gene, protein-coding, or mRNA transcript-coding region as the causative/associated SNP, other SNPs in the same exon or same intron as the causative/associated SNP, other SNPs in the same haplotype block as the causative/associated SNP, other SNPs in the same intergenic region as the causative/associated SNP, SNPs that are outside, but near a gene (e.g., within 6kb on either side, 5' or 3', of a gene boundary) that harbors a causative/associated SNP, etc. Such useful LD SNPs

can be selected from among the SNPs disclosed in Tables 1-2, and 11 (LD SNP Table), for example.

- Linkage disequilibrium in the human genome is reviewed in the following references:
- Wall *et al.*, "Haplotype blocks and linkage disequilibrium in the human genome," *Nat Rev Genet.* 2003 Aug;4(8):587-97 (Aug. 2003); Garner *et al.*, "On selecting markers for association studies: patterns of linkage disequilibrium between two and three diallelic loci," *Genet Epidemiol.* 2003 Jan;24(1):57-67 (Jan. 2003); Ardlie *et al.*, "Patterns of linkage disequilibrium in the human genome," *Nat Rev Genet.* 2002 Apr;3(4):299-309 (Apr. 2002); (erratum in *Nat Rev Genet* 2002 Jul;3(7):566 (Jul. 2002); and Remm *et al.*, "High-density genotyping and linkage disequilibrium in the human genome using chromosome 22 as a model," *Curr Opin Chem Biol.* 2002 Feb;6(1):24-30 (Feb. 2002); J.B.S. Haldane, "JBS (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors," *J Genet* 8:299-309 (1919); G. Mendel, *G. (1866) Versuche über Pflanzen-Hybriden. Verhandlungen des naturforschenden Vereines in Brünn [(Proceedings of the Natural History Society of Brünn)]* (1866); Lewin B (1990) *Genes IV*, B. Lewin, ed., . Oxford University Press, N.Y.ew York, USA (1990); D.L. Hartl DL and A.G. Clark AG (1989) *Principles of Population Genetics 2nd ed.*, . Sinauer Associates, Inc., Ma Sunderland, Mass., USA (1989); J.H. Gillespie JH (2004) *Population Genetics: A Concise Guide. 2nd ed.*,. Johns Hopkins University Press. (2004) USA; R.C. Lewontin, "RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models," *Genetics* 49:49-67 (1964); P.G. Hoel, PG (1954) *Introduction to Mathematical Statistics 2nd ed.*, John Wiley & Sons, Inc., N.Y. New York, USA (1954); R.R. Hudson, RR "(2001) Two-locus sampling distributions and their application," *Genetics* 159:1805-1817 (2001); A.P. Dempster AP, N.M. Laird, D.B. NM, Rubin, "DB (1977) Maximum likelihood from incomplete data via the EM algorithm," *J R Stat Soc* 39:1-38 (1977); L. Excoffier L, M. Slatkin, M "(1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population," *Mol Biol Evol* 12(5):921-927 (1995); D.A. Tregouet DA, S. Escolano S, L. Tirët L, A. Mallet A, J.L. Golmard, JL "(2004) A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm," *Ann Hum Genet* 68(Pt 2):165-177 (2004); A.D. Long AD and C.H. Langley CH, " (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits," *Genome Research* 9:720-731 (1999); A. Agresti, A (1990) *Categorical Data Analysis*,. John Wiley & Sons, Inc., N.Y. New York, USA (1990); K. Lange, K (1997) *Mathematical and Statistical Methods for Genetic Analysis*, . Springer-Verlag New York, Inc., N.Y. New York, USA (1997); The International HapMap Consortium, "(2003) The International HapMap Project," *Nature*

426:789-796 (2003); The International HapMap Consortium, " (2005) A haplotype map of the human genome,". *Nature* 437:1299-1320 (2005); G.A. Thorisson GA, A.V. Smith AV, L. Krishnan L, L.D. Stein LD (2005), "The International HapMap Project Web Site,". *Genome Research* 15:1591-1593 (2005); G. McVean, C.C.A. G, Spencer CCA, R. Chaix R (2005),
 5 "Perspectives on human genetic variation from the HapMap project,". *PLoS Genetics* 1(4):413-418 (2005); J.N. Hirschhorn JN, M.J. Daly, MJ "(2005) Genome-wide association studies for common diseases and complex traits,". *Nat Genet* 6:95-108 (2005); S.J. Schrodin, " SJ (2005) A probabilistic approach to large-scale association scans: a semi-Bayesian method to detect disease-predisposing alleles,". *SAGMB* 4(1):31 (2005); W.Y.S. Wang WYS, B.J. Barratt BJ,
 10 D.G. Clayton DG, J.A. Todd, "JA (2005) Genome-wide association studies: theoretical and practical concerns,". *Nat Rev Genet* 6:109-118 (2005); J.K. . Pritchard JK, M. Przeworski, " M (2001) Linkage disequilibrium in humans: models and data,". *Am J Hum Genet* 69:1-14 (2001).

As discussed above, one aspect of the present invention is the discovery that SNPs which are in certain LD distance with the interrogated SNP can also be used as valid markers for
 15 identifying an increased or decreased risks of having or developing AD. As used herein, the term "interrogated SNP" refers to SNPs that have been found to be associated with an increased or decreased risk of disease using genotyping results and analysis, or other appropriate experimental method as exemplified in the working examples described in this application. As used herein, the term "LD SNP" refers to a SNP that has been characterized as a SNP associating with an
 20 increased or decreased risk of diseases due to their being in LD with the "interrogated SNP" under the methods of calculation described in the application. Below, applicants describe the methods of calculation with which one of ordinary skilled in the art may determine if a particular SNP is in LD with an interrogated SNP. The parameter r^2 is commonly used in the genetics art to characterize the extent of linkage disequilibrium between markers (Hudson, 2001). As used
 25 herein, the term "in LD with" refers to a particular SNP that is measured at above the threshold of a parameter such as r^2 with an interrogated SNP.

It is now common place to directly observe genetic variants in a sample of chromosomes obtained from a population. Suppose one has genotype data at two genetic markers located on the same chromosome, for the markers A and B . Further suppose that two alleles segregate at
 30 each of these two markers such that alleles A_1 and A_2 can be found at marker A and alleles B_1 and B_2 at marker B . Also assume that these two markers are on a human autosome. If one is to examine a specific individual and find that they are heterozygous at both markers, such that their two-marker genotype is $A_1A_2B_1B_2$, then there are two possible configurations: the individual in question could have the alleles A_1B_1 on one chromosome and A_2B_2 on the remaining

chromosome; alternatively, the individual could have alleles A_1B_2 on one chromosome and A_2B_1 on the other. The arrangement of alleles on a chromosome is called a haplotype. In this illustration, the individual could have haplotypes A_1B_1/A_2B_2 or A_1B_2/A_2B_1 (see Hartl and Clark (1989) for a more complete description). The concept of linkage equilibrium relates the frequency of haplotypes to the allele frequencies.

Assume that a sample of individuals is selected from a larger population. Considering the two markers described above, each having two alleles, there are four possible haplotypes: A_1B_1 , A_1B_2 , A_2B_1 and A_2B_2 . Denote the frequencies of these four haplotypes with the following notation.

$$P_{11} = \text{freq}(A_1B_1) \quad (1)$$

$$P_{12} = \text{freq}(A_1B_2) \quad (2)$$

$$P_{21} = \text{freq}(A_2B_1) \quad (3)$$

$$P_{22} = \text{freq}(A_2B_2) \quad (4)$$

The allele frequencies at the two markers are then the sum of different haplotype frequencies, it is straightforward to write down a similar set of equations relating single-marker allele frequencies to two-marker haplotype frequencies:

$$p_1 = \text{freq}(A_1) = P_{11} + P_{12} \quad (5)$$

$$p_2 = \text{freq}(A_2) = P_{21} + P_{22} \quad (6)$$

$$q_1 = \text{freq}(B_1) = P_{11} + P_{21} \quad (7)$$

$$q_2 = \text{freq}(B_2) = P_{12} + P_{22} \quad (8)$$

Note that the four haplotype frequencies and the allele frequencies at each marker must sum to a frequency of 1.

$$P_{11} + P_{12} + P_{21} + P_{22} = 1 \quad (9)$$

$$p_1 + p_2 = 1 \quad (10)$$

$$q_1 + q_2 = 1 \quad (11)$$

If there is no correlation between the alleles at the two markers, one would expect that the frequency of the haplotypes would be approximately the product of the composite alleles. Therefore,

$$5 \quad P_{11} \approx p_1 q_1 \quad (12)$$

$$P_{12} \approx p_1 q_2 \quad (13)$$

$$P_{21} \approx p_2 q_1 \quad (14)$$

$$P_{22} \approx p_2 q_2 \quad (15)$$

10 These approximating equations (12)-(15) represent the concept of linkage equilibrium where there is independent assortment between the two markers – the alleles at the two markers occur together at random. These are represented as approximations because linkage equilibrium and linkage disequilibrium are concepts typically thought of as properties of a sample of chromosomes; and as such they are susceptible to stochastic fluctuations due to the sampling
15 process. Empirically, many pairs of genetic markers will be in linkage equilibrium, but certainly not all pairs.

Having established the concept of linkage equilibrium above, applicants can now describe the concept of linkage disequilibrium (LD), which is the deviation from linkage equilibrium. Since the frequency of the $A_1 B_1$ haplotype is approximately the product of the allele frequencies
20 for A_1 and B_1 under the assumption of linkage equilibrium as stated mathematically in (12), a simple measure for the amount of departure from linkage equilibrium is the difference in these two quantities, D ,

$$D = P_{11} - p_1 q_1 \quad (16)$$

25

$D = 0$ indicates perfect linkage equilibrium. Substantial departures from $D = 0$ indicates LD in the sample of chromosomes examined. Many properties of D are discussed in Lewontin (1964) including the maximum and minimum values that D can take. Mathematically, using basic algebra, it can be shown that D can also be written solely in terms of haplotypes:

30

$$D = P_{11}P_{22} - P_{12}P_{21} \quad (17)$$

If one transforms D by squaring it and subsequently dividing by the product of the allele frequencies of A_1 , A_2 , B_1 and B_2 , the resulting quantity, called r^2 , is equivalent to the square of the Pearson's correlation coefficient commonly used in statistics (*e.g.* Hoel, 1954).

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2} \quad (18)$$

As with D , values of r^2 close to 0 indicate linkage equilibrium between the two markers examined in the sample set. As values of r^2 increase, the two markers are said to be in linkage disequilibrium. The range of values that r^2 can take are from 0 to 1. $r^2 = 1$ when there is a perfect correlation between the alleles at the two markers.

In addition, the quantities discussed above are sample-specific. And as such, it is necessary to formulate notation specific to the samples studied. In the approach discussed here, three types of samples are of primary interest: (i) a sample of chromosomes from individuals affected by a disease-related phenotype (cases), (ii) a sample of chromosomes obtained from individuals not affected by the disease-related phenotype (controls), and (iii) a standard sample set used for the construction of haplotypes and calculation pairwise linkage disequilibrium. For the allele frequencies used in the development of the method described below, an additional subscript will be added to denote either the case or control sample sets.

$$p_{1,cs} = \text{freq}(A_1 \text{ in cases}) \quad (19)$$

$$p_{2,cs} = \text{freq}(A_2 \text{ in cases}) \quad (20)$$

$$q_{1,cs} = \text{freq}(B_1 \text{ in cases}) \quad (21)$$

$$q_{2,cs} = \text{freq}(B_2 \text{ in cases}) \quad (22)$$

Similarly,

$$p_{1,ct} = \text{freq}(A_1 \text{ in controls}) \quad (23)$$

$$p_{2,ct} = \text{freq}(A_2 \text{ in controls}) \quad (24)$$

$$q_{1,ct} = \text{freq}(B_1 \text{ in controls}) \quad (25)$$

$$q_{2,ct} = \text{freq}(B_2 \text{ in controls}) \quad (26)$$

As a well-accepted sample set is necessary for robust linkage disequilibrium calculations, data obtained from the International HapMap project (The International HapMap Consortium 2003, 2005; Thorisson et al, 2005; McVean et al, 2005) can be used for the calculation of pairwise r^2 values. Indeed, the samples genotyped for the International HapMap Project were
 5 selected to be representative examples from various human sub-populations with sufficient numbers of chromosomes examined to draw meaningful and robust conclusions from the patterns of genetic variation observed. The International HapMap project website (*hapmap.org*) contains a description of the project, methods utilized and samples examined. It is useful to examine empirical data to get a sense of the patterns present in such data.

10 Haplotype frequencies were explicit arguments in equation (18) above. However, knowing the 2-marker haplotype frequencies requires that phase to be determined for doubly heterozygous samples. When phase is unknown in the data examined, various algorithms can be used to infer phase from the genotype data. This issue was discussed earlier where the doubly heterozygous individual with a 2-SNP genotype of $A_1A_2B_1B_2$ could have one of two different
 15 sets of chromosomes: A_1B_1/A_2B_2 or A_1B_2/A_2B_1 . One such algorithm to estimate haplotype frequencies is the expectation-maximization (EM) algorithm first formalized by Dempster et al (1977). This algorithm is often used in genetics to infer haplotype frequencies from genotype data (e.g. Excoffier and Slatkin (, 1995); Tregouet et al (2004)). It should be noted that for the two-SNP case explored here, EM algorithms have very little error provided that the allele
 20 frequencies and sample sizes are not too small. The impact on r^2 values is typically negligible.

As correlated genetic markers share information, interrogation of SNP markers in LD with a disease-associated SNP marker can also have sufficient power to detect disease association (Long and Langley (, 1999)). The relationship between the power to directly find disease-associated alleles and the power to indirectly detect disease-association was investigated
 25 by Pritchard and Przeworski (2001). In a straight-forward derivation, it can be shown that the power to detect disease association indirectly at a marker locus in linkage disequilibrium with a disease-association locus is approximately the same as the power to detect disease-association directly at the disease- association locus if the sample size is increased by a factor of $\frac{1}{r^2}$ (the reciprocal of equation 18) at the marker in comparison with the disease- association locus.

30 Therefore, if one calculated the power to detect disease-association indirectly with an experiment having N samples, then equivalent power to directly detect disease-association (at the actual disease-susceptibility locus) would necessitate an experiment using approximately r^2N samples. This elementary relationship between power, sample size and linkage

disequilibrium can be used to derive an r^2 threshold value useful in determining whether or not genotyping markers in linkage disequilibrium with a SNP marker directly associated with disease status has enough power to indirectly detect disease-association.

To commence a derivation of the power to detect disease-associated markers through an indirect process, define the effective chromosomal sample size as

$$n = \frac{4N_{cs}N_{ct}}{N_{cs} + N_{ct}}; \quad (27)$$

where N_{cs} and N_{ct} are the numbers of diploid cases and controls, respectively. This is necessary to handle situations where the numbers of cases and controls are not equivalent. For equal case and control sample sizes, $N_{cs} = N_{ct} = N$, the value of the effective number of chromosomes is simply $n = 2N$ – as expected. Let power be calculated for a significance level α (such that traditional P-values below α will be deemed statistically significant). Define the standard Gaussian distribution function as $\Phi(\bullet)$. Mathematically,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{\theta^2}{2}} d\theta \quad (28)$$

Alternatively, the following error function notation (Erf) may also be used,

$$\Phi(x) = \frac{1}{2} \left[1 + \text{Erf} \left(\frac{x}{\sqrt{2}} \right) \right] \quad (29)$$

For example, $\Phi(1.644854) = 0.95$. The value of r^2 may be derived to yield a pre-specified minimum amount of power to detect disease association though indirect interrogation. Noting that the LD SNP marker could be the one that is carrying the disease- association allele, therefore that this approach constitutes a lower-bound model where all indirect power results are expected to be at least as large as those interrogated.

Denote by β the error rate for not detecting truly disease-associated markers. Therefore, $1 - \beta$ is the classical definition of statistical power. Substituting the Pritchard-Pzeworski result into the sample size, the power to detect disease association at a significance level of α is given by the approximation

$$1 - \beta \cong \Phi \left[\frac{|q_{1,cs} - q_{1,ct}|}{\sqrt{\frac{q_{1,cs}(1 - q_{1,cs}) + q_{1,ct}(1 - q_{1,ct})}{r^2 n}}} - Z_{1-\alpha/2} \right]; \quad (30)$$

where Z_u is the inverse of the standard normal cumulative distribution evaluated at u ($u \in (0,1)$).

$Z_u = \Phi^{-1}(u)$, where $\Phi(\Phi^{-1}(u)) = \Phi^{-1}(\Phi(u)) = u$. For example, setting $\alpha = 0.05$, and therefore

5 $1 - \alpha/2 = 0.975$, we obtain $Z_{0.975} = 1.95996$. Next, setting power equal to a threshold of a minimum power of T ,

$$T = \Phi \left[\frac{|q_{1,cs} - q_{1,ct}|}{\sqrt{\frac{q_{1,cs}(1 - q_{1,cs}) + q_{1,ct}(1 - q_{1,ct})}{r^2 n}}} - Z_{1-\alpha/2} \right] \quad (31)$$

10 and solving for r^2 , the following threshold r^2 is obtained:

$$r_T^2 = \frac{[q_{1,cs}(1 - q_{1,cs}) + q_{1,ct}(1 - q_{1,ct})]}{n(q_{1,cs} - q_{1,ct})^2} [\Phi^{-1}(T) + Z_{1-\alpha/2}]^2 \quad (32)$$

15 Or,

$$r_T^2 = \frac{(Z_T + Z_{1-\alpha/2})^2}{n} \left[\frac{q_{1,cs} - (q_{1,cs})^2 + q_{1,ct} - (q_{1,ct})^2}{(q_{1,cs} - q_{1,ct})^2} \right] \quad (33)$$

20 Suppose that r^2 is calculated between an interrogated SNP and a number of other SNPs with varying levels of LD with the interrogated SNP. The threshold value r_T^2 is the minimum value of linkage disequilibrium between the interrogated SNP and the potential LD SNPs such that the LD SNP still retains a power greater or equal to T for detecting disease-association. For example, suppose that SNP rs200 is genotyped in a case-control disease-association study and it is found to be associated with a disease phenotype. Further suppose that the minor allele

frequency in 1,000 case chromosomes was found to be 16% in contrast with a minor allele frequency of 10% in 1,000 control chromosomes. Given those measurements one could have predicted, prior to the experiment, that the power to detect disease association at a significance level of 0.05 was quite high – approximately 98% using a test of allelic association. Applying
5 equation (32) one can calculate a minimum value of r^2 to indirectly assess disease association assuming that the minor allele at SNP rs200 is truly disease-predisposing for a threshold level of power. If one sets the threshold level of power to be 80%, then $r_r^2 = 0.489$ given the same significance level and chromosome numbers as above. Hence, any SNP with a pairwise r^2 value with rs200 greater than 0.489 is expected to have greater than 80% power to detect the disease
10 association. Further, this is assuming the conservative model where the LD SNP is disease-associated only through linkage disequilibrium with the interrogated SNP rs200.

The contribution or association of particular SNPs and/or SNP haplotypes with disease phenotypes, such as AD, enables the SNPs of the present invention to be used to develop superior diagnostic tests capable of identifying individuals who express a detectable trait, such as AD, as
15 the result of a specific genotype, or individuals whose genotype places them at an increased or decreased risk of developing a detectable trait at a subsequent time as compared to individuals who do not have that genotype. As described herein, diagnostics may be based on a single SNP or a group of SNPs. Combined detection of a plurality of SNPs (for example, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 24, 25, 30, 32, 48, 50, 64, 96, 100, or any other number
20 in-between, or more, of the SNPs provided in Table 1 and/or Table 2) typically increases the probability of an accurate diagnosis. For example, the presence of a single SNP known to correlate with AD might indicate a probability of 20% that an individual has or is at risk of developing AD, whereas detection of five SNPs, each of which correlates with AD, might indicate a probability of 80% that an individual has or is at risk of developing AD. To further
25 increase the accuracy of diagnosis or predisposition screening, analysis of the SNPs of the present invention can be combined with that of other polymorphisms or other risk factors of AD, such as disease symptoms, pathological characteristics, family history, diet, environmental factors or lifestyle factors.

It will, of course, be understood by practitioners skilled in the treatment or diagnosis of
30 AD that the present invention generally does not intend to provide an absolute identification of individuals who are at risk (or less at risk) of developing AD, and/or pathologies related to AD, but rather to indicate a certain increased (or decreased) degree or likelihood of developing the disease based on statistically significant association results. However, this information is extremely valuable as it can be used to, for example, initiate preventive treatments or to allow an

individual carrying one or more significant SNPs or SNP haplotypes to foresee warning signs such as minor clinical symptoms, or to have regularly scheduled physical exams to monitor for appearance of a condition in order to identify and begin treatment of the condition at an early stage. Particularly with diseases that are extremely debilitating or fatal if not treated on time, the knowledge of a potential predisposition, even if this predisposition is not absolute, would likely contribute in a very significant manner to treatment efficacy.

The diagnostic techniques of the present invention may employ a variety of methodologies to determine whether a test subject has a SNP or a SNP pattern associated with an increased or decreased risk of developing a detectable trait or whether the individual suffers from a detectable trait as a result of a particular polymorphism/mutation, including, for example, methods which enable the analysis of individual chromosomes for haplotyping, family studies, single sperm DNA analysis, or somatic hybrids. The trait analyzed using the diagnostics of the invention may be any detectable trait that is commonly observed in pathologies and disorders related to AD.

Another aspect of the present invention relates to a method of determining whether an individual is at risk (or less at risk) of developing one or more traits or whether an individual expresses one or more traits as a consequence of possessing a particular trait-causing or trait-influencing allele. These methods generally involve obtaining a nucleic acid sample from an individual and assaying the nucleic acid sample to determine which nucleotide(s) is/are present at one or more SNP positions, wherein the assayed nucleotide(s) is/are indicative of an increased or decreased risk of developing the trait or indicative that the individual expresses the trait as a result of possessing a particular trait-causing or trait-influencing allele.

In another embodiment, the SNP detection reagents of the present invention are used to determine whether an individual has one or more SNP allele(s) affecting the level (e.g., the concentration of mRNA or protein in a sample, etc.) or pattern (e.g., the kinetics of expression, rate of decomposition, stability profile, K_m , V_{max} , etc.) of gene expression (collectively, the "gene response" of a cell or bodily fluid). Such a determination can be accomplished by screening for mRNA or protein expression (e.g., by using nucleic acid arrays, RT-PCR, TaqMan assays, or mass spectrometry), identifying genes having altered expression in an individual, genotyping SNPs disclosed in Table 1 and/or Table 2 that could affect the expression of the genes having altered expression (e.g., SNPs that are in and/or around the gene(s) having altered expression, SNPs in regulatory/control regions, SNPs in and/or around other genes that are involved in pathways that could affect the expression of the gene(s) having altered expression, or all SNPs could be genotyped), and correlating SNP genotypes with altered gene expression. In

this manner, specific SNP alleles at particular SNP sites can be identified that affect gene expression.

Pharmacogenomics and Therapeutics/Drug Development

5 The present invention provides methods for assessing the pharmacogenomics of a subject harboring particular SNP alleles or haplotypes to a particular therapeutic agent or pharmaceutical compound, or to a class of such compounds. Pharmacogenomics deals with the roles which clinically significant hereditary variations (e.g., SNPs) play in the response to drugs due to altered drug disposition and/or abnormal action in affected persons. See, e.g., Roses, *Nature* 405, 857-865
10 (2000); Gould Rothberg, *Nature Biotechnology* 19, 209-211 (2001); Eichelbaum, *Clin. Exp. Pharmacol. Physiol.* 23(10-11):983-985 (1996); and Linder, *Clin. Chem.* 43(2):254-266 (1997). The clinical outcomes of these variations can result in severe toxicity of therapeutic drugs in certain individuals or therapeutic failure of drugs in certain individuals as a result of individual variation in metabolism. Thus, the SNP genotype of an individual can determine the way a therapeutic
15 compound acts on the body or the way the body metabolizes the compound. For example, SNPs in drug metabolizing enzymes can affect the activity of these enzymes, which in turn can affect both the intensity and duration of drug action, as well as drug metabolism and clearance.

 The discovery of SNPs in drug metabolizing enzymes, drug transporters, proteins for pharmaceutical agents, and other drug targets has explained why some patients do not obtain the
20 expected drug effects, show an exaggerated drug effect, or experience serious toxicity from standard drug dosages. SNPs can be expressed in the phenotype of the extensive metabolizer and in the phenotype of the poor metabolizer. Accordingly, SNPs may lead to allelic variants of a protein in which one or more of the protein functions in one population are different from those in another population. SNPs and the encoded variant peptides thus provide targets to ascertain a genetic
25 predisposition that can affect treatment modality. For example, in a ligand-based treatment, SNPs may give rise to amino terminal extracellular domains and/or other ligand-binding regions of a receptor that are more or less active in ligand binding, thereby affecting subsequent protein activation. Accordingly, ligand dosage would necessarily be modified to maximize the therapeutic effect within a given population containing particular SNP alleles or haplotypes.

30 As an alternative to genotyping, specific variant proteins containing variant amino acid sequences encoded by alternative SNP alleles could be identified. Thus, pharmacogenomic characterization of an individual permits the selection of effective compounds and effective dosages of such compounds for prophylactic or therapeutic uses based on the individual's SNP genotype, thereby enhancing and optimizing the effectiveness of the therapy. Furthermore, the production

of recombinant cells and transgenic animals containing particular SNPs/haplotypes allow effective clinical design and testing of treatment compounds and dosage regimens. For example, transgenic animals can be produced that differ only in specific SNP alleles in a gene that is orthologous to a human disease susceptibility gene.

5 Pharmacogenomic uses of the SNPs of the present invention provide several significant advantages for patient care, particularly in treating AD. Pharmacogenomic characterization of an individual, based on an individual's SNP genotype, can identify those individuals unlikely to respond to treatment with a particular medication and thereby allows physicians to avoid prescribing the ineffective medication to those individuals. On the other hand, SNP genotyping of an individual
10 may enable physicians to select the appropriate medication and dosage regimen that will be most effective based on an individual's SNP genotype. This information increases a physician's confidence in prescribing medications and motivates patients to comply with their drug regimens. Furthermore, pharmacogenomics may identify patients predisposed to toxicity and adverse reactions to particular drugs or drug dosages. Adverse drug reactions lead to more than 100,000 avoidable
15 deaths per year in the United States alone and therefore represent a significant cause of hospitalization and death, as well as a significant economic burden on the healthcare system (Pfof et al., *Trends in Biotechnology* [Aug. 2000]). Thus, pharmacogenomics based on the SNPs disclosed herein has the potential to both save lives and reduce healthcare costs substantially.

 Pharmacogenomics in general is discussed further in Rose et al., "Pharmacogenetic
20 analysis of clinically relevant genetic polymorphisms," *Methods Mol. Med.* 85:225-37 (2003). Pharmacogenomics as it relates to AD and other neurodegenerative disorders is discussed in Cacabelos, "Pharmacogenomics for the treatment of dementia," *Ann. Med.* 34(5):357-79 (2002); Maimone et al., "Pharmacogenomics of neurodegenerative diseases," *Eur. J. Pharmacol.* 9, 413(1):11-29 (Feb 2001), and Poirier, "Apolipoprotein E: a pharmacogenetic target for the
25 treatment of AD," *Mol. Diagn.* 4(4):335-41 (Dec. 1999). Pharmacogenomics as it relates to cardiovascular disorders is discussed in Siest et al., "Pharmacogenomics of drugs affecting the cardiovascular system," *Clin. Chem. Lab Med.* 41(4):590-9 (Apr. 2003); Mukherjee et al., "Pharmacogenomics in cardiovascular diseases," *Prog. Cardiovasc. Dis.* 44(6):479-98 (May-Jun. 2002); and Mooser et al., "Cardiovascular pharmacogenetics in the SNP era," *J. Thromb.*
30 *Haemost.* 1(7):1398-402 (Jul. 2003). Pharmacogenomics as it relates to cancer is discussed in McLeod et al., "Cancer pharmacogenomics: SNPs, chips, and the individual patient," *Cancer Invest.* 21(4):630-40 (2003) and Watters et al., "Cancer pharmacogenomics: current and future applications," *Biochim. Biophys. Acta.* 17;1603(2):99-111 (Mar. 2003).

The SNPs of the present invention also can be used to identify novel therapeutic targets for AD. For example, genes containing the disease-associated variants ("variant genes") or their products, as well as genes or their products that are directly or indirectly regulated by or interacting with these variant genes or their products, can be targeted for the development of therapeutics that, for example, treat the disease or prevent or delay disease onset. The therapeutics may be composed of, for example, small molecules, proteins, protein fragments or peptides, antibodies, nucleic acids, or their derivatives or mimetics which modulate the functions or levels of the target genes or gene products.

The SNP-containing nucleic acid molecules disclosed herein, and their complementary nucleic acid molecules, may be used as antisense constructs to control gene expression in cells, tissues, and organisms. Antisense technology is well established in the art and extensively reviewed in *Antisense Drug Technology: Principles, Strategies, and Applications*, ed. Crooke, Marcel Dekker, Inc., New York (2001). An antisense nucleic acid molecule is generally designed to be complementary to a region of mRNA expressed by a gene so that the antisense molecule hybridizes to the mRNA and thereby blocks translation of mRNA into protein. Various classes of antisense oligonucleotides are used in the art, two of which are cleavers and blockers. Cleavers, by binding to target RNAs, activate intracellular nucleases (e.g., RNaseH or RNase L) that cleave the target RNA. Blockers, which also bind to target RNAs, inhibit protein translation through steric hindrance of ribosomes. Exemplary blockers include peptide nucleic acids, morpholinos, locked nucleic acids, and methylphosphonates (see, e.g., Thompson, *Drug Discovery Today*, 7 [17]: 912-917 [2002]). Antisense oligonucleotides are directly useful as therapeutic agents, and are also useful for determining and validating gene function (e.g., in gene knock-out or knock-down experiments).

Antisense technology is further reviewed in: Lavery *et al.*, "Antisense and RNAi: powerful tools in drug target discovery and validation," *Curr. Opin. Drug Discov. Devel.* 6(4):561-9 (Jul. 2003); Stephens *et al.*, "Antisense oligonucleotide therapy in cancer," *Curr. Opin. Mol. Ther.* 5(2):118-22 (Apr. 2003); Kurreck, "Antisense technologies. Improvement through novel chemical modifications," *Eur. J. Biochem.* 270(8):1628-44 (Apr. 2003); Dias *et al.*, "Antisense oligonucleotides: basic concepts and mechanisms," *Mol. Cancer Ther.* 1(5):347-55 (Mar. 2002); Chen, "Clinical development of antisense oligonucleotides as anti-cancer therapeutics," *Methods Mol. Med.* 75:621-36 (2003); Wang *et al.*, "Antisense anticancer oligonucleotide therapeutics," *Curr. Cancer Drug Targets* 1(3):177-96 (Nov. 2001); and Bennett, "Efficiency of antisense oligonucleotide drug discovery," *Antisense Nucleic Acid Drug. Dev.* 12(3):215-24 (Jun. 2002).

The SNPs of the present invention are particularly useful for designing antisense reagents that are specific for particular nucleic acid variants. Based on the SNP information disclosed herein, antisense oligonucleotides can be produced that specifically target mRNA molecules that contain one or more particular SNP nucleotides. In this manner, expression of mRNA molecules that contain one or more undesired polymorphisms (e.g., SNP nucleotides that lead to a defective protein such as an amino acid substitution in a catalytic domain) can be inhibited or completely blocked. Thus, antisense oligonucleotides can be used to specifically bind a particular polymorphic form (e.g., a SNP allele that encodes a defective protein), thereby inhibiting translation of this form, but which do not bind an alternative polymorphic form (e.g., an alternative SNP nucleotide that encodes a protein having normal function).

Antisense molecules can be used to inactivate mRNA in order to inhibit gene expression and production of defective proteins. Accordingly, these molecules can be used to treat a disorder, such as AD, characterized by abnormal or undesired gene expression or expression of certain defective proteins. This technique can involve cleavage by means of ribozymes containing nucleotide sequences complementary to one or more regions in the mRNA that attenuate the ability of the mRNA to be translated. Possible mRNA regions include, for example, protein-coding regions and particularly protein-coding regions corresponding to catalytic activities, substrate/ligand binding, or other functional activities of a protein.

The SNPs of the present invention are also useful for designing RNA interference reagents that specifically target nucleic acid molecules having particular SNP variants. RNA interference (RNAi), also referred to as gene silencing, is based on using double-stranded RNA (dsRNA) molecules to turn genes off. When introduced into a cell, dsRNAs are processed by the cell into short fragments (generally about 21, 22, or 23 nucleotides in length) known as small interfering RNAs (siRNAs) which the cell uses in a sequence-specific manner to recognize and destroy complementary RNAs (Thompson, *Drug Discovery Today*, 7 [17]: 912-917 [2002]). Accordingly, an aspect of the present invention specifically contemplates isolated nucleic acid molecules that are about 18-26 nucleotides in length, preferably 19-25 nucleotides in length, and more preferably 20, 21, 22, or 23 nucleotides in length, and the use of these nucleic acid molecules for RNAi. Because RNAi molecules, including siRNAs, act in a sequence-specific manner, the SNPs of the present invention can be used to design RNAi reagents that recognize and destroy nucleic acid molecules having specific SNP alleles/nucleotides (such as deleterious alleles that lead to the production of defective proteins), while not affecting nucleic acid molecules having alternative SNP alleles (such as alleles that encode proteins having normal function). As with antisense reagents, RNAi reagents may be directly useful as therapeutic

agents (e.g., for turning off defective, disease-causing genes), and are also useful for characterizing and validating gene function (e.g., in gene knock-out or knock-down experiments).

The following references provide a further review of RNAi: Reynolds *et al.*, "Rational siRNA design for RNA interference," *Nat. Biotechnol.* 22(3):326-30 (Mar. 2004); Epub Feb. 1 2004; Chi *et al.*, "Genomewide view of gene silencing by small interfering RNAs," *PNAS* 100(11):6343-6346 (2003); Vickers *et al.*, "Efficient Reduction of Target RNAs by Small Interfering RNA and RNase H-dependent Antisense Agents," *J. Biol. Chem.* 278: 7108-7118 (2003); Agami, "RNAi and related mechanisms and their potential use for therapy," *Curr. Opin. Chem. Biol.* 6(6):829-34 (Dec. 2002); Lavery *et al.*, "Antisense and RNAi: powerful tools in drug target discovery and validation," *Curr. Opin. Drug Discov. Devel.* 6(4):561-9 (Jul. 2003); Shi, "Mammalian RNAi for the masses," *Trends Genet.* 19(1):9-12 (Jan. 2003); Shuey *et al.*, "RNAi: gene-silencing in therapeutic intervention," *Drug Discovery Today* 7(20):1040-1046 (Oct. 2002); McManus *et al.*, *Nat. Rev. Genet.* 3(10):737-47 (Oct. 2002); Xia *et al.*, *Nat. Biotechnol.* 20(10):1006-10 (Oct. 2002); Plasterk *et al.*, *Curr. Opin. Genet. Dev.* 10(5):562-7 (Oct. 2000); Boshier *et al.*, *Nat. Cell Biol.* 2(2):E31-6 (Feb. 2000); and Hunter, *Curr. Biol.* 9(12):R440-2 (Jun. 17 1999).

A subject suffering from a pathological condition, such as AD, ascribed to a SNP may be treated so as to correct the genetic defect (see Kren *et al.*, *Proc. Natl. Acad. Sci. USA* 96:10349-10354 [1999]). Such a subject can be identified by any method that can detect the polymorphism in a biological sample drawn from the subject. Such a genetic defect may be permanently corrected by administering to such a subject a nucleic acid fragment incorporating a repair sequence that supplies the normal/wild-type nucleotide at the position of the SNP. This site-specific repair sequence can encompass an RNA/DNA oligonucleotide that operates to promote endogenous repair of a subject's genomic DNA. The site-specific repair sequence is administered in an appropriate vehicle, such as a complex with polyethylenimine, encapsulated in anionic liposomes, a viral vector such as an adenovirus, or other pharmaceutical composition that promotes intracellular uptake of the administered nucleic acid. A genetic defect leading to an inborn pathology may then be overcome, as the chimeric oligonucleotides induce incorporation of the normal sequence into the subject's genome. Upon incorporation, the normal gene product is expressed, and the replacement is propagated, thereby engendering a permanent repair and therapeutic enhancement of the clinical condition of the subject.

In cases in which a cSNP results in a variant protein that is ascribed to be the cause of, or a contributing factor to, a pathological condition, a method of treating such a condition can

include administering to a subject experiencing the pathology the wild-type/normal cognate of the variant protein. Once administered in an effective dosing regimen, the wild-type cognate provides complementation or remediation of the pathological condition.

5 The invention further provides a method for identifying a compound or agent that can be used to treat AD. The SNPs disclosed herein are useful as targets for the identification and/or development of therapeutic agents. A method for identifying a therapeutic agent or compound typically includes assaying the ability of the agent or compound to modulate the activity and/or expression of a SNP-containing nucleic acid or the encoded product and thus identifying an agent or
10 a compound that can be used to treat a disorder characterized by undesired activity or expression of the SNP-containing nucleic acid or the encoded product. The assays can be performed in cell-based and cell-free systems. Cell-based assays can include cells naturally expressing the nucleic acid molecules of interest or recombinant cells genetically engineered to express certain nucleic acid molecules.

Variant gene expression in a AD patient can include, for example, either expression of a
15 SNP-containing nucleic acid sequence (for instance, a gene that contains a SNP can be transcribed into an mRNA transcript molecule containing the SNP, which can in turn be translated into a variant protein) or altered expression of a normal/wild-type nucleic acid sequence due to one or more SNPs (for instance, a regulatory/control region can contain a SNP that affects the level or pattern of expression of a normal transcript).

20 Assays for variant gene expression can involve direct assays of nucleic acid levels (e.g., mRNA levels), expressed protein levels, or of collateral compounds involved in a signal pathway. Further, the expression of genes that are up- or down-regulated in response to the signal pathway can also be assayed. In this embodiment, the regulatory regions of these genes can be operably linked to a reporter gene such as luciferase.

25 Modulators of variant gene expression can be identified in a method wherein, for example, a cell is contacted with a candidate compound/agent and the expression of mRNA determined. The level of expression of mRNA in the presence of the candidate compound is compared to the level of expression of mRNA in the absence of the candidate compound. The candidate compound can then be identified as a modulator of variant gene expression based on this comparison and be used to treat
30 a disorder such as AD that is characterized by variant gene expression (e.g., either expression of a SNP-containing nucleic acid or altered expression of a normal/wild-type nucleic acid molecule due to one or more SNPs that affect expression of the nucleic acid molecule) due to one or more SNPs of the present invention. When expression of mRNA is statistically significantly greater in the presence of the candidate compound than in its absence, the candidate compound is identified as a

stimulator of nucleic acid expression. When nucleic acid expression is statistically significantly less in the presence of the candidate compound than in its absence, the candidate compound is identified as an inhibitor of nucleic acid expression.

5 The invention further provides methods of treatment, with the SNP or associated nucleic acid domain (e.g., catalytic domain, ligand/substrate-binding domain, regulatory/control region, etc.) or gene, or the encoded mRNA transcript, as a target, using a compound identified through drug screening as a gene modulator to modulate variant nucleic acid expression. Modulation can include either up-regulation (i.e., activation or agonization) or down-regulation (i.e., suppression or antagonization) of nucleic acid expression.

10 Expression of mRNA transcripts and encoded proteins, either wild type or variant, may be altered in individuals with a particular SNP allele in a regulatory/control element, such as a promoter or transcription factor binding domain, that regulates expression. In this situation, methods of treatment and compounds can be identified, as discussed herein, that regulate or overcome the variant regulatory/control element, thereby generating normal, or healthy, expression levels of either
15 the wild type or variant protein.

The SNP-containing nucleic acid molecules of the present invention are also useful for monitoring the effectiveness of modulating compounds on the expression or activity of a variant gene, or encoded product, in clinical trials or in a treatment regimen. Thus, the gene expression pattern can serve as an indicator for the continuing effectiveness of treatment with the compound,
20 particularly with compounds to which a patient can develop resistance, as well as an indicator for toxicities. The gene expression pattern can also serve as a marker indicative of a physiological response of the affected cells to the compound. Accordingly, such monitoring would allow either increased administration of the compound or the administration of alternative compounds to which the patient has not become resistant. Similarly, if the level of nucleic acid expression falls below a
25 desirable level, administration of the compound could be commensurately decreased.

In another aspect of the present invention, there is provided a pharmaceutical pack comprising a therapeutic agent (e.g., a small molecule drug, antibody, peptide, antisense or RNAi nucleic acid molecule, etc.) and a set of instructions for administration of the therapeutic agent to humans diagnostically tested for one or more SNPs or SNP haplotypes provided by the present
30 invention.

The SNPs/haplotypes of the present invention are also useful for improving many different aspects of the drug development process. For instance, an aspect of the present invention includes selecting individuals for clinical trials based on their SNP genotype. For example, individuals with SNP genotypes that indicate that they are likely to positively respond

to a drug can be included in the trials, whereas those individuals whose SNP genotypes indicate that they are less likely to or would not respond to the drug, or who are at risk for suffering toxic effects or other adverse reactions, can be excluded from the clinical trials. This not only can improve the safety of clinical trials, but also can enhance the chances that the trial will demonstrate statistically significant efficacy. Furthermore, the SNPs of the present invention may explain why certain previously developed drugs performed poorly in clinical trials and may help identify a subset of the population that would benefit from a drug that had previously performed poorly in clinical trials, thereby “rescuing” previously developed drugs, and enabling the drug to be made available to a particular AD patient population that can benefit from it.

SNPs have many important uses in drug discovery, screening, and development. A high probability exists that, for any gene/protein selected as a potential drug target, variants of that gene/protein will exist in a patient population. Thus, determining the impact of gene/protein variants on the selection and delivery of a therapeutic agent should be an integral aspect of the drug discovery and development process. (Jazwinska, *A Trends Guide to Genetic Variation and Genomic Medicine* S30-S36 [Mar. 2002]).

Knowledge of variants (e.g., SNPs and any corresponding amino acid polymorphisms) of a particular therapeutic target (e.g., a gene, mRNA transcript, or protein) enables parallel screening of the variants in order to identify therapeutic candidates (e.g., small molecule compounds, antibodies, antisense or RNAi nucleic acid compounds, etc.) that demonstrate efficacy across variants (Rothberg, *Nat. Biotechnol.* 19[3]:209-11 [Mar. 2001]). Such therapeutic candidates would be expected to show equal efficacy across a larger segment of the patient population, thereby leading to a larger potential market for the therapeutic candidate.

Furthermore, identifying variants of a potential therapeutic target enables the most common form of the target to be used for selection of therapeutic candidates, thereby helping to ensure that the experimental activity that is observed for the selected candidates reflects the real activity expected in the largest proportion of a patient population (Jazwinska, *A Trends Guide to Genetic Variation and Genomic Medicine* S30-S36 [Mar. 2002]).

Additionally, screening therapeutic candidates against all known variants of a target can enable the early identification of potential toxicities and adverse reactions relating to particular variants. For example, variability in drug absorption, distribution, metabolism and excretion (ADME) caused by, for example, SNPs in therapeutic targets or drug metabolizing genes, can be identified, and this information can be utilized during the drug development process to minimize variability in drug disposition and develop therapeutic agents that are safer across a wider range of a patient population. The SNPs of the present invention, including the variant proteins and

encoding polymorphic nucleic acid molecules provided in Tables 1-2, are useful in conjunction with a variety of toxicology methods established in the art, such as those set forth in *Current Protocols in Toxicology*, John Wiley & Sons, Inc., New York.

Furthermore, therapeutic agents that target any art-known proteins (or nucleic acid molecules, either RNA or DNA) may cross-react with the variant proteins (or polymorphic nucleic acid molecules) disclosed in Table 1, thereby significantly affecting the pharmacokinetic properties of the drug. Consequently, the protein variants and the SNP-containing nucleic acid molecules disclosed in Tables 1-2 are useful in developing, screening, and evaluating therapeutic agents that target corresponding art-known protein forms (or nucleic acid molecules).

Additionally, as discussed above, knowledge of all polymorphic forms of a particular drug target enables the design of therapeutic agents that are effective against most or all such polymorphic forms of the drug target.

Pharmaceutical Compositions and Administration Thereof

Any of the AD-associated proteins, and encoding nucleic acid molecules, disclosed herein can be used as therapeutic targets (or directly used themselves as therapeutic compounds) for treating AD and related pathologies, and the present disclosure enables therapeutic compounds (e.g., small molecules, antibodies, therapeutic proteins, RNAi and antisense molecules, etc.) to be developed that target (or are comprised of) any of these therapeutic targets.

In general, a therapeutic compound will be administered in a therapeutically effective amount by any of the accepted modes of administration for agents that serve similar utilities. The actual amount of the therapeutic compound of this invention, i.e., the active ingredient, will depend upon numerous factors such as the severity of the disease to be treated, the age and relative health of the subject, the potency of the compound used, the route and form of administration, and other factors.

Therapeutically effective amounts of therapeutic compounds may range from, for example, approximately 0.01-50 mg per kilogram body weight of the recipient per day; preferably about 0.1-20 mg/kg/day. Thus, as an example, for administration to a 70 kg person, the dosage range would most preferably be about 7 mg to 1.4 g per day.

In general, therapeutic compounds will be administered as pharmaceutical compositions by any one of the following routes: oral, systemic (e.g., transdermal, intranasal, or by suppository), or parenteral (e.g., intramuscular, intravenous, or subcutaneous) administration. The preferred manner of administration is oral or parenteral using a convenient daily dosage regimen, which can be adjusted according to the degree of affliction. Oral compositions can take

the form of tablets, pills, capsules, semisolids, powders, sustained release formulations, solutions, suspensions, elixirs, aerosols, or any other appropriate compositions.

The choice of formulation depends on various factors such as the mode of drug administration (e.g., for oral administration, formulations in the form of tablets, pills, or capsules are preferred) and the bioavailability of the drug substance. Recently, pharmaceutical formulations have been developed especially for drugs that show poor bioavailability based upon the principle that bioavailability can be increased by increasing the surface area, i.e., decreasing particle size. For example, U.S. Patent No. 4,107,288 describes a pharmaceutical formulation having particles in the size range from 10 to 1,000 nm in which the active material is supported on a cross-linked matrix of macromolecules. U.S. Patent No. 5,145,684 describes the production of a pharmaceutical formulation in which the drug substance is pulverized to nanoparticles (average particle size of 400 nm) in the presence of a surface modifier and then dispersed in a liquid medium to give a pharmaceutical formulation that exhibits remarkably high bioavailability.

Pharmaceutical compositions are comprised of, in general, a therapeutic compound in combination with at least one pharmaceutically acceptable excipient. Acceptable excipients are non-toxic, aid administration, and do not adversely affect the therapeutic benefit of the therapeutic compound. Such excipients may be any solid, liquid, semi-solid or, in the case of an aerosol composition, gaseous excipient that is generally available to one skilled in the art.

Solid pharmaceutical excipients include starch, cellulose, talc, glucose, lactose, sucrose, gelatin, malt, rice, flour, chalk, silica gel, magnesium stearate, sodium stearate, glycerol monostearate, sodium chloride, dried skim milk and the like. Liquid and semisolid excipients may be selected from glycerol, propylene glycol, water, ethanol and various oils, including those of petroleum, animal, vegetable or synthetic origin, e.g., peanut oil, soybean oil, mineral oil, sesame oil, etc. Preferred liquid carriers, particularly for injectable solutions, include water, saline, aqueous dextrose, and glycols.

Compressed gases may be used to disperse a compound of this invention in aerosol form. Inert gases suitable for this purpose are nitrogen, carbon dioxide, etc.

Other suitable pharmaceutical excipients and their formulations are described in *Remington's Pharmaceutical Sciences* 18th edition, ed. E. W. Martin (Mack Publishing Company [1990]).

The amount of the therapeutic compound in a formulation can vary within the full range employed by those skilled in the art. Typically, the formulation will contain, on a weight percent (wt %) basis, from about 0.01-99.99 wt % of the therapeutic compound based on the total

formulation, with the balance being one or more suitable pharmaceutical excipients. Preferably, the compound is present at a level of about 1-80 wt %.

Therapeutic compounds can be administered alone or in combination with other therapeutic compounds or in combination with one or more other active ingredient(s). For example, an inhibitor or stimulator of a AD-associated protein can be administered in combination with another agent that inhibits or stimulates the activity of the same or a different AD-associated protein to thereby counteract the affects of AD.

For further information regarding pharmacology, see *Current Protocols in Pharmacology*, John Wiley & Sons, Inc., New York.

Human Identification Applications

In addition to their diagnostic and therapeutic uses in AD and related pathologies, the SNPs provided by the present invention are also useful as human identification markers for such applications as forensics, paternity testing, and biometrics (see, e.g., Gill, "An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes," *Int. J. Legal Med.* 114[4-5]:204-10 [2001]). Genetic variations in the nucleic acid sequences between individuals can be used as genetic markers to identify individuals and to associate a biological sample with an individual. Determination of which nucleotides occupy a set of SNP positions in an individual identifies a set of SNP markers that distinguishes the individual. The more SNP positions that are analyzed, the lower the probability that the set of SNPs in one individual is the same as that in an unrelated individual. Preferably, if multiple sites are analyzed, the sites are unlinked (i.e., inherited independently). Thus, preferred sets of SNPs can be selected from among the SNPs disclosed herein, which may include SNPs on different chromosomes, SNPs on different chromosome arms, and/or SNPs that are dispersed over substantial distances along the same chromosome arm.

Furthermore, among the SNPs disclosed herein, preferred SNPs for use in certain forensic/human identification applications include SNPs located at degenerate codon positions (i.e., the third position in certain codons which can be one of two or more alternative nucleotides and still encode the same amino acid), since these SNPs do not affect the encoded protein. SNPs that do not affect the encoded protein are expected to be under less selective pressure and are therefore expected to be more polymorphic in a population, which is typically an advantage for forensic/human identification applications. However, for certain forensics/human identification applications, such as predicting phenotypic characteristics (e.g., inferring ancestry or inferring

one or more physical characteristics of an individual) from a DNA sample, it may be desirable to utilize SNPs that affect the encoded protein.

For many of the SNPs disclosed in Tables 1-2 (which are identified as “Applera” SNP source), Tables 1-2 provide SNP allele frequencies obtained by re-sequencing the DNA of chromosomes from 39 individuals (Tables 1-2 also provide allele frequency information for “Celera” source SNPs and, where available, public SNPs from dbEST, HGBASE, and/or HGMD). The allele frequencies provided in Tables 1-2 enable these SNPs to be readily used for human identification applications. Although any SNP disclosed in Table 1 and/or Table 2 could be used for human identification, the closer that the frequency of the minor allele at a particular SNP site is to 50%, the greater the ability of that SNP to discriminate between different individuals in a population since it becomes increasingly likely that two randomly selected individuals would have different alleles at that SNP site. Using the SNP allele frequencies provided in Tables 1-2, one of ordinary skill in the art could readily select a subset of SNPs for which the frequency of the minor allele is, for example, at least 1%, 2%, 5%, 10%, 20%, 25%, 30%, 40%, 45%, or 50%, or any other frequency in-between. Thus, since Tables 1-2 provide allele frequencies based on the re-sequencing of the chromosomes from 39 individuals, a subset of SNPs could readily be selected for human identification in which the total allele count of the minor allele at a particular SNP site is, for example, at least 1, 2, 4, 8, 10, 16, 20, 24, 30, 32, 36, 38, 39, 40, or any other number in-between.

Furthermore, Tables 1-2 also provide population group (interchangeably referred to herein as ethnic or racial groups) information coupled with the extensive allele frequency information. For example, the group of 39 individuals whose DNA was re-sequenced was made-up of 20 Caucasians and 19 African-Americans. This population group information enables further refinement of SNP selection for human identification. For example, preferred SNPs for human identification can be selected from Tables 1-2 that have similar allele frequencies in both the Caucasian and African-American populations; thus, for example, SNPs can be selected that have equally high discriminatory power in both populations. Alternatively, SNPs can be selected for which there is a statistically significant difference in allele frequencies between the Caucasian and African-American populations (as an extreme example, a particular allele may be observed only in either the Caucasian or the African-American population group but not observed in the other population group); such SNPs are useful, for example, for predicting the race/ethnicity of an unknown perpetrator from a biological sample such as a hair or blood stain recovered at a crime scene. For a discussion of using SNPs to predict ancestry from a DNA sample, including

statistical methods, see Frudakis *et al.*, "A Classifier for the SNP-Based Inference of Ancestry," *Journal of Forensic Sciences* 48[4]:771-782 [2003]).

SNPs have numerous advantages over other types of polymorphic markers, such as short tandem repeats (STRs). For example, SNPs can be easily scored and are amenable to automation, making SNPs the markers of choice for large-scale forensic databases. SNPs are found in much greater abundance throughout the genome than repeat polymorphisms. Population frequencies of two polymorphic forms can usually be determined with greater accuracy than those of multiple polymorphic forms at multi-allelic loci. SNPs are mutationally more stable than repeat polymorphisms. SNPs are not susceptible to artefacts such as stutter bands that can hinder analysis. Stutter bands are frequently encountered when analyzing repeat polymorphisms, and are particularly troublesome when analyzing samples such as crime scene samples that may contain mixtures of DNA from multiple sources. Another significant advantage of SNP markers over STR markers is the much shorter length of nucleic acid needed to score a SNP. For example, STR markers are generally several hundred base pairs in length. A SNP, on the other hand, comprises a single nucleotide, and generally a short conserved region on either side of the SNP position for primer and/or probe binding. This makes SNPs more amenable to typing in highly degraded or aged biological samples that are frequently encountered in forensic casework in which DNA may be fragmented into short pieces.

SNPs also are not subject to microvariant and "off-ladder" alleles frequently encountered when analyzing STR loci. Microvariants are deletions or insertions within a repeat unit that change the size of the amplified DNA product so that the amplified product does not migrate at the same rate as reference alleles with normal sized repeat units. When separated by size, such as by electrophoresis on a polyacrylamide gel, microvariants do not align with a reference allelic ladder of standard sized repeat units, but rather migrate between the reference alleles. The reference allelic ladder is used for precise sizing of alleles for allele classification; therefore alleles that do not align with the reference allelic ladder lead to substantial analysis problems. Furthermore, when analyzing multi-allelic repeat polymorphisms, occasionally an allele is found that consists of more or less repeat units than has been previously seen in the population, or more or less repeat alleles than are included in a reference allelic ladder. These alleles will migrate outside the size range of known alleles in a reference allelic ladder, and therefore are referred to as "off-ladder" alleles. In extreme cases, the allele may contain so few or so many repeats that it migrates well out of the range of the reference allelic ladder. In this situation, the allele may not even be observed, or, with multiplex analysis, it may migrate within or close to the size range for another locus, further confounding analysis.

SNP analysis avoids the problems of microvariants and off-ladder alleles encountered in STR analysis. Importantly, microvariants and off-ladder alleles may provide significant problems, and may be completely missed, when using analysis methods such as oligonucleotide hybridization arrays, which utilize oligonucleotide probes specific for certain known alleles.

5 Furthermore, off-ladder alleles and microvariants encountered with STR analysis, even when correctly typed, may lead to improper statistical analysis, since their frequencies in the population are generally unknown or poorly characterized, and therefore the statistical significance of a matching genotype may be questionable. All these advantages of SNP analysis are considerable in light of the consequences of most DNA identification cases, which may lead
10 to life imprisonment for an individual, or re-association of remains to the family of a deceased individual.

DNA can be isolated from biological samples such as blood, bone, hair, saliva, or semen, and compared with the DNA from a reference source at particular SNP positions. Multiple SNP markers can be assayed simultaneously in order to increase the power of discrimination and the
15 statistical significance of a matching genotype. For example, oligonucleotide arrays can be used to genotype a large number of SNPs simultaneously. The SNPs provided by the present invention can be assayed in combination with other polymorphic genetic markers, such as other SNPs known in the art or STRs, in order to identify an individual or to associate an individual with a particular biological sample.

20 Furthermore, the SNPs provided by the present invention can be genotyped for inclusion in a database of DNA genotypes, for example, a criminal DNA databank such as the FBI's Combined DNA Index System (CODIS) database. A genotype obtained from a biological sample of unknown source can then be queried against the database to find a matching genotype, with the SNPs of the present invention providing nucleotide positions at which to compare the
25 known and unknown DNA sequences for identity. Accordingly, the present invention provides a database comprising novel SNPs or SNP alleles of the present invention (e.g., the database can comprise information indicating which alleles are possessed by individual members of a population at one or more novel SNP sites of the present invention), such as for use in forensics, biometrics, or other human identification applications. Such a database typically comprises a
30 computer-based system in which the SNPs or SNP alleles of the present invention are recorded on a computer readable medium (see the section of the present specification entitled "Computer-Related Embodiments").

The SNPs of the present invention can also be assayed for use in paternity testing. The object of paternity testing is usually to determine whether a male is the father of a child. In most

cases, the mother of the child is known and thus, the mother's contribution to the child's genotype can be traced. Paternity testing investigates whether the part of the child's genotype not attributable to the mother is consistent with that of the putative father. Paternity testing can be performed by analyzing sets of polymorphisms in the putative father and the child, with the SNPs of the present invention providing nucleotide positions at which to compare the putative father's and child's DNA sequences for identity. If the set of polymorphisms in the child attributable to the father does not match the set of polymorphisms of the putative father, it can be concluded, barring experimental error, that the putative father is not the father of the child. If the set of polymorphisms in the child attributable to the father match the set of polymorphisms of the putative father, a statistical calculation can be performed to determine the probability of coincidental match, and a conclusion drawn as to the likelihood that the putative father is the true biological father of the child.

In addition to paternity testing, SNPs are also useful for other types of kinship testing, such as for verifying familial relationships for immigration purposes, or for cases in which an individual alleges to be related to a deceased individual in order to claim an inheritance from the deceased individual, etc. For further information regarding the utility of SNPs for paternity testing and other types of kinship testing, including methods for statistical analysis, see Krawczak, "Informativity assessment for biallelic single nucleotide polymorphisms," *Electrophoresis* 20(8):1676-81 (Jun. 1999).

The use of the SNPs of the present invention for human identification further extends to various authentication systems, commonly referred to as biometric systems, which typically convert physical characteristics of humans (or other organisms) into digital data. Biometric systems include various technological devices that measure such unique anatomical or physiological characteristics as finger, thumb, or palm prints; hand geometry; vein patterning on the back of the hand; blood vessel patterning of the retina and color and texture of the iris; facial characteristics; voice patterns; signature and typing dynamics; and DNA. Such physiological measurements can be used to verify identity and, for example, restrict or allow access based on the identification. Examples of applications for biometrics include physical area security, computer and network security, aircraft passenger check-in and boarding, financial transactions, medical records access, government benefit distribution, voting, law enforcement, passports, visas and immigration, prisons, various military applications, and for restricting access to expensive or dangerous items, such as automobiles or guns (see, for example, O'Connor, *Stanford Technology Law Review* and U.S. Patent No. 6,119,096).

Groups of SNPs, particularly the SNPs provided by the present invention, can be typed to uniquely identify an individual for biometric applications such as those described above. Such

SNP typing can readily be accomplished using, for example, DNA chips/arrays. Preferably, a minimally invasive means for obtaining a DNA sample is utilized. For example, PCR amplification enables sufficient quantities of DNA for analysis to be obtained from buccal swabs or fingerprints, which contain DNA-containing skin cells and oils that are naturally transferred during contact.

Further information regarding techniques for using SNPs in forensic/human identification applications can be found in, for example, *Current Protocols in Human Genetics* 14.1-14.7, John Wiley & Sons, New York (2002).

VARIANT PROTEINS, ANTIBODIES, VECTORS & HOST CELLS, & USES THEREOF

Variant Proteins Encoded by SNP-Containing Nucleic Acid Molecules

The present invention provides SNP-containing nucleic acid molecules, many of which encode proteins having variant amino acid sequences as compared to the art-known (i.e., wild-type) proteins. Amino acid sequences encoded by the polymorphic nucleic acid molecules of the present invention are provided as SEQ ID NOS:49-96 in Table 1 and the Sequence Listing. These variants will generally be referred to herein as variant proteins/peptides/polypeptides, or polymorphic proteins/peptides/polypeptides of the present invention. The terms "protein," "peptide," and "polypeptide" are used herein interchangeably.

A variant protein of the present invention may be encoded by, for example, a nonsynonymous nucleotide substitution at any one of the cSNP positions disclosed herein. In addition, variant proteins may also include proteins whose expression, structure, and/or function is altered by a SNP disclosed herein, such as a SNP that creates or destroys a stop codon, a SNP that affects splicing, and a SNP in control/regulatory elements, e.g. promoters, enhancers, or transcription factor binding domains.

As used herein, a protein or peptide is said to be "isolated" or "purified" when it is substantially free of cellular material or chemical precursors or other chemicals. The variant proteins of the present invention can be purified to homogeneity or other lower degrees of purity. The level of purification will be based on the intended use. The key feature is that the preparation allows for the desired function of the variant protein, even if in the presence of considerable amounts of other components.

As used herein, "substantially free of cellular material" includes preparations of the variant protein having less than about 30% (by dry weight) other proteins (i.e., contaminating protein), less than about 20% other proteins, less than about 10% other proteins, or less than about 5% other proteins. When the variant protein is recombinantly produced, it can also be substantially free of

culture medium, i.e., culture medium represents less than about 20% of the volume of the protein preparation.

The language "substantially free of chemical precursors or other chemicals" includes preparations of the variant protein in which it is separated from chemical precursors or other chemicals that are involved in its synthesis. In one embodiment, the language "substantially free of chemical precursors or other chemicals" includes preparations of the variant protein having less than about 30% (by dry weight) chemical precursors or other chemicals, less than about 20% chemical precursors or other chemicals, less than about 10% chemical precursors or other chemicals, or less than about 5% chemical precursors or other chemicals.

An isolated variant protein may be purified from cells that naturally express it, purified from cells that have been altered to express it (recombinant host cells), or synthesized using known protein synthesis methods. For example, a nucleic acid molecule containing SNP(s) encoding the variant protein can be cloned into an expression vector, the expression vector introduced into a host cell, and the variant protein expressed in the host cell. The variant protein can then be isolated from the cells by any appropriate purification scheme using standard protein purification techniques. Examples of these techniques are described in detail below (Sambrook and Russell, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY [2000]).

The present invention provides isolated variant proteins that comprise, consist of or consist essentially of amino acid sequences that contain one or more variant amino acids encoded by one or more codons which contain a SNP of the present invention.

Accordingly, the present invention provides variant proteins that consist of amino acid sequences that contain one or more amino acid polymorphisms (or truncations or extensions due to creation or destruction of a stop codon, respectively) encoded by the SNPs provided in Table 1 and/or Table 2. A protein consists of an amino acid sequence when the amino acid sequence is the entire amino acid sequence of the protein.

The present invention further provides variant proteins that consist essentially of amino acid sequences that contain one or more amino acid polymorphisms (or truncations or extensions due to creation or destruction of a stop codon, respectively) encoded by the SNPs provided in Table 1 and/or Table 2. A protein consists essentially of an amino acid sequence when such an amino acid sequence is present with only a few additional amino acid residues in the final protein.

The present invention further provides variant proteins that comprise amino acid sequences that contain one or more amino acid polymorphisms (or truncations or extensions due to creation or destruction of a stop codon, respectively) encoded by the SNPs provided in Table 1 and/or Table 2.

A protein comprises an amino acid sequence when the amino acid sequence is at least part of the final amino acid sequence of the protein. In such a fashion, the protein may contain only the variant amino acid sequence or have additional amino acid residues, such as a contiguous encoded sequence that is naturally associated with it or heterologous amino acid residues. Such a protein can have a few additional amino acid residues or can comprise many more additional amino acids. A brief description of how various types of these proteins can be made and isolated is provided below.

The variant proteins of the present invention can be attached to heterologous sequences to form chimeric or fusion proteins. Such chimeric and fusion proteins comprise a variant protein operatively linked to a heterologous protein having an amino acid sequence not substantially homologous to the variant protein. "Operatively linked" indicates that the coding sequences for the variant protein and the heterologous protein are ligated in-frame. The heterologous protein can be fused to the N-terminus or C-terminus of the variant protein. In another embodiment, the fusion protein is encoded by a fusion polynucleotide that is synthesized by conventional techniques including automated DNA synthesizers. Alternatively, PCR amplification of gene fragments can be carried out using anchor primers which give rise to complementary overhangs between two consecutive gene fragments which can subsequently be annealed and re-amplified to generate a chimeric gene sequence (see Ausubel *et al.*, *Current Protocols in Molecular Biology* [1992]). Moreover, many expression vectors are commercially available that already encode a fusion moiety (e.g., a GST protein). A variant protein-encoding nucleic acid can be cloned into such an expression vector such that the fusion moiety is linked in-frame to the variant protein.

In many uses, the fusion protein does not affect the activity of the variant protein. The fusion protein can include, but is not limited to, enzymatic fusion proteins, for example, beta-galactosidase fusions, yeast two-hybrid GAL fusions, poly-His fusions, MYC-tagged, HI-tagged and Ig fusions. Such fusion proteins, particularly poly-His fusions, can facilitate their purification following recombinant expression. In certain host cells (e.g., mammalian host cells), expression and/or secretion of a protein can be increased by using a heterologous signal sequence. Fusion proteins are further described in, for example, Terpe, "Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems," *Appl. Microbiol. Biotechnol.* 60(5):523-33 (Jan. 2003); Epub Nov. 7 2002; Graddis *et al.*, "Designing proteins that work using recombinant technologies," *Curr. Pharm. Biotechnol.* 3(4):285-97 (Dec. 2002); and Nilsson *et al.*, "Affinity fusion strategies for detection, purification, and immobilization of recombinant proteins," *Protein Expr. Purif.* 11(1):1-16 (Oct. 1997).

The present invention also relates to further obvious variants of the variant polypeptides of the present invention, such as naturally-occurring mature forms (e.g., allelic variants), non-naturally occurring recombinantly-derived variants, and orthologs and paralogs of such proteins that share sequence homology. Such variants can readily be generated using art-known techniques in the fields of recombinant nucleic acid technology and protein biochemistry. It is understood, however, that variants exclude those known in the prior art before the present invention.

Further variants of the variant polypeptides disclosed in Table 1 can comprise an amino acid sequence that shares at least 70-80%, 80-85%, 85-90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% sequence identity with an amino acid sequence disclosed in Table 1 (or a fragment thereof) and that includes a novel amino acid residue (allele) disclosed in Table 1 (which is encoded by a novel SNP allele). Thus, an aspect of the present invention that is specifically contemplated are polypeptides that have a certain degree of sequence variation compared with the polypeptide sequences shown in Table 1, but that contain a novel amino acid residue (allele) encoded by a novel SNP allele disclosed herein. In other words, as long as a polypeptide contains a novel amino acid residue disclosed herein, other portions of the polypeptide that flank the novel amino acid residue can vary to some degree from the polypeptide sequences shown in Table 1.

Full-length pre-processed forms, as well as mature processed forms, of proteins that comprise one of the amino acid sequences disclosed herein can readily be identified as having complete sequence identity to one of the variant proteins of the present invention as well as being encoded by the same genetic locus as the variant proteins provided herein.

Orthologs of a variant peptide can readily be identified as having some degree of significant sequence homology/identity to at least a portion of a variant peptide as well as being encoded by a gene from another organism. Preferred orthologs will be isolated from non-human mammals, preferably primates, for the development of human therapeutic targets and agents. Such orthologs can be encoded by a nucleic acid sequence that hybridizes to a variant peptide-encoding nucleic acid molecule under moderate to stringent conditions depending on the degree of relatedness of the two organisms yielding the homologous proteins.

Variant proteins include, but are not limited to, proteins containing deletions, additions and substitutions in the amino acid sequence caused by the SNPs of the present invention. One class of substitutions is conserved amino acid substitutions in which a given amino acid in a polypeptide is substituted for another amino acid of like characteristics. Typical conservative substitutions are replacements, one for another, among the aliphatic amino acids Ala, Val, Leu, and Ile; interchange of the hydroxyl residues Ser and Thr; exchange of the acidic residues Asp and Glu; substitution

between the amide residues Asn and Gln; exchange of the basic residues Lys and Arg; and replacements among the aromatic residues Phe and Tyr. Guidance concerning which amino acid changes are likely to be phenotypically silent are found in, for example, Bowie *et al.*, *Science* 247:1306-1310 (1990).

5 Variant proteins can be fully functional or can lack function in one or more activities, e.g. ability to bind another molecule, ability to catalyze a substrate, ability to mediate signaling, etc. Fully functional variants typically contain only conservative variations or variations in non-critical residues or in non-critical regions. Functional variants can also contain substitution of similar amino acids that result in no change or an insignificant change in function. Alternatively,
10 such substitutions may positively or negatively affect function to some degree. Non-functional variants typically contain one or more non-conservative amino acid substitutions, deletions, insertions, inversions, truncations or extensions, or a substitution, insertion, inversion, or deletion of a critical residue or in a critical region.

Amino acids that are essential for function of a protein can be identified by methods known
15 in the art, such as site-directed mutagenesis or alanine-scanning mutagenesis (Cunningham *et al.*, *Science* 244:1081-1085 [1989]), particularly using the amino acid sequence and polymorphism information provided in Table 1. The latter procedure introduces single alanine mutations at every residue in the molecule. The resulting mutant molecules are then tested for biological activity such as enzyme activity or in assays such as an *in vitro* proliferative activity. Sites that are critical for
20 binding partner/substrate binding can also be determined by structural analysis such as crystallization, nuclear magnetic resonance or photoaffinity labeling (Smith *et al.*, *J. Mol. Biol.* 224:899-904 [1992]; de Vos *et al.*, *Science* 255:306-312 [1992]).

Polypeptides can contain amino acids other than the 20 amino acids commonly referred to as the 20 naturally occurring amino acids. Further, many amino acids, including the terminal
25 amino acids, may be modified by natural processes, such as processing and other post-translational modifications, or by chemical modification techniques well known in the art. Accordingly, the variant proteins of the present invention also encompass derivatives or analogs in which a substituted amino acid residue is not one encoded by the genetic code, in which a substituent group is included, in which the mature polypeptide is fused with another compound,
30 such as a compound to increase the half-life of the polypeptide (e.g., polyethylene glycol), or in which additional amino acids are fused to the mature polypeptide, such as a leader or secretory sequence or a sequence for purification of the mature polypeptide or a pro-protein sequence.

Known protein modifications include, but are not limited to, acetylation, acylation, ADP-ribosylation, amidation, covalent attachment of flavin, covalent attachment of a heme moiety,

covalent attachment of a nucleotide or nucleotide derivative, covalent attachment of a lipid or lipid derivative, covalent attachment of phosphatidylinositol, cross-linking, cyclization, disulfide bond formation, demethylation, formation of covalent crosslinks, formation of cystine, formation of pyroglutamate, formylation, gamma carboxylation, glycosylation, GPI anchor formation, hydroxylation, iodination, methylation, myristoylation, oxidation, proteolytic processing, phosphorylation, prenylation, racemization, selenoylation, sulfation, transfer-RNA mediated addition of amino acids to proteins such as arginylation, and ubiquitination.

Such protein modifications are well known to those of skill in the art and have been described in great detail in the scientific literature. Several particularly common modifications, glycosylation, lipid attachment, sulfation, gamma-carboxylation of glutamic acid residues, hydroxylation and ADP-ribosylation, for instance, are described in most basic texts, such as *Proteins - Structure and Molecular Properties*, 2nd Ed., T.E. Creighton, W. H. Freeman and Company, New York (1993); F. Wold, *Posttranslational Covalent Modification of Proteins* 1-12, ed. B.C. Johnson, Academic Press, New York (1983); Seifter *et al.*, *Meth. Enzymol.* 182: 626-646 (1990); and Rattan *et al.*, *Ann. N.Y. Acad. Sci.* 663:48-62 (1992).

The present invention further provides fragments of the variant proteins in which the fragments contain one or more amino acid sequence variations (e.g., substitutions, or truncations or extensions due to creation or destruction of a stop codon) encoded by one or more SNPs disclosed herein. The fragments to which the invention pertains, however, are not to be construed as encompassing fragments that have been disclosed in the prior art before the present invention.

As used herein, a fragment may comprise at least about 4, 8, 10, 12, 14, 16, 18, 20, 25, 30, 50, 100 (or any other number in-between) or more contiguous amino acid residues from a variant protein, wherein at least one amino acid residue is affected by a SNP of the present invention, e.g., a variant amino acid residue encoded by a nonsynonymous nucleotide substitution at a cSNP position provided by the present invention. The variant amino acid encoded by a cSNP may occupy any residue position along the sequence of the fragment. Such fragments can be chosen based on the ability to retain one or more of the biological activities of the variant protein or the ability to perform a function, e.g., act as an immunogen. Particularly important fragments are biologically active fragments. Such fragments will typically comprise a domain or motif of a variant protein of the present invention, e.g., active site, transmembrane domain, or ligand/substrate binding domain. Other fragments include, but are not limited to, domain or motif-containing fragments, soluble peptide fragments, and fragments containing immunogenic structures. Predicted domains and functional sites are readily identifiable by computer programs well known to those of skill in the art

(e.g., PROSITE analysis) (*Current Protocols in Protein Science*, John Wiley & Sons, New York [2002]).

Uses of Variant Proteins

5 The variant proteins of the present invention can be used in a variety of ways, including but not limited to, in assays to determine the biological activity of a variant protein, such as in a panel of multiple proteins for high-throughput screening; to raise antibodies or to elicit another type of immune response; as a reagent (including the labeled reagent) in assays designed to quantitatively determine levels of the variant protein (or its binding partner) in biological fluids;
10 as a marker for cells or tissues in which it is preferentially expressed (either constitutively or at a particular stage of tissue differentiation or development or in a disease state); as a target for screening for a therapeutic agent; and as a direct therapeutic agent to be administered into a human subject. Any of the variant proteins disclosed herein may be developed into reagent grade or kit format for commercialization as research products. Methods for performing the uses listed
15 above are well known to those skilled in the art (see, e.g., *Molecular Cloning: A Laboratory Manual*, Sambrook and Russell, Cold Spring Harbor Laboratory Press, New York [2000], and *Methods in Enzymology: Guide to Molecular Cloning Techniques*, eds. S.L. Berger and A.R. Kimmel, Academic Press [1987]).

 In a specific embodiment of the invention, the methods of the present invention include
20 detection of one or more variant proteins disclosed herein. Variant proteins are disclosed in Table 1 and in the Sequence Listing as SEQ ID NOS: 49-96. Detection of such proteins can be accomplished using, for example, antibodies, small molecule compounds, aptamers, ligands/substrates, other proteins or protein fragments, or other protein-binding agents. Preferably, protein detection agents are specific for a variant protein of the present invention and
25 can therefore discriminate between a variant protein of the present invention and the wild-type protein or another variant form. This can generally be accomplished by, for example, selecting or designing detection agents that bind to the region of a protein that differs between the variant and wild-type protein, such as a region of a protein that contains one or more amino acid substitutions that is/are encoded by a non-synonymous cSNP of the present invention, or a region
30 of a protein that follows a nonsense mutation-type SNP that creates a stop codon thereby leading to a shorter polypeptide, or a region of a protein that follows a read-through mutation-type SNP that destroys a stop codon thereby leading to a longer polypeptide in which a portion of the polypeptide is present in one version of the polypeptide but not the other.

In another specific aspect of the invention, the variant proteins of the present invention are used as targets for diagnosing AD or for determining predisposition to AD in a human.

Accordingly, the invention provides methods for detecting the presence of, or levels of, one or more variant proteins of the present invention in a cell, tissue, or organism. Such methods typically

5 involve contacting a test sample with an agent (e.g., an antibody, small molecule compound, or peptide) capable of interacting with the variant protein such that specific binding of the agent to the variant protein can be detected. Such an assay can be provided in a single detection format or a multi-detection format such as an array, for example, an antibody or aptamer array (arrays for protein detection may also be referred to as "protein chips"). The variant protein of interest can be
10 isolated from a test sample and assayed for the presence of a variant amino acid sequence encoded by one or more SNPs disclosed by the present invention. The SNPs may cause changes to the protein and the corresponding protein function/activity, such as through non-synonymous substitutions in protein coding regions that can lead to amino acid substitutions, deletions, insertions, and/or rearrangements; formation or destruction of stop codons; or alteration of control elements
15 such as promoters. SNPs may also cause inappropriate post-translational modifications.

One preferred agent for detecting a variant protein in a sample is an antibody capable of selectively binding to a variant form of the protein (antibodies are described in greater detail in the next section). Such samples include, for example, tissues, cells, and biological fluids isolated from a subject, as well as tissues, cells and fluids present within a subject.

20 *In vitro* methods for detection of the variant proteins associated with AD that are disclosed herein and fragments thereof include, but are not limited to, enzyme linked immunosorbent assays (ELISAs), radioimmunoassays (RIA), Western blots, immunoprecipitations, immunofluorescence, and protein arrays/chips (e.g., arrays of antibodies or aptamers). For further information regarding immunoassays and related protein detection methods, see *Current Protocols in Immunology*, John
25 Wiley & Sons, New York, and Hage, "Immunoassays," *Anal. Chem.* 71(12):294R-304R (Jun. 15 1999).

Additional analytic methods of detecting amino acid variants include, but are not limited to, altered electrophoretic mobility, altered tryptic peptide digest, altered protein activity in cell-based or cell-free assay, alteration in ligand or antibody-binding pattern, altered isoelectric point, and
30 direct amino acid sequencing.

Alternatively, variant proteins can be detected *in vivo* in a subject by introducing into the subject a labeled antibody (or other type of detection reagent) specific for a variant protein. For example, the antibody can be labeled with a radioactive marker whose presence and location in a subject can be detected by standard imaging techniques.

Other uses of the variant peptides of the present invention are based on the class or action of the protein. For example, proteins isolated from humans and their mammalian orthologs serve as targets for identifying agents (e.g., small molecule drugs or antibodies) for use in therapeutic applications, particularly for modulating a biological or pathological response in a cell or tissue
5 that expresses the protein. Pharmaceutical agents can be developed that modulate protein activity.

As an alternative to modulating gene expression, therapeutic compounds can be developed that modulate protein function. For example, many SNPs disclosed herein affect the amino acid sequence of the encoded protein (e.g., non-synonymous cSNPs and nonsense mutation-type SNPs). Such alterations in the encoded amino acid sequence may affect protein function, particularly if such
10 amino acid sequence variations occur in functional protein domains, such as catalytic domains, ATP-binding domains, or ligand/substrate binding domains. It is well established in the art that variant proteins having amino acid sequence variations in functional domains can cause or influence pathological conditions. In such instances, compounds (e.g., small molecule drugs or antibodies) can be developed that target the variant protein and modulate (e.g., up- or down-regulate) protein
15 function/activity.

The therapeutic methods of the present invention further include methods that target one or more variant proteins of the present invention. Variant proteins can be targeted using, for example, small molecule compounds, antibodies, aptamers, ligands/substrates, other proteins, or other protein-binding agents. Additionally, the skilled artisan will recognize that the novel
20 protein variants (and polymorphic nucleic acid molecules) disclosed in Table 1 may themselves be directly used as therapeutic agents by acting as competitive inhibitors of corresponding art-known proteins (or nucleic acid molecules such as mRNA molecules).

The variant proteins of the present invention are particularly useful in drug screening assays, in cell-based or cell-free systems. Cell-based systems can utilize cells that naturally express the
25 protein, a biopsy specimen, or cell cultures. In one embodiment, cell-based assays involve recombinant host cells expressing the variant protein. Cell-free assays can be used to detect the ability of a compound to directly bind to a variant protein or to the corresponding SNP-containing nucleic acid fragment that encodes the variant protein.

A variant protein of the present invention, as well as appropriate fragments thereof, can be
30 used in high-throughput screening assays to test candidate compounds for the ability to bind and/or modulate the activity of the variant protein. These candidate compounds can be further screened against a protein having normal function (e.g., a wild-type/non-variant protein) to further determine the effect of the compound on the protein activity. Furthermore, these compounds can be tested in animal or invertebrate systems to determine *in vivo* activity/effectiveness. Compounds can be

identified that activate (agonists) or inactivate (antagonists) the variant protein, and different compounds can be identified that cause various degrees of activation or inactivation of the variant protein.

Further, the variant proteins can be used to screen a compound for the ability to stimulate or inhibit interaction between the variant protein and a target molecule that normally interacts with the protein. The target can be a ligand, a substrate or a binding partner that the protein normally interacts with (for example, epinephrine or norepinephrine). Such assays typically include the steps of combining the variant protein with a candidate compound under conditions that allow the variant protein, or fragment thereof, to interact with the target molecule, and to detect the formation of a complex between the protein and the target or to detect the biochemical consequence of the interaction with the variant protein and the target, such as any of the associated effects of signal transduction.

Candidate compounds include, for example, 1) peptides such as soluble peptides, including Ig-tailed fusion peptides and members of random peptide libraries (see, e.g., Lam *et al.*, *Nature* 354:82-84 [1991]; Houghten *et al.*, *Nature* 354:84-86 [1991]) and combinatorial chemistry-derived molecular libraries made of D- and/or L- configuration amino acids; 2) phosphopeptides (e.g., members of random and partially degenerate, directed phosphopeptide libraries, see, e.g., Songyang *et al.*, *Cell* 72:767-778 [1993]); 3) antibodies (e.g., polyclonal, monoclonal, humanized, anti-idiotypic, chimeric, and single chain antibodies as well as Fab, F(ab')₂, Fab expression library fragments, and epitope-binding fragments of antibodies); and 4) small organic and inorganic molecules (e.g., molecules obtained from combinatorial and natural product libraries).

One candidate compound is a soluble fragment of the variant protein that competes for ligand binding. Other candidate compounds include mutant proteins or appropriate fragments containing mutations that affect variant protein function and thus compete for ligand. Accordingly, a fragment that competes for ligand, for example with a higher affinity, or a fragment that binds ligand but does not allow release, is encompassed by the invention.

The invention further includes other end point assays to identify compounds that modulate (stimulate or inhibit) variant protein activity. The assays typically involve an assay of events in the signal transduction pathway that indicate protein activity. Thus, the expression of genes that are up or down-regulated in response to the variant protein dependent signal cascade can be assayed. In one embodiment, the regulatory region of such genes can be operably linked to a marker that is easily detectable, such as luciferase. Alternatively, phosphorylation of the variant protein, or a variant protein target, could also be measured. Any of the biological or biochemical functions mediated by the variant protein can be used as an endpoint assay. These include all of the

biochemical or biological events described herein, in the references cited herein, incorporated by reference for these endpoint assay targets, and other functions known to those of ordinary skill in the art.

Binding and/or activating compounds can also be screened by using chimeric variant proteins in which an amino terminal extracellular domain or parts thereof, an entire transmembrane domain or subregions, and/or the carboxyl terminal intracellular domain or parts thereof, can be replaced by heterologous domains or subregions. For example, a substrate-binding region can be used that interacts with a different substrate than that which is normally recognized by a variant protein. Accordingly, a different set of signal transduction components is available as an end-point assay for activation. This allows for assays to be performed in other than the specific host cell from which the variant protein is derived.

The variant proteins are also useful in competition binding assays in methods designed to discover compounds that interact with the variant protein. Thus, a compound can be exposed to a variant protein under conditions that allow the compound to bind or to otherwise interact with the variant protein. A binding partner, such as ligand, that normally interacts with the variant protein is also added to the mixture. If the test compound interacts with the variant protein or its binding partner, it decreases the amount of complex formed or activity from the variant protein. This type of assay is particularly useful in screening for compounds that interact with specific regions of the variant protein (Hodgson, *Bio/technology*, 10[9], 973-80 [Sept. 1992]).

To perform cell-free drug screening assays, it is sometimes desirable to immobilize either the variant protein or a fragment thereof, or its target molecule, to facilitate separation of complexes from uncomplexed forms of one or both of the proteins, as well as to accommodate automation of the assay. Any method for immobilizing proteins on matrices can be used in drug screening assays. In one embodiment, a fusion protein containing an added domain allows the protein to be bound to a matrix. For example, glutathione-S-transferase/¹²⁵I fusion proteins can be adsorbed onto glutathione sepharose beads (Sigma Chemical, St. Louis, MO) or glutathione derivatized microtitre plates, which are then combined with the cell lysates (e.g., ³⁵S-labeled) and a candidate compound, such as a drug candidate, and the mixture incubated under conditions conducive to complex formation (e.g., at physiological conditions for salt and pH). Following incubation, the beads can be washed to remove any unbound label, and the matrix immobilized and radiolabel determined directly, or in the supernatant after the complexes are dissociated. Alternatively, the complexes can be dissociated from the matrix, separated by SDS-PAGE, and the level of bound material found in the bead fraction quantitated from the gel using standard electrophoretic techniques.

Either the variant protein or its target molecule can be immobilized utilizing conjugation of biotin and streptavidin. Alternatively, antibodies reactive with the variant protein but which do not interfere with binding of the variant protein to its target molecule can be derivatized to the wells of the plate, and the variant protein trapped in the wells by antibody conjugation. Preparations of the target molecule and a candidate compound are incubated in the variant protein-presenting wells and the amount of complex trapped in the well can be quantitated. Methods for detecting such complexes, in addition to those described above for the GST-immobilized complexes, include immunodetection of complexes using antibodies reactive with the protein target molecule, or which are reactive with variant protein and compete with the target molecule, and enzyme-linked assays that rely on detecting an enzymatic activity associated with the target molecule.

Modulators of variant protein activity identified according to these drug screening assays can be used to treat a subject with a disorder mediated by the protein pathway, such as AD. These methods of treatment typically include the steps of administering the modulators of protein activity in a pharmaceutical composition to a subject in need of such treatment.

The variant proteins, or fragments thereof, disclosed herein can themselves be directly used to treat a disorder characterized by an absence of, inappropriate, or unwanted expression or activity of the variant protein. Accordingly, methods for treatment include the use of a variant protein disclosed herein or fragments thereof.

In yet another aspect of the invention, variant proteins can be used as "bait proteins" in a two-hybrid assay or three-hybrid assay (see, e.g., U.S. Patent No. 5,283,317; Zervos *et al.*, *Cell* 72:223-232 [1993]; Madura *et al.*, *J. Biol. Chem.* 268:12046-12054 [1993]; Bartel *et al.*, *Biotechniques* 14:920-924 [1993]; Iwabuchi *et al.*, *Oncogene* 8:1693-1696 [1993]; and Brent WO94/10300) to identify other proteins that bind to or interact with the variant protein and are involved in variant protein activity. Such variant protein-binding proteins are also likely to be involved in the propagation of signals by the variant proteins or variant protein targets as, for example, elements of a protein-mediated signaling pathway. Alternatively, such variant protein-binding proteins are inhibitors of the variant protein.

The two-hybrid system is based on the modular nature of most transcription factors, which typically consist of separable DNA-binding and activation domains. Briefly, the assay typically utilizes two different DNA constructs. In one construct, the gene that codes for a variant protein is fused to a gene encoding the DNA binding domain of a known transcription factor (e.g., GAL-4). In the other construct, a DNA sequence, from a library of DNA sequences, that encodes an unidentified protein ("prey" or "sample") is fused to a gene that codes for the activation domain of the known transcription factor. If the "bait" and the "prey" proteins are able

to interact, *in vivo*, forming a variant protein-dependent complex, the DNA-binding and activation domains of the transcription factor are brought into close proximity. This proximity allows transcription of a reporter gene (e.g., LacZ) that is operably linked to a transcriptional regulatory site responsive to the transcription factor. Expression of the reporter gene can be detected, and cell colonies containing the functional transcription factor can be isolated and used to obtain the cloned gene that encodes the protein that interacts with the variant protein.

Antibodies Directed to Variant Proteins

The present invention also provides antibodies that selectively bind to the variant proteins disclosed herein and fragments thereof. Such antibodies may be used to quantitatively or qualitatively detect the variant proteins of the present invention. As used herein, an antibody selectively binds a target variant protein when it binds the variant protein and does not significantly bind to non-variant proteins, i.e., the antibody does not significantly bind to normal, wild-type, or art-known proteins that do not contain a variant amino acid sequence due to one or more SNPs of the present invention (variant amino acid sequences may be due to, for example, nonsynonymous cSNPs, nonsense SNPs that create a stop codon, thereby causing a truncation of a polypeptide or SNPs that cause read-through mutations resulting in an extension of a polypeptide).

As used herein, an antibody is defined in terms consistent with that recognized in the art: they are multi-subunit proteins produced by an organism in response to an antigen challenge. The antibodies of the present invention include both monoclonal antibodies and polyclonal antibodies, as well as antigen-reactive proteolytic fragments of such antibodies, such as Fab, F(ab)₂, and Fv fragments. In addition, an antibody of the present invention further includes any of a variety of engineered antigen-binding molecules such as a chimeric antibody (U.S. Patent Nos. 4,816,567 and 4,816,397; Morrison *et al.*, *Proc. Natl. Acad. Sci. USA*, 81:6851 [1984]; Neuberger *et al.*, *Nature* 312:604 [1984]), a humanized antibody (U.S. Patent Nos. 5,693,762; 5,585,089; and 5,565,332), a single-chain Fv (U.S. Patent No. 4,946,778; Ward *et al.*, *Nature* 334:544 [1989]), a bispecific antibody with two binding specificities (Segal *et al.*, *J. Immunol. Methods* 248:1 [2001]; Carter, *J. Immunol. Methods* 248:7 [2001]), a diabody, a triabody, and a tetrabody (Todorovska *et al.*, *J. Immunol. Methods*, 248:47 [2001]), as well as a Fab conjugate (dimer or trimer), and a minibody.

Many methods are known in the art for generating and/or identifying antibodies to a given target antigen (Harlow, *Antibodies*, Cold Spring Harbor Press, New York [1989]). In general, an isolated peptide (e.g., a variant protein of the present invention) is used as an immunogen and is administered to a mammalian organism, such as a rat, rabbit, hamster or mouse. Either a full-length protein, an antigenic peptide fragment (e.g., a peptide fragment containing a region that varies

between a variant protein and a corresponding wild-type protein), or a fusion protein can be used. A protein used as an immunogen may be naturally-occurring, synthetic or recombinantly produced, and may be administered in combination with an adjuvant, including but not limited to, Freund's (complete and incomplete), mineral gels such as aluminum hydroxide, surface active substance such as lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, keyhole limpet hemocyanin, dinitrophenol, and the like.

Monoclonal antibodies can be produced by hybridoma technology (Kohler and Milstein, *Nature*, 256:495 [1975]), which immortalizes cells secreting a specific monoclonal antibody. The immortalized cell lines can be created *in vitro* by fusing two different cell types, typically lymphocytes, and tumor cells. The hybridoma cells may be cultivated *in vitro* or *in vivo*. Additionally, fully human antibodies can be generated by transgenic animals (He *et al.*, *J. Immunol.*, 169:595 [2002]). Fd phage and Fd phagemid technologies may be used to generate and select recombinant antibodies *in vitro* (Hoogenboom and Chames, *Immunol. Today* 21:371 [2000]; Liu *et al.*, *J. Mol. Biol.* 315:1063 [2002]). The complementarity-determining regions of an antibody can be identified, and synthetic peptides corresponding to such regions may be used to mediate antigen binding (U.S. Patent No. 5,637,677).

Antibodies are preferably prepared against regions or discrete fragments of a variant protein containing a variant amino acid sequence as compared to the corresponding wild-type protein (e.g., a region of a variant protein that includes an amino acid encoded by a nonsynonymous cSNP, a region affected by truncation caused by a nonsense SNP that creates a stop codon, or a region resulting from the destruction of a stop codon due to read-through mutation caused by a SNP). Furthermore, preferred regions will include those involved in function/activity and/or protein/binding partner interaction. Such fragments can be selected on a physical property, such as fragments corresponding to regions that are located on the surface of the protein, e.g., hydrophilic regions, or can be selected based on sequence uniqueness, or based on the position of the variant amino acid residue(s) encoded by the SNPs provided by the present invention. An antigenic fragment will typically comprise at least about 8-10 contiguous amino acid residues in which at least one of the amino acid residues is an amino acid affected by a SNP disclosed herein. The antigenic peptide can comprise, however, at least 12, 14, 16, 20, 25, 50, 100 (or any other number in-between) or more amino acid residues, provided that at least one amino acid is affected by a SNP disclosed herein.

Detection of an antibody of the present invention can be facilitated by coupling (i.e., physically linking) the antibody or an antigen-reactive fragment thereof to a detectable substance. Detectable substances include, but are not limited to, various enzymes, prosthetic groups,

fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase, β -galactosidase, or acetylcholinesterase; examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin; examples of suitable fluorescent materials include
 5 umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin; an example of a luminescent material includes luminol; examples of bioluminescent materials include luciferase, luciferin, and aequorin, and examples of suitable radioactive material include ^{125}I , ^{131}I , ^{35}S or ^3H .

Antibodies, particularly the use of antibodies as therapeutic agents, are reviewed in: Morgan,
 10 "Antibody therapy for Alzheimer's Disease," *Expert Rev. Vaccines*. 2(1):53-9 (Feb. 2003); Ross et al., "Anticancer antibodies," *Am. J. Clin. Pathol.* 119(4):472-85 (Apr. 2003); Goldenberg, "Advancing role of radiolabeled antibodies in the therapy of cancer", *Cancer Immunol. Immunother.* 52(5):281-96 (May 2003); Epub Mar. 11 2003; Ross et al., "Antibody-based therapeutics in oncology," *Expert Rev. Anticancer Ther.* 3(1):107-21 (Feb. 2003); Cao et al., "Bispecific antibody
 15 conjugates in therapeutics," *Adv. Drug Deliv. Rev.* 55(2):171-97 (Feb. 10 2003); von Mehren et al., "Monoclonal antibody therapy for cancer," *Annu. Rev. Med.* 54:343-69 (2003); Epub Dec. 3 2001; Hudson et al., "Engineered antibodies," *Nat. Med.* 9(1):129-34 (Jan. 2003); Brekke et al., "Therapeutic antibodies for human diseases at the dawn of the twenty-first century," *Nat. Rev. Drug
 Discov.* 2(1):52-62 (Jan. 2003) (Erratum in: *Nat. Rev. Drug Discov.* 2[3]:240 [Mar. 2003]);
 20 Houdebine, "Antibody manufacture in transgenic animals and comparisons with other systems," *Curr. Opin. Biotechnol.* 13(6):625-9 (Dec. 2002); Andreacos et al., "Monoclonal antibodies in immune and inflammatory diseases," *Curr. Opin. Biotechnol.* 13(6):615-20 (Dec. 2002); Kellermann et al., "Antibody discovery: the use of transgenic mice to generate human monoclonal antibodies for therapeutics," *Curr. Opin. Biotechnol.* 13(6):593-7 (Dec. 2002); Pini et al., "Phage
 25 display and colony filter screening for high-throughput selection of antibody libraries," *Comb. Chem. High Throughput Screen* 5(7):503-10 (Nov. 2002); Batra et al., "Pharmacokinetics and biodistribution of genetically engineered antibodies," *Curr. Opin. Biotechnol.* 13(6):603-8 (Dec. 2002); and Tangri et al., "Rationally engineered proteins or antibodies with absent or reduced immunogenicity," *Curr. Med. Chem.* 9(24):2191-9 (Dec. 2002).

30

Uses of Antibodies

Antibodies can be used to isolate the variant proteins of the present invention from a natural cell source or from recombinant host cells by standard techniques, such as affinity chromatography or immunoprecipitation. In addition, antibodies are useful for detecting the presence of a variant

protein of the present invention in cells or tissues to determine the pattern of expression of the variant protein among various tissues in an organism and over the course of normal development or disease progression. Further, antibodies can be used to detect variant protein *in situ*, *in vitro*, in a bodily fluid, or in a cell lysate or supernatant in order to evaluate the amount and pattern of expression. Also, antibodies can be used to assess abnormal tissue distribution, abnormal expression during development, or expression in an abnormal condition, such as AD. Additionally, antibody detection of circulating fragments of the full-length variant protein can be used to identify turnover.

Antibodies to the variant proteins of the present invention are also useful in pharmacogenomic analysis. Thus, antibodies against variant proteins encoded by alternative SNP alleles can be used to identify individuals that require modified treatment modalities.

Further, antibodies can be used to assess expression of the variant protein in disease states such as in active stages of the disease or in an individual with a predisposition to a disease related to the protein's function, particularly AD. Antibodies specific for a variant protein encoded by a SNP-containing nucleic acid molecule of the present invention can be used to assay for the presence of the variant protein, such as to screen for predisposition to AD as indicated by the presence of the variant protein.

Antibodies are also useful as diagnostic tools for evaluating the variant proteins in conjunction with analysis by electrophoretic mobility, isoelectric point, tryptic peptide digest, and other physical assays well known in the art.

Antibodies are also useful for tissue typing. Thus, where a specific variant protein has been correlated with expression in a specific tissue, antibodies that are specific for this protein can be used to identify a tissue type.

Antibodies can also be used to assess aberrant subcellular localization of a variant protein in cells in various tissues. The diagnostic uses can be applied, not only in genetic testing, but also in monitoring a treatment modality. Accordingly, where treatment is ultimately aimed at correcting the expression level or the presence of variant protein or aberrant tissue distribution or developmental expression of a variant protein, antibodies directed against the variant protein or relevant fragments can be used to monitor therapeutic efficacy.

The antibodies are also useful for inhibiting variant protein function, for example, by blocking the binding of a variant protein to a binding partner. These uses can also be applied in a therapeutic context in which treatment involves inhibiting a variant protein's function. An antibody can be used, for example, to block or competitively inhibit binding, thus modulating (agonizing or antagonizing) the activity of a variant protein. Antibodies can be prepared against specific variant protein fragments containing sites required for function or against an intact variant protein that is

associated with a cell or cell membrane. For *in vivo* administration, an antibody may be linked with an additional therapeutic payload such as a radionuclide, an enzyme, an immunogenic epitope, or a cytotoxic agent. Suitable cytotoxic agents include, but are not limited to, bacterial toxin such as diphtheria, and plant toxin such as ricin. The *in vivo* half-life of an antibody or a fragment thereof
5 may be lengthened by pegylation through conjugation to polyethylene glycol (Leong *et al.*, *Cytokine* 16:106 [2001]).

The invention also encompasses kits for using antibodies, such as kits for detecting the presence of a variant protein in a test sample. An exemplary kit can comprise antibodies such as a labeled or labelable antibody and a compound or agent for detecting variant proteins in a biological
10 sample; means for determining the amount, or presence/absence of variant protein in the sample; means for comparing the amount of variant protein in the sample with a standard; and instructions for use.

Vectors and Host Cells

15 The present invention also provides vectors containing the SNP-containing nucleic acid molecules described herein. The term "vector" refers to a vehicle, preferably a nucleic acid molecule, which can transport a SNP-containing nucleic acid molecule. When the vector is a nucleic acid molecule, the SNP-containing nucleic acid molecule can be covalently linked to the vector nucleic acid. Such vectors include, but are not limited to, a plasmid, single or double
20 stranded phage, a single or double stranded RNA or DNA viral vector, or artificial chromosome, such as a BAC, PAC, YAC, or MAC.

A vector can be maintained in a host cell as an extrachromosomal element where it replicates and produces additional copies of the SNP-containing nucleic acid molecules. Alternatively, the vector may integrate into the host cell genome and produce additional copies of
25 the SNP-containing nucleic acid molecules when the host cell replicates.

The invention provides vectors for the maintenance (cloning vectors) or vectors for expression (expression vectors) of the SNP-containing nucleic acid molecules. The vectors can function in prokaryotic or eukaryotic cells or in both (shuttle vectors).

Expression vectors typically contain *cis*-acting regulatory regions that are operably linked in
30 the vector to the SNP-containing nucleic acid molecules such that transcription of the SNP-containing nucleic acid molecules is allowed in a host cell. The SNP-containing nucleic acid molecules can also be introduced into the host cell with a separate nucleic acid molecule capable of affecting transcription. Thus, the second nucleic acid molecule may provide a trans-acting factor interacting with the *cis*-regulatory control region to allow transcription of the SNP-containing

nucleic acid molecules from the vector. Alternatively, a trans-acting factor may be supplied by the host cell. Finally, a trans-acting factor can be produced from the vector itself. It is understood, however, that in some embodiments, transcription and/or translation of the nucleic acid molecules can occur in a cell-free system.

5 The regulatory sequences to which the SNP-containing nucleic acid molecules described herein can be operably linked include promoters for directing mRNA transcription. These include, but are not limited to, the left promoter from bacteriophage λ , the lac, TRP, and TAC promoters from *E. coli*, the early and late promoters from SV40, the CMV immediate early promoter, the adenovirus early and late promoters, and retrovirus long-terminal repeats.

10 In addition to control regions that promote transcription, expression vectors may also include regions that modulate transcription, such as repressor binding sites and enhancers. Examples include the SV40 enhancer, the cytomegalovirus immediate early enhancer, polyoma enhancer, adenovirus enhancers, and retrovirus LTR enhancers.

 In addition to containing sites for transcription initiation and control, expression vectors can
15 also contain sequences necessary for transcription termination and, in the transcribed region, a ribosome-binding site for translation. Other regulatory control elements for expression include initiation and termination codons as well as polyadenylation signals. A person of ordinary skill in the art would be aware of the numerous regulatory sequences that are useful in expression vectors (see, e.g., Sambrook and Russell, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor
20 Laboratory Press, Cold Spring Harbor, New York [2000]).

 A variety of expression vectors can be used to express a SNP-containing nucleic acid molecule. Such vectors include chromosomal, episomal, and virus-derived vectors, for example, vectors derived from bacterial plasmids, from bacteriophage, from yeast episomes, from yeast chromosomal elements, including yeast artificial chromosomes, from viruses such as baculoviruses,
25 papovaviruses such as SV40, Vaccinia viruses, adenoviruses, poxviruses, pseudorabies viruses, and retroviruses. Vectors can also be derived from combinations of these sources such as those derived from plasmid and bacteriophage genetic elements, e.g., cosmids and phagemids. Appropriate cloning and expression vectors for prokaryotic and eukaryotic hosts are described in Sambrook and Russell, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold
30 Spring Harbor, New York (2000).

 The regulatory sequence in a vector may provide constitutive expression in one or more host cells (e.g., tissue specific expression) or may provide for inducible expression in one or more cell types such as by temperature, nutrient additive, or exogenous factor, e.g., a hormone or other ligand.

A variety of vectors that provide constitutive or inducible expression of a nucleic acid sequence in prokaryotic and eukaryotic host cells are well known to those of ordinary skill in the art.

A SNP-containing nucleic acid molecule can be inserted into the vector by methodology well-known in the art. Generally, the SNP-containing nucleic acid molecule that will ultimately be expressed is joined to an expression vector by cleaving the SNP-containing nucleic acid molecule and the expression vector with one or more restriction enzymes and then ligating the fragments together. Procedures for restriction enzyme digestion and ligation are well known to those of ordinary skill in the art.

The vector containing the appropriate nucleic acid molecule can be introduced into an appropriate host cell for propagation or expression using well-known techniques. Bacterial host cells include, but are not limited to, *E. coli*, *Streptomyces*, and *Salmonella typhimurium*. Eukaryotic host cells include, but are not limited to, yeast, insect cells such as *Drosophila*, animal cells such as COS and CHO cells, and plant cells.

As described herein, it may be desirable to express the variant peptide as a fusion protein. Accordingly, the invention provides fusion vectors that allow for the production of the variant peptides. Fusion vectors can, for example, increase the expression of a recombinant protein, increase the solubility of the recombinant protein, and aid in the purification of the protein by acting, for example, as a ligand for affinity purification. A proteolytic cleavage site may be introduced at the junction of the fusion moiety so that the desired variant peptide can ultimately be separated from the fusion moiety. Proteolytic enzymes suitable for such use include, but are not limited to, factor Xa, thrombin, and enterokinase. Typical fusion expression vectors include pGEX (Smith *et al.*, *Gene* 67:31-40 [1988]), pMAL (New England Biolabs, Beverly, MA) and pRIT5 (Pharmacia, Piscataway, NJ) which fuse glutathione S-transferase (GST), maltose E binding protein, or protein A, respectively, to the target recombinant protein. Examples of suitable inducible non-fusion *E. coli* expression vectors include pTrc (Amann *et al.*, *Gene* 69:301-315 [1988]) and pET 11d (Studier *et al.*, *Gene Expression Technology: Methods in Enzymology* 185:60-89 [1990]).

Recombinant protein expression can be maximized in a bacterial host by providing a genetic background wherein the host cell has an impaired capacity to proteolytically cleave the recombinant protein (S. Gottesman, *Gene Expression Technology: Methods in Enzymology* 185 119-128, Academic Press, San Diego, California [1990]). Alternatively, the sequence of the SNP-containing nucleic acid molecule of interest can be altered to provide preferential codon usage for a specific host cell, for example, *E. coli* (Wada *et al.*, *Nucleic Acids Res.* 20:2111-2118 [1992]).

The SNP-containing nucleic acid molecules can also be expressed by expression vectors that are operative in yeast. Examples of vectors for expression in yeast (e.g., *S. cerevisiae*) include

pYepSec1 (Baldari, *et al.*, *EMBO J.* 6:229-234 [1987]), pMFa (Kurjan *et al.*, *Cell* 30:933-943 [1982]), pJRY88 (Schultz *et al.*, *Gene* 54:113-123 [1987]), and pYES2 (Invitrogen Corporation, San Diego, CA).

5 The SNP-containing nucleic acid molecules can also be expressed in insect cells using, for example, baculovirus expression vectors. Baculovirus vectors available for expression of proteins in cultured insect cells (e.g., Sf 9 cells) include the pAc series (Smith *et al.*, *Mol. Cell Biol.* 3:2156-2165 [1983]) and the pVL series (Lucklow *et al.*, *Virology* 170:31-39 [1989]).

10 In certain embodiments of the invention, the SNP-containing nucleic acid molecules described herein are expressed in mammalian cells using mammalian expression vectors. Examples of mammalian expression vectors include pCDM8 (B. Seed, *Nature* 329:840 [1987]) and pMT2PC (Kaufman *et al.*, *EMBO J.* 6:187-195 [1987]).

15 The invention also encompasses vectors in which the SNP-containing nucleic acid molecules described herein are cloned into the vector in reverse orientation, but operably linked to a regulatory sequence that permits transcription of antisense RNA. Thus, an antisense transcript can be produced to the SNP-containing nucleic acid sequences described herein, including both coding and non-coding regions. Expression of this antisense RNA is subject to each of the parameters described above in relation to expression of the sense RNA (regulatory sequences, constitutive or inducible expression, tissue-specific expression).

20 The invention also relates to recombinant host cells containing the vectors described herein. Host cells therefore include, for example, prokaryotic cells, lower eukaryotic cells such as yeast, other eukaryotic cells such as insect cells, and higher eukaryotic cells such as mammalian cells.

25 The recombinant host cells can be prepared by introducing the vector constructs described herein into the cells by techniques readily available to persons of ordinary skill in the art. These include, but are not limited to, calcium phosphate transfection, DEAE-dextran-mediated transfection, cationic lipid-mediated transfection, electroporation, transduction, infection, lipofection, and other techniques such as those described in Sambrook and Russell, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (2000).

30 Host cells can contain more than one vector. Thus, different SNP-containing nucleotide sequences can be introduced in different vectors into the same cell. Similarly, the SNP-containing nucleic acid molecules can be introduced either alone or with other nucleic acid molecules that are not related to the SNP-containing nucleic acid molecules, such as those providing trans-acting factors for expression vectors. When more than one vector is introduced into a cell, the vectors can be introduced independently, co-introduced, or joined to the nucleic acid molecule vector.

In the case of bacteriophage and viral vectors, these can be introduced into cells as packaged or encapsulated virus by standard procedures for infection and transduction. Viral vectors can be replication-competent or replication-defective. In the case in which viral replication is defective, replication can occur in host cells that provide functions that complement the defects.

5 Vectors generally include selectable markers that enable the selection of the subpopulation of cells that contain the recombinant vector constructs. The marker can be inserted in the same vector that contains the SNP-containing nucleic acid molecules described herein or may be in a separate vector. Markers include, for example, tetracycline or ampicillin-resistance genes for prokaryotic host cells, and dihydrofolate reductase or neomycin resistance genes for eukaryotic host
10 cells. However, any marker that provides selection for a phenotypic trait can be effective.

 While the mature variant proteins can be produced in bacteria, yeast, mammalian cells, and other cells under the control of the appropriate regulatory sequences, cell-free transcription and translation systems can also be used to produce these variant proteins using RNA derived from the DNA constructs described herein.

15 Where secretion of the variant protein is desired, which is difficult to achieve with multi-transmembrane domain containing proteins such as G-protein-coupled receptors (GPCRs), appropriate secretion signals can be incorporated into the vector. The signal sequence can be endogenous to the peptides or heterologous to these peptides.

 Where the variant protein is not secreted into the medium, the protein can be isolated from
20 the host cell by standard disruption procedures, including freeze/thaw, sonication, mechanical disruption, use of lysing agents, and the like. The variant protein can then be recovered and purified by well-known purification methods including, for example, ammonium sulfate precipitation, acid extraction, anion or cationic exchange chromatography, phosphocellulose chromatography, hydrophobic-interaction chromatography, affinity chromatography, hydroxylapatite
25 chromatography, lectin chromatography, or high performance liquid chromatography.

 It is also understood that, depending upon the host cell in which recombinant production of the variant proteins described herein occurs, they can have various glycosylation patterns, or may be non-glycosylated, as when produced in bacteria. In addition, the variant proteins may include an initial modified methionine in some cases as a result of a host-mediated process.

30 For further information regarding vectors and host cells, see *Current Protocols in Molecular Biology*, John Wiley & Sons, New York.

Uses of Vectors and Host Cells, and Transgenic Animals

Recombinant host cells that express the variant proteins described herein have a variety of uses. For example, the cells are useful for producing a variant protein that can be further purified into a preparation of desired amounts of the variant protein or fragments thereof. Thus, host cells
5 containing expression vectors are useful for variant protein production.

Host cells are also useful for conducting cell-based assays involving the variant protein or variant protein fragments, such as those described above as well as other formats known in the art. Thus, a recombinant host cell expressing a variant protein is useful for assaying compounds that stimulate or inhibit variant protein function. Such an ability of a compound to modulate
10 variant protein function may not be apparent from assays of the compound on the native/wild-type protein, or from cell-free assays of the compound. Recombinant host cells are also useful for assaying functional alterations in the variant proteins as compared with a known function.

Genetically-engineered host cells can be further used to produce non-human transgenic animals. A transgenic animal is preferably a non-human mammal, for example, a rodent, such as a
15 rat or mouse, in which one or more of the cells of the animal include a transgene. A transgene is exogenous DNA containing a SNP of the present invention which is integrated into the genome of a cell from which a transgenic animal develops and which remains in the genome of the mature animal in one or more of its cell types or tissues. Such animals are useful for studying the function of a variant protein *in vivo*, and identifying and evaluating modulators of variant protein activity.
20 Other examples of transgenic animals include, but are not limited to, non-human primates, sheep, dogs, cows, goats, chickens, and amphibians. Transgenic non-human mammals such as cows and goats can be used to produce variant proteins which can be secreted in the animal's milk and then recovered.

A transgenic animal can be produced by introducing a SNP-containing nucleic acid
25 molecule into the male pronuclei of a fertilized oocyte, e.g., by microinjection or retroviral infection, and allowing the oocyte to develop in a pseudopregnant female foster animal. Any nucleic acid molecules that contain one or more SNPs of the present invention can potentially be introduced as a transgene into the genome of a non-human animal.

Any of the regulatory or other sequences useful in expression vectors can form part of the
30 transgenic sequence. This includes intronic sequences and polyadenylation signals, if not already included. A tissue-specific regulatory sequence(s) can be operably linked to the transgene to direct expression of the variant protein in particular cells or tissues.

Methods for generating transgenic animals via embryo manipulation and microinjection, particularly animals such as mice, have become conventional in the art and are described in, for

example, U.S. Patent Nos. 4,736,866 and 4,870,009, both by Leder *et al.*, U.S. Patent No. 4,873,191 by Wagner *et al.*, and in B. Hogan, *Manipulating the Mouse Embryo*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1986). Similar methods are used for production of other transgenic animals. A transgenic founder animal can be identified based upon the presence of the transgene in its genome and/or expression of transgenic mRNA in tissues or cells of the animals. A transgenic founder animal can then be used to breed additional animals carrying the transgene. Moreover, transgenic animals carrying a transgene can further be bred to other transgenic animals carrying other transgenes. A transgenic animal also includes a non-human animal in which the entire animal or tissues in the animal have been produced using the homologously recombinant host cells described herein.

In another embodiment, transgenic non-human animals can be produced which contain selected systems that allow for regulated expression of the transgene. One example of such a system is the cre/loxP recombinase system of bacteriophage P1 (Lakso *et al.* *PNAS* 89:6232-6236 [1992]). Another example of a recombinase system is the FLP recombinase system of *S. cerevisiae* (O'Gorman *et al.*, *Science* 251:1351-1355 [1991]). If a cre/loxP recombinase system is used to regulate expression of the transgene, animals containing transgenes encoding both the Cre recombinase and a selected protein are generally needed. Such animals can be provided through the construction of "double" transgenic animals, e.g., by mating two transgenic animals, one containing a transgene encoding a selected variant protein and the other containing a transgene encoding a recombinase.

Clones of the non-human transgenic animals described herein can also be produced according to the methods described in, for example, I. Wilmut *et al.*, *Nature* 385:810-813 (1997) and PCT International Publication Nos. WO 97/07668 and WO 97/07669. In brief, a cell (e.g., a somatic cell) from the transgenic animal can be isolated and induced to exit the growth cycle and enter G₀ phase. The quiescent cell can then be fused, e.g., through the use of electrical pulses, to an enucleated oocyte from an animal of the same species from which the quiescent cell is isolated. The reconstructed oocyte is then cultured such that it develops to morula or blastocyst and then transferred to pseudopregnant female foster animal. The offspring born of this female foster animal will be a clone of the animal from which the cell (e.g., a somatic cell) is isolated.

Transgenic animals containing recombinant cells that express the variant proteins described herein are useful for conducting the assays described herein in an *in vivo* context. Accordingly, the various physiological factors that are present *in vivo* and that could influence ligand or substrate binding, variant protein activation, signal transduction, or other processes or interactions, may not be evident from *in vitro* cell-free or cell-based assays. Thus, non-human transgenic animals of the

present invention may be used to assay *in vivo* variant protein function as well as the activities of a therapeutic agent or compound that modulates variant protein function/activity or expression. Such animals are also suitable for assessing the effects of null mutations (i.e., mutations that substantially or completely eliminate one or more variant protein functions).

For further information regarding transgenic animals, see Houdebine, "Antibody manufacture in transgenic animals and comparisons with other systems," *Curr. Opin. Biotechnol.* 13(6):625-9 (Dec. 2002); Petters *et al.*, "Transgenic animals as models for human disease," *Transgenic Res.* 9(4-5):347-51 (2000); discussion 345-6; Wolf *et al.*, "Use of transgenic animals in understanding molecular mechanisms of toxicity," *J. Pharm. Pharmacol.* 50(6):567-74 (Jun. 1998); Echelard, "Recombinant protein production in transgenic animals," *Curr. Opin. Biotechnol.* 7(5):536-40 (Oct. 1996); Houdebine, "Transgenic animal bioreactors," *Transgenic Res.* 9(4-5):305-20 (2000); Purity *et al.*, "Embryonic stem cells, creating transgenic animals," *Methods Cell Biol.* 57:279-93 (1998); and Robl *et al.*, "Artificial chromosome vectors and expression of complex proteins in transgenic animals," *Theriogenology* 59(1):107-13 (Jan. 1 2003).

EXAMPLE: STATISTICAL ANALYSIS OF SNPs ASSOCIATED WITH ALZHEIMER'S DISEASE

The following example is offered to illustrate, but not limit the claimed invention.

Results from previous whole-genome screens have identified several markers that show significant association with late-onset Alzheimer's disease (LOAD) (M.A. Pericak-Vance *et al.*, *Exp. Gerontol.* 35[9-10]:1343-1352 [2000]; D. Curtis *et al.*, *Ann. Hum. Genet.* 65[Pt. 5]:473-481 [2001]; A. Myers *et al.*, *Am. J. Med. Genet.* 114[2]:235-244 [2002]; J.M. Olson *et al.*, *Am. J. Hum. Genet.* 71[1]:154-161 [2002]; D. Blacker *et al.*, *Hum. Mol. Genet.* 12[1]:23-32 [2003]).

Introduction

Late-onset Alzheimer disease (LOAD), is an age-related neurodegenerative disease primarily affecting learning and memory. Two recently published studies suggest that the heritability of AD (proportion of phenotypic variance in the population attributable to genes) is in the range of 60% - 80%, with the balance of variance attributable to unique (typically adult) environmental influences [Gatz *et al.*, 2005, *Neurobiol Aging*, 26: 439-447; Pedersen *et al.*, 2004, *Ann Neurol*, 55: 180-185]. The only known genetic risk factor is the apolipoprotein E (*APOE*) gene. The *APOE4* allele is highly enriched in AD cases compared to non-demented individuals, whereas the *APOE2* allele is under-represented in cases [Strittmatter *et al.*, 1993, *Proc Natl Acad Sci U S A*, 90: 1977-1981]. However, only 50% of AD cases carry an *APOE4* allele and several

additional putative LOAD susceptibility loci have been identified through linkage studies [Pastor and Goate, 2004, Curr Psychiatry Rep, 6: 125-133].

In addition to linkage studies many association studies have examined specific candidate genes as risk factors for disease, often through single nucleotide polymorphism (SNP)-based disease association studies (<http://www.alzgene.org>). Early association studies generally tested SNPs in one or a small number of selected biological and/or positional candidate genes at a time, but recent reports, including those from our groups, have expanded the search by testing a large number of SNPs from within linkage regions and/or whole chromosomes [Grupe et al., 2006, Am J Hum Genet, 78: 78-88; Kuwano et al., 2006, Hum Mol Genet, 15: 2170-2182; Li et al., 2004, Proc Natl Acad Sci U S A, 101: 15688-15693]. These recent studies have led to the identification of various novel candidate genes that may contribute to the etiology of LOAD, including *DAPK1* on chromosome 9 [Li et al., 2006, Hum Mol Genet, 15: 2560-2568], *GAPD* and its homolog on chromosome 12 [Li et al., 2004, Proc Natl Acad Sci U S A, 101: 15688-15693], and a RPS3A homologue and several others on chromosome 10 [Grupe et al., 2006, Am J Hum Genet, 78: 78-88; Kuwano et al., 2006, Hum Mol Genet]. Both, *DAPK1* and *GAPD*, are regulators of neuronal apoptosis and RPS3A is also involved in the apoptotic process. These genetic variants have been replicated to various degrees in several of the tested case-control sample sets but show relatively small to moderate effect sizes. As such, definitive conclusions concerning of their potential involvement in conferring susceptibility will require further studies and meta-analyses.

In the absence of convincing evidence that explains most of the genetic contribution to LOAD risk, a more thorough investigation is warranted. Perhaps the best and most comprehensive approach may be a SNP-based, genome-wide association study. Fortunately, such a study design has become feasible and has been applied successfully in other studies in the search for risk factors of common complex genetic diseases [Klein et al., 2005, Science, 308: 385-389; Smyth et al., 2006, Nat Genet, 38: 617-619]. Two strategies that utilize current knowledge and technology are commonly used: The linkage disequilibrium (LD) approach interrogates independent position specific markers that tag a large number of haplotype blocks to narrow the location of susceptibility alleles. The other study design directly tests potentially causative mutations, which requires making assumptions about the effect a mutation may have on gene or protein function. The latter study design may provide a focused strategy to directly identify a disease relevant mutation. Both approaches test tens to hundreds of thousands of SNPs, making it hard to distinguish genuine from spurious associations unless a strongly associated genetic marker stands out from the many expected false-positives. Unfortunately,

many of the genetic effects may be small, at least for those genetic factors that fit the common disease – common variant hypothesis, making it difficult to unambiguously demonstrate association. Results from previous whole-genome screens have identified several markers that show significant association with late-onset Alzheimer's disease (LOAD) (M.A. Pericak-Vance *et al.*, *Exp. Gerontol.* 35[9-10]:1343-1352 [2000]; D. Curtis *et al.*, *Ann. Hum. Genet.* 65[Pt. 5]:473-481 [2001]; A. Myers *et al.*, *Am. J. Med. Genet.* 114[2]:235-244 [2002]; J.M. Olson *et al.*, *Am. J. Hum. Genet.* 71[1]:154-161 [2002]; D. Blacker *et al.*, *Hum. Mol. Genet.* 12[1]:23-32 [2003]).

In an effort to uncover more SNPs associated with the development of AD, a large-scale, genome-wide SNP-based association study of LOAD was undertaken as described below, where about 20,000 putative functional variants were screened across the genome using a large, well-characterized case-control sample set collected in the UK and followed up on a subset of markers in 3 sample sets from the UK and US. The results here show that markers in LD with *APOE* were successfully identified from this screen and that other novel markers are significantly associated with LOAD.

15 **Methods**

Clinical Samples

Five LOAD case-control sample sets were used in this study. Demographic information for these samples is shown in Table 6. Cases had a minimum age at disease onset (AAO) of 60 years and a diagnosis of probable or definite Alzheimer's type according to National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) (G. McKhann *et al.*, "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease," *Neurology* 34, 939-44 (1984)), while controls were ascertained at the age of 65 years or older and screened for evidence of dementia (MMSE $\geq$ 28, Clinical Dementia Rating = 0, or full neurological exam). All individuals were of Caucasian descent. All sample sets show the expected *APOE4* distribution in cases and controls.

The five sample sets, designated as the WU, SD, UK1, UK2 and UK3 sample sets respectively, are the WU sample set, obtained through the Washington University Alzheimer's Disease Research Center (ADRC) patient registry, the SD sample set, obtained from the ADRC of the University of California, San Diego, and the UK1, UK2 and UK3 sample sets, obtained from the Medical Research Council Late Onset AD Genetic Resource that included those from Cardiff University, Cambridge University, King's College London, and Trinity College Dublin. These sample sets have been used in our other recent studies and detailed phenotypic information

can be found therein (Li, Y. et al. "Association of late-onset Alzheimer's disease with genetic variation in multiple members of the GAPD gene family," *Proc Natl Acad Sci U S A* 101, 15688-93 (2004); Li, Y. et al., "Genetic association of the APP binding protein 2 gene (APBB2) with late onset Alzheimer disease," *Hum Mutat* 25, 270-277 (2005), Li et al., *Hum Mol Genet*, 15: 2560-2568, (2006)). They all show an expected *APOE* ϵ 4-genotype distribution.

Table 6. Sample set characteristics

Sample Set	Sample Size(Cs/Co)	AAO/AEE (Cs/Co)	ApoE4 Allele % (Cs/Co)	Female% (Cs/Co)
UK1	380/396	75.7 $\pm$ 7.1/76.6 $\pm$ 6.2	36.9/13.3	76.6/75
UK2	309/349	76.3 $\pm$ 7.0/76.7 $\pm$ 5.6	36.7/11.8	72.8/56.7
UK3	503/643	75.1 $\pm$ 6.5/76.1 $\pm$ 6.4	37.1/15.2	64.0/59.0
WU	376/344	76.3 $\pm$ 6.9/77.6 $\pm$ 7.5	33.5/12.4	62.7/61.6
SD	240/330	72.0 $\pm$ 6.2/79.3 $\pm$ 7.2	38.9/11.8	45.4/63.3

Cs/Co: Cases/Controls

AAO: average age at disease onset for cases ($\pm$ Standard Deviation)

AAE: average age at exam for controls ($\pm$ Standard Deviation)

Cases AAO $\geq$ 60 yrs; Controls AAE $\geq$ 65 yrs

SNP Selection

The majority of SNPs that were chosen for the scan are predicted to modulate gene or protein function (Table 7); missense and nonsense variants account for 70% of the analyzed markers and SNPs within putative transcription factor binding sites account for 13%. Allele frequencies of 19,863 unique autosomal markers and 390 unique X-chromosome markers were determined in DNA pools of the discovery sample set (UK1). Markers with low minor allele frequencies ($<0.02\%$ in case and control pools) and those for which assay quality was poor (significantly inconsistent allele frequencies among three orthogonal pools [see below]), were not further analyzed, resulting in 17,343 SNPs analyzed for association with LOAD. The allele frequencies of these markers are fairly evenly distributed (Figure 1A) based on the data in control pools. The number of SNPs analyzed on each chromosome is given in Figure 1B. The SNPs map to 11,221 unique genes based on the Celera annotation (Assembly R27). The distribution of these genes on specific chromosomes is given in Figure 1C. Although SNP coverage varies from chromosome to chromosome, the relative gene coverage per chromosome is similar with an average of 35.5% of genes containing at least one SNP (Figure 1D).

Table 7. SNP Categories

SNP_OVERALL_TYPE	Count Of SNP_OVERALL_TYPE
Missense Mutation	12066
Nonsense Mutation	228

Acceptor splice site	42
Donor splice site	58
ESS	232
UTR3	581
UTR5	391
putative UTR5	22
Transcription Factor Binding Site	2318
Silent Mutation	468
Human-Mouse Syntenic region	32
Intron	559
Intergenic/Unknown	346
SUM	17343

Allele Frequency Determination in Pools and Statistical Analysis

From each of the UK1 and WU samples, 12 DNA pools were constructed. Overlapping pools for each sample were constructed to contain 1) all male cases 2) all female cases 3) all cases with AAO < 75 years, 4) all cases with AAO ≥ 75 years 5) all cases who carry 1 or more copies of the *APOE* 4 allele 6) all cases who do not carry an *APOE* 4 allele. The 6 corresponding pools of control subjects were also constructed with the exception that control subjects were stratified by age at examination rather than AAO. The numbers of individuals in each pool are given in Table 7. The allele frequencies of each of the 12 pools were determined from the average of two pooled genotyping assays based on SYBR green real time PCR as described by Germer et al. [Germer et al., 2000, Genome Res, 10: 258-266]. Allele numbers (rounded to the nearest integer) for each pool were estimated by multiplying the chromosome count in the pool by the estimated allele frequency. Our design resulted in a three-fold redundancy (orthogonal pools), that is there are three ways of reconstructing the allele frequencies in the complete sample based on combining the gender pools, the AAO pools or the *APOE* status pools. Based upon the average from each of the three methods, allelic odds ratios were estimated for each marker as was a statistical test for allelic association (Fisher's exact test). It should be noted that the *P*-value from this test is expected to have an inflated type-I error rate due to the measurement error generated when estimating allele frequencies from the replicate pools used in this study and from the technical error associated with pool construction. Analyzing the average of the three orthogonal pools helps to reduce the inflation by averaging out the measurement and technical error. In addition, χ^2 -tests were performed to calculate an assay quality score separately for cases and controls to test the hypothesis that the 3 replicate pools have equal allele frequency. For SNPs residing on the X chromosome, allele counts were estimated assuming females contribute 2 alleles and males 1 allele and statistical tests were modified accordingly. The quality score *p*-value for SNPs on the X chromosome is based on a chi-square test, which tests the hypothesis

that male and female pools have equal allele frequency since replicate pools do not exist for these SNPs. Using sex as a stratifying variable, the Mantel-Haenzsel odds ratio and *P*-value were used to estimate and test for the overall association of SNPs on the X chromosome with disease status. Mantel-Haenzsel odds ratios and *P*-values were also used to estimate and test for the overall association of SNPs when combining results of multiple sample sets into a meta-analysis.

Table 8. DNA Pool Composition

Sample Set	Pool	Sample Size (Cs/Co)
UK1		380/396
	ApoE4+	229/96
	ApoE4-	149/297
	GE75	218/246
	LT75	162/150
	Male	89/99
	Female	291/297
WU		376/344
	ApoE4+	217/81
	ApoE4-	159/262
	GE75	208/200
	LT75	168/144
	Male	140/132
	Female	236/212

APOE4+: carriers of one or more APOE4 alleles

APOE4-: no APOE4 allele is present

GE75: Age at onset (age at exam) is 75years of age or more for cases (controls)

LT75: Age at onset (age at exam) is less than 75years of age for cases (controls)

Cs: cases

Co: controls

Samples without APOE genotype information were excluded from the APOE pools
(UK1: 2 Cs, 3 Co; WU: 1 Co)

Replication Criteria for DNA Pools

The following criteria were applied to identify autosomal markers that were then followed up in DNA pools of the WU sample set: a non-significant assay quality score *P*-value ($P \geq 0.01$) in case pools and control pools and identical risk alleles in all three orthogonal pools (i.e. combined APOE, sex, AAO) together with one of the following: (i) a significant allelic *P*-value ($P < 0.05$) and an odds ratio (OR) greater than 1.25 or smaller than 0.8 (ii) an allelic *P*-value of < 0.075 , an OR > 1.15 or OR < 0.87 , and minor allele frequency $> 5\%$ in either case or control pools (iii) for markers that map to a biological candidate gene, a greater than 5% minor allele frequency in either case or control pools, a *P*-value < 0.1 and an OR > 1.25 or OR < 0.8 . This

weighting scheme was introduced to reduce the number of false positive markers that may due to measurement or technical errors for low frequency markers between 2% and 5% allele frequency.

X-chromosome assays were followed up in WU pools when the following criteria were met: assay quality score P -value ≥ 0.01 in case and control pools, identical risk allele in male
 5 pools and female pools and either (i) a significant allelic Mantel-Haenszel P -value ($P < 0.05$) and an OR > 1.25 or OR < 0.8 or (ii) minor allele frequency of greater than 5% in either cases or controls, a Mantel Haenszel P -value of $P < 0.075$ and a Mantel-Haenszel OR > 1.15 or < 0.87 .

The threshold criteria in the WU DNA pools to select 119 markers for individual genotyping in all four sample sets (UK1, WU, SD, UK2) were allelic P -value $P < 0.15$ (Fisher's
 10 exact test for autosomal markers, Mantel Haaenszel test for X-chromosome markers) and identical risk alleles in the UK1 and WU pools.

Individual Genotyping and Statistical Analysis

Individual genotyping and statistical analysis were performed as previously described [Grupe et al., 2006, Am J Hum Genet, 78: 78-88; Li et al., 2005, Hum Mutat, 25: 270-277; Li et
 15 al., 2004, Proc Natl Acad Sci U S A, 101: 15688-15693]. Accuracy of our individual genotyping was greater than 99% [Li et al., 2004, Proc Natl Acad Sci U S A, 101: 15688-15693].

Results

To identify novel genetic variants that predispose to LOAD, SNP-based association
 20 studies were performed in up to 5 independently collected case control sample sets, totaling 1808 cases and 2062 controls. Our overall strategy was to first generate a short-list of candidate variants that meet the follow up criteria, through testing a large number of SNPs in DNA pools of one sample set (UK1, Figure 2A). Those SNPs that met previously defined significance criteria were typed in DNA pools from a second set of cases and controls (WU) (hypothesis generation
 25 study) (Figure 2B). Markers that satisfied threshold criteria from these first two rounds of pooled genotyping were individually genotyped in 4 sets of cases and controls (UK1, WU, SD, and UK2). Finally, markers showing the most evidence of association ($p < 0.01$) were genotyped in the fifth sample (UK3). The effect size and significance of the association with LOAD was calculated for each sample set and for all sample sets combined.

30 The markers tested were primarily located within genes or promoter regions and targeted putative functional mutations (e.g. non-synonymous, putative transcription factor binding site, 5'/3'UTR, mouse-man conserved sequence) (Table 7). By prioritizing putative functional mutations for the scan, there was an increased chance of finding disease-associated variants that

have a direct biological effect and therefore may show a stronger effect than markers that are in LD with the causal mutation.

Specifically, the hypothesis generation study was carried out by first genotyping ~20,000 SNPs in pooled cases and controls of the UK1 sample set. Excluding markers of low frequency (minor allele frequency of < 2% in both cases and controls) and inconsistent allele frequencies between orthogonal pools (assay quality $P < 0.01$), 17,343 SNPs were tested, located in 11,221 unique genes, for association with disease risk. Since small measurement errors in allele frequencies determined from DNA pools can result in false negative as well as false positive findings, a relaxed significance threshold (generally, $P < 0.075$, see Methods above) was adopted to select markers for further testing in the WU sample set and relied on replication in additional sample sets to help reduce false positive findings. The significance threshold in pools was further relaxed for markers in pre-defined biological candidate genes ($P < 0.1$), which were selected based on our understanding of AD biology. By applying these thresholds for minor allele frequency, allelic P -value, and OR, a total of 1544 markers were identified from the UK1 study for subsequent analysis in the WU pools. Of these, 119 markers met our threshold criteria for the WU sample set (allelic $P < 0.15$, same risk allele as UK1).

To validate the association of these markers with LOAD, all 119 markers were individually genotyped in the SD sample set and the UK2 sample set. Because spurious associations may arise from measurement and technical errors in the pool derived allele frequency estimates, these 119 markers were also individually genotyped in the UK1 and WU sample sets and required $P < 0.05$ for the meta-analysis of UK1 and WU to assess the marker further (90 SNPs). A meta-analysis combining these four sample sets was then performed and the 19 most significant markers ($P_{\text{meta}} < 0.005$) were genotyped in a fifth sample set (UK3), consisting of 503 LOAD cases and 643 matched controls. The results of a meta-analysis of all five sample sets on these 19 markers are presented in Table 8a, along with P -values for each sample separately. The P -values given in Table 9 for the meta-analysis were estimated by simulation, in order to account for the multistage study design. Of these 19 markers, in the 5 control samples, Hardy-Weinberg equilibrium was violated at a $p < 0.05$ level in one test out of 94 (hCV15746640 in UK1), which is less than would be expected by chance.

Table 9. The significant markers from a meta-analysis of 5 LOAD case control sample sets

Accession No.	Gene Symbol	Allelic p-value				Odds Ratio (95% CI)				Risk Allele	Allele	
		UK1	WU	SD	UK2	UK3	Full Sample	Full Sample	Risk Allele		1	2
1	rs405509	APOE	0.000081	0.0035	0.00010	0.000300	0.002	<1.0E-8	1.43 (1.3-1.57)	T	G	T
2	rs157581	TOMM40	6.86E-14	3.38E-14	5.2E-22	2.939E-15	4.82E-24	<1.0E-8	2.73 (2.46-3.05)	C	C	T
3	rs3745833	GALP	0.008	0.036	0.512	0.261	0.061	0.000050	1.2 (1.09-1.32)	G	C	G
4	rs1554948	TNK1	0.130	0.0090	0.699	0.054	0.021	0.000063	1.19 (1.08-1.3)	T	A	T
5	rs1132899	APOC2	0.137	0.0075	0.056	0.013	0.975	0.000076	1.19 (1.08-1.3)	C	C	T
6	rs11622883		0.144	0.010	0.979	0.033	0.091	0.000094	1.19 (1.08-1.3)	T	A	T
7	rs8192708	PCK1	0.0033	0.026	0.396	0.791	0.051	0.000099	1.29 (1.12-1.49)	G	A	G
8	rs505058	LMNA	0.0119	0.150	0.357	0.171	0.261	0.00020	1.35 (1.13-1.63)	C	C	T
9	rs3800324	PGBD1	0.204	0.026	0.502	0.317	0.146	0.00032	1.43 (1.13-1.8)	A	A	G
10	rs6907175	LOC651924	0.017	0.304	0.042	0.957	n/a	0.00047	1.16 (1.04-1.3)	A	G	A
11	rs1859849		0.0047	0.168	0.776	0.031	0.575	0.00058	1.15 (1.03-1.28)	C	C	T
12	hCV22274641	THEM5	0.000031	0.109	0.647	0.169	0.467	0.00061	1.24 (1.05-1.47)	T	A	T
13	rs2074877	MYH13	0.017	0.029	0.664	0.931	0.742	0.00072	1.12 (1.02-1.23)	T	C	T
14	hCV15746640	CTSS	0.0008	0.031	0.595	0.116	0.293	0.00078	1.21 (1.02-1.42)	A	A	G
15	rs444013	UBD	0.010	0.042	0.299	0.741	0.845	0.00079	1.11 (1.01-1.22)	C	G	C
16	rs9608099	BCR	0.261	0.0018	0.651	0.134	0.197	0.00097	1.09 (0.99-1.21)	C	C	T
17	rs2882676	AGC1	0.076	0.033	0.358	0.375	0.05	0.0010	1.07 (0.97-1.18)	A	A	C
18	rs13022344	TRAK2	0.0051	0.0077	0.821	0.628	0.015	0.0010	1.07 (0.97-1.18)	C	C	T
19	rs11016976	EBF3	0.050	0.024	0.294	0.966	0.033	0.0010	1.08 (0.95-1.24)	C	C	T

Of the identified significant markers in Table 9 (SNP genomic position and function are in Table 10), 3 are located near the *APOE* locus on chromosome 19 (rs157581, rs405509, rs1132899). The LD of these 3 markers with the *APOE* ϵ 4 and ϵ 2/3 markers is between 0.92 and 0.088 (D') and 0.66 and 0.001 (r^2 , Table 11). To determine whether these markers increased disease risk independently of *APOE* ϵ 2/3/4, *P*-values were calculated adjusted for either rs429358 (ϵ 4-allele) or rs7412 (ϵ 2/3-allele). Although these 3 markers show a wide spread of LD with the *APOE* alleles, none of the 3 markers increases LOAD risk independently of rs429358 and rs7412 (See Table 12). The sixteen other markers had effect sizes ranging between 1.07 and 1.45 in the multi-sample meta-analysis and several map to known linkage regions, biological candidate genes and novel genes.

Table 10. Genomic position and function of the significant markers from a meta-analysis of five LOAD case-control sample sets

<i>Accession No.</i>	<i>SNP TYPE</i>	<i>Gene Symbol</i>	<i>Gene Name</i>	<i>Chr</i>	<i>Position (Mbp)</i>
hCV15746640	Missense (Celera)	CTSS	cathepsin S	1	147.5
hCV22274641	Missense (Celera)	THEM5	thioesterase superfamily member 5	1	147.7
rs505058	Silent	LMNA	lamin A/C	1	152.9
rs13022344	Missense	TRAK2	trafficking protein	2	202.0
rs3800324	Missense	PGBD1	piggyBac transposable element derived 1	6	28.37
rs444013	UTR3	UBD	ubiquitin D	6	29.63
rs6907175	UTR3 (Celera)	LOC651924	homolog to ubiquitin-like 1-activating enzyme E1B (SUMO-1-activating enzyme subunit 2)	6	142.4
rs1859849	Intergenic			7	25.45
rs11016976	TFBS (Celera)	EBF3	early B-cell factor 3	10	131.5
rs11622883	Intergenic		near EST DA712292 and SERPINA13	14	94.23
rs2882676	Missense	AGC1	aggreCAN 1	15	87.2
rs1554948	Silent (Celera)	TNK1	tyrosine kinase	17	7.23
rs2074877	Missense	MYH13	myosin	17	10.16
rs157581	TFBS (Celera)	TOMM40	translocase of outer mitochondrial membrane 40 homolog	19	50.05
rs405509	Intergenic	APOE	apolipoprotein E	19	50.1
rs1132899	Missense	APOC2	apolipoprotein C-II	19	50.14
rs3745833	Missense	GALP	galanin-like peptide precursor	19	61.35
rs8192708	Missense	PCK1	phosphoenolpyruvate carboxykinase 1 (soluble)	20	55.57
rs9608099	UTR5 (Celera)	BCR	breakpoint cluster region	22	21.96

Table 11. LD Profile of Markers near *APOE*

		D'				
		rs157581	rs405509	Rs429358	rs7412	rs1132899
r ²	rs157581		0.36	0.92	0.088	0.11
	rs405509	0.048		0.6	0.92	0.007
	rs429358	0.66	0.11		1	0.19
	rs7412	0.001	0.062	0.019		0.45
	rs1132899	0.004	0	0.011	0.013	

Legend for Table 11: The upper right triangle shows D' values, the lower left triangle shows r² values

5 **Table 12. Significance of Markers near the APOE locus after correction by other markers**

		Correction SNP				
		rs157581	rs405509	rs429358	rs7412	rs1132899
Evaluated SNP	rs157581	7.1E-59	1.3E-46	8.5E-01	4.8E-55	1.6E-51
	rs405509	1.3E-03	1.9E-13	7.4E-01	9.8E-08	4.4E-12
	rs429358	4.5E-35	2.2E-74	2.6E-97	2.9E-79	1.6E-79
	rs7412	2.6E-13	2.1E-06	2.3E-04	4.1E-12	3.9E-11
	rs1132899	4.9E-03	1.9E-04	1.3E-01	2.8E-05	1.7E-05

rs429358 determines whether the epsilon 4 allele of APOE is present or absent

rs7412 determines whether the epsilon 3 or the epsilon 2 allele of APOE is present

The diagonal contains the single marker P-values

Note: Results are for UK1,WU,SD,UK2 sample sets combined

Genetic association of the SNPs with LOAD observed in this study specifically implicates the genes and regions of the chromosome in which these SNPs reside are potential contributor to disease susceptibility. A genotyping kit for the detection of various SNPs may be designed for the diagnosis of AD and treatment response in the management of AD and other neurological pathologies. Novel medicines including, but not limited to, small molecules, proteins, protein fragments or peptides, antibodies, or nucleotide acids, may be designed to target the risk factors for the treatment of AD and other neurological pathologies.

15

Example Two: Additional LD SNPs associated with AD

Another investigation was conducted to identify SNP markers in linkage disequilibrium (LD) with SNPs which have been found to be associated with AD as shown in Tables 1-2. The methodology is described earlier in the instant application. To summarize briefly, the power threshold (*T*) was set at an appropriate level such as 51%, or 70%, for detecting disease association using LD markers. This power threshold is based on equation (31) above, which incorporates allele frequency data from previous disease association studies, the predicted error

20

rate for not detecting truly disease-associated markers, and a significance level of 0.05. Using this power calculation and the sample size, for each interrogated SNP, a threshold level of LD, or r^2 value, was derived (r_T^2 , equations (32) and (33)). The threshold value r_T^2 is the minimum value of linkage disequilibrium between the interrogated SNP and its LD SNPs possible such that the non-interrogated SNP still retains a power greater or equal to T for detecting disease-association.

Based on the above methodology, LD SNPs were found for all interrogated SNPs shown in Tables 1- 2. These LD SNPs are listed in Table 13; each is associated with its respective Interrogated SNP. Also shown are hCV numbers and the corresponding public SNP IDs (rs numbers) for the interrogated and LD SNPs, the threshold r_T^2 value and the power used to determine this, and the r^2 value of linkage disequilibrium between the interrogated SNP and its matching LD SNP. As an example in Table 13, the interrogated, AD-associated SNP hCV11827509 was calculated to be in LD with hCV11170193 at a r^2 value of 1, based on a 51% power calculation, thus making the latter a marker associated with AD as well. Another interrogated, AD-associated SNP hCV8717889 was calculated to be in LD with hCV1247823 at a r^2 value of 0.57362, based on a 51% power calculation, thus making hCV1247823 a marker associated with AD as well because the threshold r_T^2 value is only 0.287.

Table 13: LD SNPs

Interrogated SNP	Interrogated rs	LD SNP	LD SNP rs	Power	Threshold r_T^2	r^2
hCV11827509	rs1859849	hCV11170193	rs2058024	0.51	0.977	1
hCV11827509	rs1859849	hCV11170195	rs1859848	0.51	0.977	1
hCV11827509	rs1859849	hCV15738	rs740345	0.51	0.977	1
hCV11827509	rs1859849	hCV15854511	rs2214976	0.51	0.977	1
hCV11827509	rs1859849	hCV16151469	rs2190532	0.51	0.977	1
hCV11827509	rs1859849	hCV2576008	rs10807830	0.51	0.977	1
hCV11827509	rs1859849	hCV2576009	rs11762312	0.51	0.977	1
hCV11827509	rs1859849	hCV2576011	rs758012	0.51	0.977	1
hCV11827509	rs1859849	hCV2576021	rs1544480	0.51	0.977	1
hCV11827509	rs1859849	hCV2576027	rs886836	0.51	0.977	1
hCV11827509	rs1859849	hCV2576036	rs10951092	0.51	0.977	1
hCV11827509	rs1859849	hCV2576041	rs10951098	0.51	0.977	1
hCV11827509	rs1859849	hCV2576042	rs2391151	0.51	0.977	1
hCV11827509	rs1859849	hCV2576044	rs2391152	0.51	0.977	1
hCV11827509	rs1859849	hCV2576045	rs6953294	0.51	0.977	1
hCV11827509	rs1859849	hCV27263228	rs2893212	0.51	0.977	1
hCV11827509	rs1859849	hCV27263241	rs2158932	0.51	0.977	1
hCV11827509	rs1859849	hCV29401758	rs6942970	0.51	0.977	1
hCV11827509	rs1859849	hCV32157908	rs11973825	0.51	0.977	1
hCV11827509	rs1859849	hCV32157909	rs11489283	0.51	0.977	1

hCV11827509	rs1859849	hCV32399625	rs4141279	0.51	0.977	1
hCV11827509	rs1859849	hCV345686	rs9942690	0.51	0.977	1
hCV11827509	rs1859849	hCV346553	rs4722492	0.51	0.977	1
hCV11827509	rs1859849	hCV346555	rs11772927	0.51	0.977	1
hCV11827509	rs1859849	hCV346556	rs11762724	0.51	0.977	1
hCV11827509	rs1859849	hCV349359	rs11983371	0.51	0.977	1
hCV11827509	rs1859849	hCV7600851	rs1106876	0.51	0.977	1
hCV11827509	rs1859849	hCV7600852	rs1558077	0.51	0.977	1
hCV11827509	rs1859849	hCV7600855	rs1558075	0.51	0.977	1
hCV11827509	rs1859849	hCV9949982	rs12667142	0.51	0.977	1
hCV11827509	rs1859849	hDV70903693	rs17152187	0.51	0.977	1
hCV11827509	rs1859849	hDV70903706	rs17152208	0.51	0.977	1
hCV11827509	rs1859849	hDV70903726	rs17152235	0.51	0.977	1
hCV3084827	rs157581	hCV27899309	rs4801914	0.51	0.01047	0.01425
hCV3084827	rs157581	hCV29207239	rs8110693	0.51	0.01047	0.0177
hCV3084827	rs157581	hCV2931396	rs2560983	0.51	0.01047	0.01338
hCV3084827	rs157581	hCV31444359	rs7252479	0.51	0.01047	0.02632
hCV3084827	rs157581	hDV70782148	rs16983943	0.51	0.01047	0.03065
hCV3084827	rs157581	hDV71600369	rs16983796	0.51	0.01047	0.02174
hCV8717889	rs1554948	hCV1247820	rs12948090	0.51	0.287	1
hCV8717889	rs1554948	hCV1247823	rs2269762	0.51	0.287	0.57362
hCV8717889	rs1554948	hCV1247824	rs1554947	0.51	0.287	0.63112
hCV8717889	rs1554948	hCV1477271	rs11655156	0.51	0.287	0.40191
hCV8717889	rs1554948	hCV26960061	rs4796409	0.51	0.287	0.5414
hCV8717889	rs1554948	hCV26960066	rs4325594	0.51	0.287	0.40191
hCV8717889	rs1554948	hCV30677672	rs7219773	0.51	0.287	0.93522
hCV905013	rs405509	hCV3066768	rs162129	0.51	0.069	0.07287
hCV905013	rs405509	hCV32392600	rs544769	0.51	0.069	0.08005

Tables 1-5 are attached hereinafter.

All publications and patents cited in this specification are herein incorporated by reference in their entirety. Various modifications and variations of the described compositions, methods and systems of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments and certain working examples, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the above-described modes for carrying out the invention that are obvious to those skilled in the field of molecular biology, genetics and related fields are intended to be within the scope of the following claims.

Table 1

	Gene Number:	1
5	Celera Gene:	hCG1640936 - 67000129056428
	Celera Transcript:	hCT1641063 - 67000129056429
	Public Transcript Accession:	NM_003802
	Celera Protein:	hCP1602520 - 197000069434185
	Public Protein Accession:	NP_003793
10	Gene Symbol:	MYH13
	Protein Name:	myosin, heavy polypeptide 13, skeletal muscle
	Celera Genomic Axis:	GA_x5YUV32W2JD (5134352..5226503)
	Chromosome:	17
	OMIM NUMBER:	603487
15	OMIM Information:	
	Transcript Sequence (SEQ ID NO: 1):	
	Protein Sequence (SEQ ID NO: 49):	
20	SNP Information	
	Context (SEQ ID NO: 97):	
25	TCTTGAGGGTTCCTTAGAGCAGGAGAAGAACTGCGGGCGGACTTGGAAGGGCGA	
	AGAGGAAGCTGGAAGGAGATCTGAAAATGTCCCAGGAATCCATT	
	R	
	TGGATCTAGAAAATGACAAGCAGCAAATAGAAGAGAAATTGAAAAAGAAGGAGTTT	
	GAACTCAGTCAGTTACAAGCCAAAATAGATGACGAACAAGTCCA	
	Celera SNP ID:	hCV2179030
30	Public SNP ID:	rs2074877
	SNP in Transcript Sequence SEQ ID NO: 1	
	SNP Position Transcript: 3301	
	SNP Source: Applera	
	Population (Allele,Count): caucasian (G,14 A,2) african american (G,2 A,4) total (G,16 A,6)	
35	SNP Type:	Missense Mutation
	Protein Coding:	SEQ ID NO: 49, at position 1071,(M,ATG) (V,GTG)
	SNP Source:	dbSNP; Celera; HapMap; HGBASE
	Population (Allele,Count): caucasian (A,46 G,74)	
	SNP Type:	Missense Mutation
40	Protein Coding:	SEQ ID NO: 49, at position 1071,(M,ATG) (V,GTG)
	Gene Number:	2
	Celera Gene:	hCG1641280 - 65000099415023
	Celera Transcript:	hCT1641407 - 65000099415024
45	Public Transcript Accession:	NM_033106
	Celera Protein:	hCP1631004 - 197000069408245
	Public Protein Accession:	NP_149097
	Gene Symbol:	GALP
	Protein Name:	galanin-like peptide precursor
50	Celera Genomic Axis:	GA_x5YUV32VY4T (6028380..6058135)
	Chromosome:	19

OMIM NUMBER:
OMIM Information:

Transcript Sequence (SEQ ID NO: 2):

5

Protein Sequence (SEQ ID NO: 50):

SNP Information

10 Context (SEQ ID NO: 98):

CTGGCTACCTTCTGGGTCCCGTCCTCCACCTTCCCCAAATGGGTGACCAAGACGGAA
AGAGGGAGACAGCCCTTGAGATCCTAGACCTGTGGAAGGCCAT
S

15 GACGGGCTCCCCTACTCCCACCCTCCACAGCCCTCCAAGAGGAATGTGATGGAGACG
TTTGCCAAACCAGAGATTGGAGATCTGGGCATGCTCAGCATGA

Celera SNP ID: hCV1936853

Public SNP ID: rs3745833

SNP in Transcript Sequence SEQ ID NO: 2

SNP Position Transcript: 298

20 SNP Source: Applera

Population (Allele,Count): caucasian (C,27|G,13) african american (C,23|G,15) total
(C,50|G,28)

SNP Type: Missense Mutation

Protein Coding: SEQ ID NO: 50, at position 72,(I,ATC) (M,ATG)

25 SNP Source: dbSNP; Celera; HapMap; HGBASE

Population (Allele,Count): caucasian (G,62|C,58)

SNP Type: Missense Mutation

Protein Coding: SEQ ID NO: 50, at position 72,(I,ATC) (M,ATG)

30 Gene Number: 3

Celera Gene: hCG16992 - 145000147109979

Celera Transcript: hCT2303293 - 145000147110000

Public Transcript Accession:

Celera Protein: hCP1809140 - 197000064951781

35 Public Protein Accession:

Gene Symbol: TRAK2

Protein Name: amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidat
e 3

40 Celera Genomic Axis: GA_x5YUV32VWPT (6898299..6992749)

Chromosome: 2

OMIM NUMBER:

OMIM Information:

45 Transcript Sequence (SEQ ID NO: 3):

Protein Sequence (SEQ ID NO: 51):

SNP Information

50

Context (SEQ ID NO: 99):

TTACAATGACATCGACATGGTTACACATCTCCTGGCAGAGAGGGATCGTGATCTGGA
ACTCGCTGCTCGAATTGGACAAGCTCTCTTAAAGCGGAACCAT

R

5 TCTTATCTGAGCAGAACGAATCCCTGGAGGAGCAATTGGGACAAGCCTTTGATCAAG
TTAATCAGCTGCAGCATGAGCTATGCAAGAAAGATGAGTTACT

Celera SNP ID: hCV1226613

Public SNP ID: rs13022344

SNP in Transcript Sequence SEQ ID NO: 3

SNP Position Transcript: 825

10 SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (G,50|A,70)

SNP Type: Missense Mutation

Protein Coding: SEQ ID NO: 51, at position 142,(I,ATC) (V,GTC)

15 Gene Number: 3

Celera Gene: hCG16992 - 145000147109979

Celera Transcript: hCT8039 - 145000147109980

Public Transcript Accession: NM_015049

Celera Protein: hCP34834 - 197000064951780

20 Public Protein Accession: NP_055864

Gene Symbol: TRAK2

Protein Name: amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidat
e 3

25 Celera Genomic Axis: GA_x5YUV32VWPT (6898299..6992749)

Chromosome: 2

OMIM NUMBER:

OMIM Information:

30 Transcript Sequence (SEQ ID NO: 4):

Protein Sequence (SEQ ID NO: 52):

SNP Information

35

Context (SEQ ID NO: 100):

TTACAATGACATCGACATGGTTACACATCTCCTGGCAGAGAGGGATCGTGATCTGGA
ACTCGCTGCTCGAATTGGACAAGCTCTCTTAAAGCGGAACCAT

R

40 TCTTATCTGAGCAGAACGAATCCCTGGAGGAGCAATTGGGACAAGCCTTTGATCAAG
TTAATCAGCTGCAGCATGAGCTATGCAAGAAAGATGAGTTACT

Celera SNP ID: hCV1226613

Public SNP ID: rs13022344

SNP in Transcript Sequence SEQ ID NO: 4

45 SNP Position Transcript: 825

SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (G,50|A,70)

SNP Type: Missense Mutation

Protein Coding: SEQ ID NO: 52, at position 142,(I,ATC) (V,GTC)

50

Gene Number: 4

Celera Gene: hCG1996194 - 65000099360912

Celera Transcript: hCT2277005 - 65000099360942

Public Transcript Accession:

Celera Protein: hCP1891917 - 197000069435181

5 Public Protein Accession:

Gene Symbol: TOMM40

Protein Name: translocase of outer mitochondrial membrane 40 homolog (yeast)

Celera Genomic Axis: GA_x5YUV32VTL1 (7595592..7628038)

Chromosome: 19

10 OMIM NUMBER: 608061

OMIM Information:

Transcript Sequence (SEQ ID NO: 5):

15 Protein Sequence (SEQ ID NO: 53):

SNP Information

Context (SEQ ID NO: 101):

20 CGGGCACATTCGAGGAGTGCCACCGGAAGTGCAAGGAGCTGTTTCCCATTTCAGATG
GAGGGTGTCAAGCTCACAGTCAACAAAGGGTTGAGTAACCATT
Y

CAGGTCAACCACACAGTAGCCCTCAGCACAATCGGGGAGTCCAACTACCACTTCGG
GGTCACATATGTGGGGACAAAGCAGCTGAGTCCCACAGAGGCGT

25 Celera SNP ID: hCV3084827

Public SNP ID: rs157581

SNP in Transcript Sequence SEQ ID NO: 5

SNP Position Transcript: 438

SNP Source: dbSNP; Celera; HGBASE

30 Population (Allele,Count): caucasian (C,28|T,92)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 53, at position 113,(F,TTT) (F,TTC)

Gene Number: 4

35 Celera Gene: hCG1996194 - 65000099360912

Celera Transcript: hCT2277006 - 65000099360913

Public Transcript Accession: NM_006114

Celera Protein: hCP1891915 - 197000069435179

Public Protein Accession: NP_006105

40 Gene Symbol: TOMM40

Protein Name: translocase of outer mitochondrial membrane 40 homolog (yeast)

Celera Genomic Axis: GA_x5YUV32VTL1 (7595592..7628038)

Chromosome: 19

OMIM NUMBER: 608061

45 OMIM Information:

Transcript Sequence (SEQ ID NO: 6):

Protein Sequence (SEQ ID NO: 54):

50

SNP Information

Context (SEQ ID NO: 102):
 CGGGCACATTCGAGGAGTGCCACCGGAAGTGCAAGGAGCTGTTTCCCATTCAGATG
 GAGGGTGTCAAGCTCACAGTCAACAAAGGGTTGAGTAACCATT
 5 Y
 CAGGTCAACCACACAGTAGCCCTCAGCACAATCGGGGAGTCCAACTACCACTTCGG
 GGTACATATGTGGGGACAAAGCAGCTGAGTCCCACAGAGGCGT
 Celera SNP ID: hCV3084827
 Public SNP ID: rs157581
 10 SNP in Transcript Sequence SEQ ID NO: 6
 SNP Position Transcript: 438
 SNP Source: dbSNP; Celera; HGBASE
 Population (Allele,Count): caucasian (C,28|T,92)
 SNP Type: Silent Mutation
 15 Protein Coding: SEQ ID NO: 54, at position 113,(F,TTT) (F,TTC)

Gene Number: 4
 Celera Gene: hCG1996194 - 65000099360912
 Celera Transcript: hCT2277007 - 65000099360927
 20 Public Transcript Accession:
 Celera Protein: hCP1891916 - 197000069435180
 Public Protein Accession:
 Gene Symbol: TOMM40
 Protein Name: translocase of outer mitochondrial membrane 40 homolog (yeast)
 25 Celera Genomic Axis: GA_x5YUV32VTL1 (7595592..7628038)
 Chromosome: 19
 OMIM NUMBER: 608061
 OMIM Information:

30 Transcript Sequence (SEQ ID NO: 7):
 Protein Sequence (SEQ ID NO: 55):

SNP Information

35 Context (SEQ ID NO: 103):
 CGGGCACATTCGAGGAGTGCCACCGGAAGTGCAAGGAGCTGTTTCCCATTCAGATG
 GAGGGTGTCAAGCTCACAGTCAACAAAGGGTTGAGTAACCATT
 Y
 40 CAGGTCAACCACACAGTAGCCCTCAGCACAATCGGGGAGTCCAACTACCACTTCGG
 GGTACATATGTGGGGACAAAGCAGCTGAGTCCCACAGAGGCGT
 Celera SNP ID: hCV3084827
 Public SNP ID: rs157581
 SNP in Transcript Sequence SEQ ID NO: 7
 45 SNP Position Transcript: 401
 SNP Source: dbSNP; Celera; HGBASE
 Population (Allele,Count): caucasian (C,28|T,92)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 55, at position 16,(F,TTT) (F,TTC)
 50 Gene Number: 5

Celera Gene: hCG1778763 - 106000122211945
 Celera Transcript: hCT1817541 - 106000122211960
 Public Transcript Accession: NM_018379
 Celera Protein: hCP1713033 - 197000069372377
 5 Public Protein Accession: NP_060849
 Gene Symbol: FAM63A
 Protein Name: hypothetical protein FLJ11280
 Celera Genomic Axis: GA_x5YUV32VSLS (10371159..10402785)
 Chromosome: 1
 10 OMIM NUMBER:
 OMIM Information:

Transcript Sequence (SEQ ID NO: 8):

15 Protein Sequence (SEQ ID NO: 56):

SNP Information

Context (SEQ ID NO: 104):
 20 ATCCAGCCCCTGGTAAGGCCGGGACTGCAGAAGCAGTCATCCCTGAAAACCATGAG
 GTTCTGGCAGGCCAGATGAGCACCCCTCAGGACACAGATGCAAG
 W
 GATGCTGATGGGGAGGCTAGAGAACGGGAGCCAGCAGACCAAGCTTTGCTGCCTAG
 CCAGTGTGGGGACAACCTTGAGTCCCCTCTGCCTGAAGCTAGCT
 25 Celera SNP ID: hCV22274641
 Public SNP ID:
 SNP in Transcript Sequence SEQ ID NO: 8
 SNP Position Transcript: 1288
 SNP Source: Applera
 30 Population (Allele,Count): caucasian (A,27|T,1) african american (A27,|T,7) total (A,54|T,8)
 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 56, at position 41,(R,AGA) (S,AGT)

Gene Number: 5
 35 Celera Gene: hCG1778763 - 106000122211945
 Celera Transcript: hCT2280498 - 106000122211975
 Public Transcript Accession: NM_018379
 Celera Protein: hCP1859244 - 197000069372378
 Public Protein Accession: NP_060849
 40 Gene Symbol: FAM63A
 Protein Name: hypothetical protein FLJ11280
 Celera Genomic Axis: GA_x5YUV32VSLS (10371159..10402785)
 Chromosome: 1
 OMIM NUMBER:
 45 OMIM Information:

Transcript Sequence (SEQ ID NO: 9):

50 Protein Sequence (SEQ ID NO: 57):

SNP Information

Context (SEQ ID NO: 105):

ATCCAGCCCCCTGGTAAGGCCGGGACTGCAGAAGCAGTCATCCCTGAAAACCATGAG
GTTCTGGCAGGCCCAGATGAGCACCCCTCAGGACACAGATGCAAG

5 W

GATGCTGATGGGGAGGCTAGAGAACGGGAGCCAGCAGACCAAGCTTTGCTGCCTAG
CCAGTGTGGGGACAACCTTGAGTCCCCTCTGCCTGAAGCTAGCT

Celera SNP ID: hCV22274641

Public SNP ID:

10 SNP in Transcript Sequence SEQ ID NO: 9

SNP Position Transcript: 1078

SNP Source: Applera

Population (Allele,Count): caucasian (A,27|T,1) african american (A27,|T,7) total (A,54|T,8)

SNP Type: Missense Mutation

15 Protein Coding: SEQ ID NO: 57, at position 41,(R,AGA) (S,AGT)

Gene Number: 6

Celera Gene: hCG2010706 - 207000007279826

Celera Transcript: hCT2300385 - 207000007279667

20 Public Transcript Accession:

Celera Protein: hCP1797579 - 207000007279782

Public Protein Accession:

Gene Symbol: BCR

Protein Name: breakpoint cluster region

25 Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)

Chromosome: 22

OMIM NUMBER: 151410

OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute
lymphocytic/(3)

30

Transcript Sequence (SEQ ID NO: 10):

Protein Sequence (SEQ ID NO: 58):

35 SNP Information

Context (SEQ ID NO: 106):

AGTGTTTCAGAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
GTGTGAAACTCCAGACTGTCCACAGCATTCCGCTGACCATCAA

40 Y

AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA
GCCACTGGATTTAAGCAGAGTTCAAATCTGTACTGCACCCTGG

Celera SNP ID: hCV2447571

Public SNP ID: rs9608099

45 SNP in Transcript Sequence SEQ ID NO: 10

SNP Position Transcript: 1554

SNP Source: Applera

Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total
(C,22|T,38)

50 SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 58, at position 473,(N,AAT) (N,AAC)

SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (C,36|T,84)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 58, at position 473,(N,AAT) (N,AAC)

5

Gene Number: 6
 Celera Gene: hCG2010706 - 207000007279826
 Celera Transcript: hCT2300386 - 207000007279625
 Public Transcript Accession: NM_004327

10

Celera Protein: hCP1797576 - 207000007279774
 Public Protein Accession: NP_004318

Gene Symbol: BCR
 Protein Name: breakpoint cluster region
 Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)

15

Chromosome: 22
 OMIM NUMBER: 151410
 OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute lymphocytic/(3)

20

Transcript Sequence (SEQ ID NO: 11):

Protein Sequence (SEQ ID NO: 59):

SNP Information

25

Context (SEQ ID NO: 107):
 AGTGTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
 GTGTGAAACTCCAGACTGTCCACAGCATTCGCTGACCATCAA
 Y

30

AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA
 GCCACTGGATTTAAGCAGAGTTCAAATCTGTACTGCACCCTGG

Celera SNP ID: hCV2447571

Public SNP ID: rs9608099

SNP in Transcript Sequence SEQ ID NO: 11

35

SNP Position Transcript: 3296

SNP Source: Applera

Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total (C,22|T,38)

SNP Type: Silent Mutation

40

Protein Coding: SEQ ID NO: 59, at position 900,(N,AAT) (N,AAC)

SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (C,36|T,84)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 59, at position 900,(N,AAT) (N,AAC)

45

Gene Number: 6
 Celera Gene: hCG2010706 - 207000007279826
 Celera Transcript: hCT2300387 - 207000007279781
 Public Transcript Accession:

50

Celera Protein: hCP1797575 - 207000007279772
 Public Protein Accession:

Gene Symbol: BCR
 Protein Name: breakpoint cluster region
 Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)
 Chromosome: 22
 5 OMIM NUMBER: 151410
 OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute lymphocytic/(3)

Transcript Sequence (SEQ ID NO: 12):

10

Protein Sequence (SEQ ID NO: 60):

SNP Information

15 Context (SEQ ID NO: 108):
 AGTGTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
 GTGTGAAACTCCAGACTGTCCACAGCATTCGCTGACCATCAA
 Y
 AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA
 20 GCCACTGGATTTAAGCAGAGTTCAAATCTGTACTGCACCCTGG
 Celera SNP ID: hCV2447571
 Public SNP ID: rs9608099
 SNP in Transcript Sequence . SEQ ID NO: 12
 SNP Position Transcript: 1707
 25 SNP Source: Applera
 Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total (C,22|T,38)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 60, at position 477,(N,AAT) (N,AAC)
 30 SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (C,36|T,84)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 60, at position 477,(N,AAT) (N,AAC)

35 Gene Number: 6
 Celera Gene: hCG2010706 - 207000007279826
 Celera Transcript: hCT2300388 - 207000007279864
 Public Transcript Accession:
 Celera Protein: hCP1797572 - 207000007279768
 40 Public Protein Accession:
 Gene Symbol: BCR
 Protein Name: breakpoint cluster region
 Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)
 Chromosome: 22
 45 OMIM NUMBER: 151410
 OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute lymphocytic/(3)

Transcript Sequence (SEQ ID NO: 13):

50

Protein Sequence (SEQ ID NO: 61):

SNP Information

Context (SEQ ID NO: 109):

5 AGTGTTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
GTGTGAAACTCCAGACTGTCCACAGCATTCCGCTGACCATCAA
Y

AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA
GCCACTGGATTTAAGCAGAGTTCAAATCTGTACTGCACCCTGG

10 Celera SNP ID: hCV2447571

Public SNP ID: rs9608099

SNP in Transcript Sequence SEQ ID NO: 13

SNP Position Transcript: 1707

SNP Source: Applera

15 Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total
(C,22|T,38)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 61, at position 477,(N,AAT) (N,AAC)

SNP Source: dbSNP; Celera; HapMap

20 Population (Allele,Count): caucasian (C,36|T,84)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 61, at position 477,(N,AAT) (N,AAC)

Gene Number: 6

25 Celera Gene: hCG2010706 - 207000007279826

Celera Transcript: hCT2300389 - 207000007279567

Public Transcript Accession:

Celera Protein: hCP1797570 - 207000007279755

Public Protein Accession:

30 Gene Symbol: BCR

Protein Name: breakpoint cluster region

Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)

Chromosome: 22

OMIM NUMBER: 151410

35 OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute
lymphocytic/(3)

Transcript Sequence (SEQ ID NO: 14):

40 Protein Sequence (SEQ ID NO: 62):

SNP Information

Context (SEQ ID NO: 110):

45 AGTGTTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
GTGTGAAACTCCAGACTGTCCACAGCATTCCGCTGACCATCAA
Y

AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA
GCCACTGGATTTAAGCAGAGTTCAAATCTGTACTGCACCCTGG

50 Celera SNP ID: hCV2447571

Public SNP ID: rs9608099

SNP in Transcript Sequence SEQ ID NO: 14

SNP Position Transcript: 1554

SNP Source: Applera

5 Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total (C,22|T,38)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 62, at position 473,(N,AAT) (N,AAC)

SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (C,36|T,84)

10 SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 62, at position 473,(N,AAT) (N,AAC)

Gene Number: 6

Celera Gene: hCG2010706 - 207000007279826

15 Celera Transcript: hCT2300390 - 207000007279688

Public Transcript Accession:

Celera Protein: hCP1797574 - 207000007279770

Public Protein Accession:

Gene Symbol: BCR

20 Protein Name: breakpoint cluster region

Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)

Chromosome: 22

OMIM NUMBER: 151410

25 OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute lymphocytic/(3)

Transcript Sequence (SEQ ID NO: 15):

Protein Sequence (SEQ ID NO: 63):

30

SNP Information

Context (SEQ ID NO: 111):

35 AGTGTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
GTGTGAAACTCCAGACTGTCCACAGCATTCCGCTGACCATCAA

Y

AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA

GCCACTGGATTTAAGCAGAGTTCAAGTAAGTACTGATCTGTAC

Celera SNP ID: hCV2447571

40 Public SNP ID: rs9608099

SNP in Transcript Sequence SEQ ID NO: 15

SNP Position Transcript: 1707

SNP Source: Applera

45 Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total (C,22|T,38)

SNP Type: UTR5

SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (C,36|T,84)

50 SNP Type: UTR5

Gene Number: 6

Celera Gene: hCG2010706 - 207000007279826
 Celera Transcript: hCT2300392 - 207000007279726
 Public Transcript Accession:
 Celera Protein: hCP1797577 - 207000007279776
 5 Public Protein Accession:
 Gene Symbol: BCR
 Protein Name: breakpoint cluster region
 Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)
 Chromosome: 22
 10 OMIM NUMBER: 151410
 OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute lymphocytic/(3)

 Transcript Sequence (SEQ ID NO: 16):
 15 Protein Sequence (SEQ ID NO: 64):

 SNP Information
 20 Context (SEQ ID NO: 112):
 AGTGTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
 GTGTGAAACTCCAGACTGTCCACAGCATTCCGCTGACCATCAA
 Y
 AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA
 25 GCCACTGGATTTAAGCAGAGTTCAAATCTGTACTGCACCCTGG
 Celera SNP ID: hCV2447571
 Public SNP ID: rs9608099
 SNP in Transcript Sequence SEQ ID NO: 16
 SNP Position Transcript: 3298
 30 SNP Source: Applera
 Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total (C,22|T,38)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 64, at position 403,(N,AAT) (N,AAC)
 35 SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (C,36|T,84)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 64, at position 403,(N,AAT) (N,AAC)

 40 Gene Number: 6
 Celera Gene: hCG2010706 - 207000007279826
 Celera Transcript: hCT2300393 - 207000007279598
 Public Transcript Accession: NM_021574
 Celera Protein: hCP1797573 - 207000007279769
 45 Public Protein Accession: NP_067585
 Gene Symbol: BCR
 Protein Name: breakpoint cluster region
 Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)
 Chromosome: 22
 50 OMIM NUMBER: 151410

OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute lymphocytic/(3)

Transcript Sequence (SEQ ID NO: 17):

5

Protein Sequence (SEQ ID NO: 65):

SNP Information

10 Context (SEQ ID NO: 113):

AGTGTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
GTGTGAAACTCCAGACTGTCCACAGCATTCCGCTGACCATCAA
Y

15 AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA
GCCACTGGATTAAAGCAGAGTTCAAATCTGTACTGCACCCTGG

Celera SNP ID: hCV2447571

Public SNP ID: rs9608099

SNP in Transcript Sequence SEQ ID NO: 17

SNP Position Transcript: 3296

20 SNP Source: Applera

Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total (C,22|T,38)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 65, at position 900,(N,AAT) (N,AAC)

25 SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (C,36|T,84)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 65, at position 900,(N,AAT) (N,AAC)

30 Gene Number: 6

Celera Gene: hCG2010706 - 207000007279826

Celera Transcript: hCT2300394 - 207000007279827

Public Transcript Accession: NM_004327

Celera Protein: hCP1797568 - 207000007279749

35 Public Protein Accession: NP_004318

Gene Symbol: BCR

Protein Name: breakpoint cluster region

Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)

Chromosome: 22

40 OMIM NUMBER: 151410

OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute lymphocytic/(3)

Transcript Sequence (SEQ ID NO: 18):

45

Protein Sequence (SEQ ID NO: 66):

SNP Information

50 Context (SEQ ID NO: 114):

AGTGTTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
GTGTGAAACTCCAGACTGTCCACAGCATTCCGCTGACCATCAA

Y

AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA
5 GCCACTGGATTTAAGCAGAGTTCAAATCTGTACTGCACCCTGG

Celera SNP ID: hCV2447571

Public SNP ID: rs9608099

SNP in Transcript Sequence SEQ ID NO: 18

SNP Position Transcript: 3296

10 SNP Source: Applera

Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total
(C,22|T,38)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 66, at position 900,(N,AAT) (N,AAC)

15 SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (C,36|T,84)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 66, at position 900,(N,AAT) (N,AAC)

20 Gene Number: 6

Celera Gene: hCG2010706 - 207000007279826

Celera Transcript: hCT2300395 - 207000007279853

Public Transcript Accession: NM_021574

Celera Protein: hCP1797569 - 207000007279752

25 Public Protein Accession: NP_067585

Gene Symbol: BCR

Protein Name: breakpoint cluster region

Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)

Chromosome: 22

30 OMIM NUMBER: 151410

OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute
lymphocytic/(3)

Transcript Sequence (SEQ ID NO: 19):

35 Protein Sequence (SEQ ID NO: 67):

SNP Information

40 Context (SEQ ID NO: 115):

AGTGTTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
GTGTGAAACTCCAGACTGTCCACAGCATTCCGCTGACCATCAA

Y

AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA
45 GCCACTGGATTTAAGCAGAGTTCAAATCTGTACTGCACCCTGG

Celera SNP ID: hCV2447571

Public SNP ID: rs9608099

SNP in Transcript Sequence SEQ ID NO: 19

SNP Position Transcript: 3296

50 SNP Source: Applera

- Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total (C,22|T,38)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 67, at position 900,(N,AAT) (N,AAC)
- 5 SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (C,36|T,84)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 67, at position 900,(N,AAT) (N,AAC)
- 10 Gene Number: 6
 Celera Gene: hCG2010706 - 207000007279826
 Celera Transcript: hCT2300396 - 207000007279798
 Public Transcript Accession:
 Celera Protein: hCP1797578 - 207000007279779
- 15 Public Protein Accession:
 Gene Symbol: BCR
 Protein Name: breakpoint cluster region
 Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)
 Chromosome: 22
- 20 OMIM NUMBER: 151410
 OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute lymphocytic/(3)
- Transcript Sequence (SEQ ID NO: 20):
- 25 Protein Sequence (SEQ ID NO: 68):
- SNP Information
- 30 Context (SEQ ID NO: 116):
 AGTGTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
 GTGTGAAACTCCAGACTGTCCACAGCATTCCGCTGACCATCAA
 Y
 AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA
- 35 GCCACTGGATTAAAGCAGAGTTCAAATCTGTACTGCACCCTGG
 Celera SNP ID: hCV2447571
 Public SNP ID: rs9608099
 SNP in Transcript Sequence SEQ ID NO: 20
 SNP Position Transcript: 1690
- 40 SNP Source: Applera
 Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total (C,22|T,38)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 68, at position 449,(N,AAT) (N,AAC)
- 45 SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (C,36|T,84)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 68, at position 449,(N,AAT) (N,AAC)
- 50 Gene Number: 6
 Celera Gene: hCG2010706 - 207000007279826

Celera Transcript: hCT2300397 - 207000007279636

Public Transcript Accession:

Celera Protein: hCP1797571 - 207000007279767

Public Protein Accession:

5 Gene Symbol: BCR

Protein Name: breakpoint cluster region

Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)

Chromosome: 22

OMIM NUMBER: 151410

10 OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute lymphocytic/(3)

Transcript Sequence (SEQ ID NO: 21):

15 Protein Sequence (SEQ ID NO: 69):

SNP Information

Context (SEQ ID NO: 117):

20 AGTGTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
GTGTGAAACTCCAGACTGTCCACAGCATTCGCTGACCATCAA
Y

AAGGAAGATGATGAGTCTCCGGGGCTCTATGGGTTTCTGAATGTCATCGTCCACTCA
GCCACTGGATTTAAGCAGAGTTCAAATCTGTACTGCACCCTGG

25 Celera SNP ID: hCV2447571

Public SNP ID: rs9608099

SNP in Transcript Sequence SEQ ID NO: 21

SNP Position Transcript: 1696

SNP Source: Applera

30 Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total
(C,22|T,38)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 69, at position 473,(N,AAT) (N,AAC)

SNP Source: dbSNP; Celera; HapMap

35 Population (Allele,Count): caucasian (C,36|T,84)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 69, at position 473,(N,AAT) (N,AAC)

Gene Number: 7

40 Celera Gene: hCG1811794 - 84000313720751

Celera Transcript: hCT1954680 - 84000313720752

Public Transcript Accession: NM_032507

Celera Protein: hCP1765540 - 197000069366270

Public Protein Accession: NP_115896

45 Gene Symbol: PGBD1

Protein Name: piggyBac transposable element derived 1

Celera Genomic Axis: GA_x5YUV32W6W6 (1310121..1351328)

Chromosome: 6

OMIM NUMBER:

50 OMIM Information:

Transcript Sequence (SEQ ID NO: 22):

Protein Sequence (SEQ ID NO: 70):

5 SNP Information

Context (SEQ ID NO: 118):

GTCAGGTCCCAGACATTGGTGAAGACTGAGGAAGAAACAGCCCAGGCCGTTGCTGC
AGAGAAGTGGTCACATCTGAGTCTGACTCGGAGGAACCTCTGTG

10

R
GAACTCAGCTCAGGAGACAGTTATGAGCCTCAGTCCGATGACTGAAGAAATTGTAA
CTAAAGATAGATTGTTTAAAGCAAAGCAAGAACTTCTGAAGAA

Celera SNP ID: hCV25942411

Public SNP ID: rs3800324

15

SNP in Transcript Sequence SEQ ID NO: 22

SNP Position Transcript: 1110

SNP Source: Applera

Population (Allele,Count): caucasian (A,0|G,34) african american (A,2|G,34) total (A,2|G,68)

SNP Type: Missense Mutation

20

Protein Coding: SEQ ID NO: 70, at position 244,(G,GGG) (E,GAG)

SNP Source: dbSNP; HapMap; HGBASE

Population (Allele,Count): caucasian (G,109|A,7)

SNP Type: Missense Mutation

Protein Coding: SEQ ID NO: 70, at position 244,(G,GGG) (E,GAG)

25

Gene Number: 7

Celera Gene: hCG1811794 - 84000313720751

Celera Transcript: hCT2281921 - 84000313720763

Public Transcript Accession: NM_032507

30

Celera Protein: hCP1853138 - 197000069366272

Public Protein Accession: NP_115896

Gene Symbol: PGBD1

Protein Name: piggyBac transposable element derived 1

Celera Genomic Axis: GA_x5YUV32W6W6 (1310121..1351328)

35

Chromosome: 6

OMIM NUMBER:

OMIM Information:

Transcript Sequence (SEQ ID NO: 23):

40

Protein Sequence (SEQ ID NO: 71):

SNP Information

45 Context (SEQ ID NO: 119):

GTCAGGTCCCAGACATTGGTGAAGACTGAGGAAGAAACAGCCCAGGCCGTTGCTGC
AGAGAAGTGGTCACATCTGAGTCTGACTCGGAGGAACCTCTGTG

R

GAACTCAGCTCAGGAGACAGTTATGAGCCTCAGTCCGATGACTGAAGAAATTGTAA

50

CTAAAGATAGATTGTTTAAAGCAAAGCAAGAACTTCTGAAGAA

Celera SNP ID: hCV25942411

Public SNP ID: rs3800324
 SNP in Transcript Sequence SEQ ID NO: 23
 SNP Position Transcript: 769
 SNP Source: Applera
 5 Population (Allele,Count): caucasian (A,0|G,34) african american (A,2|G,34) total (A,2|G,68)
 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 71, at position 244,(G,GGG) (E,GAG)
 SNP Source: dbSNP; HapMap; HGBASE
 Population (Allele,Count): caucasian (G,109|A,7)
 10 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 71, at position 244,(G,GGG) (E,GAG)

 Gene Number: 7
 Celera Gene: hCG1811794 - 84000313720751
 15 Celera Transcript: hCT2281924 - 84000313720802
 Public Transcript Accession: NM_032507
 Celera Protein: hCP1853143 - 197000069366277
 Public Protein Accession: NP_115896
 Gene Symbol: PGBD1
 20 Protein Name: piggyBac transposable element derived 1
 Celera Genomic Axis: GA_x5YUV32W6W6 (1310121..1351328)
 Chromosome: 6
 OMIM NUMBER:
 OMIM Information:
 25 Transcript Sequence (SEQ ID NO: 24):

 Protein Sequence (SEQ ID NO: 72):

 30 SNP Information

 Context (SEQ ID NO: 120):
 GTCAGGTCCCAGACATTGGTGAAGACTGAGGAAGAAACAGCCCAGGCCGTTGCTGC
 AGAGAAGTGGTCACATCTGAGTCTGACTCGGAGGAACCTCTGTG
 35 R
 GAACTCAGCTCAGGAGACAGTTATGAGCCTCAGTCCGATGACTGAAGAAATTGTAA
 CTAAAGATAGATTGTTTAAAGCAAAGCAAGAACTTCTGAAGAA
 Celera SNP ID: hCV25942411
 Public SNP ID: rs3800324
 40 SNP in Transcript Sequence SEQ ID NO: 24
 SNP Position Transcript: 1110
 SNP Source: Applera
 Population (Allele,Count): caucasian (A,0|G,34) african american (A,2|G,34) total (A,2|G,68)
 SNP Type: Missense Mutation
 45 Protein Coding: SEQ ID NO: 72, at position 244,(G,GGG) (E,GAG)
 SNP Source: dbSNP; HapMap; HGBASE
 Population (Allele,Count): caucasian (G,109|A,7)
 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 72, at position 244,(G,GGG) (E,GAG)
 50 Gene Number: 8

Celera Gene: hCG1997343 - 63000132525707
 Celera Transcript: hCT2278868 - 63000132525756
 Public Transcript Accession:
 Celera Protein: hCP1859535 - 197000069372669
 5 Public Protein Accession:
 Gene Symbol: LMNA
 Protein Name: lamin A/C
 Celera Genomic Axis: GA_x5YUV32VSLS (5284831..5367979)
 Chromosome: 1
 10 OMIM NUMBER: 150330
 OMIM Information: Emery-Dreifuss muscular dystrophy, AD, 181350 (3);
 Cardiomyopathy, dilated, 1A, 115200 (3); Lipodystrophy, familial partial, 151660 (3); Emery-Dreifuss muscular
 15 dystrophy, AR, 604929 (3); Charcot-Marie-Tooth disease, type 2B1, 605588 (3); Muscular dystrophy, limb-girdle, type 1B, 159001 (3); Mandibuloacral dysplasia, 248370 (3); Hutchinson-Gilford progeria, 176670 (3); LDHCP syndrome, 608056; Myopathy, early-onset, and progeroid features, 176670 (3);
 20 Restrictive dermopathy, lethal, 275210 (3); Cardiomyopathy, dilated, with quadriceps myopathy, 607920 (3)
 Transcript Sequence (SEQ ID NO: 25):
 25 Protein Sequence (SEQ ID NO: 73):
 SNP Information
 Context (SEQ ID NO: 121):
 30 GCAGCGTCACCAAAAAGCGCAAACCTGGAGTCCACTGAGAGCCGCAGCAGCTTCTCA
 CAGCACGCACGCACTAGCGGGCGCGTGGCCGTGGAGGAGGTGGA
 Y
 GAGGAGGGCAAGTTTGTCCGGCTGCGCAACAAGTCCAATGAGGACCAGTCCATGGG
 CAATTGGCAGATCAAGCGCCAGAATGGAGATGATCCCTTGCTGA
 35 Celera SNP ID: hCV630374
 Public SNP ID: rs505058
 SNP in Transcript Sequence SEQ ID NO: 25
 SNP Position Transcript: 1830
 SNP Source: Applera
 40 Population (Allele,Count): caucasian (C,2|T,36) african american (C,22|T,16) total (C,24|T,52)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 73, at position 572,(D,GAT) (D,GAC)
 SNP Source: dbSNP; HGBASE
 45 Population (Allele,Count): caucasian (T,107|C,9)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 73, at position 572,(D,GAT) (D,GAC)
 Gene Number: 8
 50 Celera Gene: hCG1997343 - 63000132525707
 Celera Transcript: hCT2278869 - 63000132525774

Public Transcript Accession:

Celera Protein: hCP1859536 - 197000069372670

Public Protein Accession:

Gene Symbol: LMNA

5 Protein Name: lamin A/C

Celera Genomic Axis: GA_x5YUV32VSLS (5284831..5367979)

Chromosome: 1

OMIM NUMBER: 150330

OMIM Information: Emery-Dreifuss muscular dystrophy, AD, 181350 (3);

10 Cardiomyopathy, dilated, 1A, 115200 (3); Lipodystrophy, familial partial, 151660 (3); Emery-Dreifuss muscular

dystrophy

, AR, 604929 (3); Charcot-Marie-Tooth disease, type 2B1, 605588 (3); Muscular dystrophy, limb-girdle

15 , type 1B, 159001 (3); Mandibuloacral dysplasia, 248370 (3); Hutchinson-Gilford progeria, 176670 (3)

; LDHCP syndrome, 608056; Myopathy, early-onset, and progeroid features, 176670 (3);

Restrictive der

mopathy, lethal, 275210 (3); Cardiomyopathy, dilated, with quadriceps myopathy, 607920 (3)

20

Transcript Sequence (SEQ ID NO: 26):

Protein Sequence (SEQ ID NO: 74):

25 SNP Information

Context (SEQ ID NO: 122):

GCAGCGTCACCAAAAAGCGCAAAGTGGAGTCCACTGAGAGCCGCAGCAGCTTCTCA
CAGCACGCACGCACTAGCGGGCGCGTGGCCGTGGAGGAGGTGGA

30 Y

GAGGAGGGCAAGTTTGTCCGGCTGCGCAACAAGTCCAATGAGGACCAGTCCATGGG
CAATTGGCAGATCAAGCGCCAGAATGGAGATGATCCCTTGCTGA

Celera SNP ID: hCV630374

Public SNP ID: rs505058

35 SNP in Transcript Sequence SEQ ID NO: 26

SNP Position Transcript: 1830

SNP Source: Applera

Population (Allele,Count): caucasian (C,2|T,36) african american (C,22|T,16) total
(C,24|T,52)

40 SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 74, at position 572,(D,GAT) (D,GAC)

SNP Source: dbSNP; HGBASE

Population (Allele,Count): caucasian (T,107|C,9)

SNP Type: Silent Mutation

45 Protein Coding: SEQ ID NO: 74, at position 572,(D,GAT) (D,GAC)

Gene Number: 8

Celera Gene: hCG1997343 - 63000132525707

Celera Transcript: hCT2278870 - 63000132525708

50 Public Transcript Accession: NM_005572

Celera Protein: hCP1859532 - 197000069372666

Public Protein Accession: NP_005563
 Gene Symbol: LMNA
 Protein Name: lamin A/C
 Celera Genomic Axis: GA_x5YUV32VSLS (5284831..5367979)
 5 Chromosome: 1
 OMIM NUMBER: 150330
 OMIM Information: Emery-Dreifuss muscular dystrophy, AD, 181350 (3);
 Cardiomyopathy, dilated, 1A, 115200 (3); Lipodystrophy, familial partial, 151660 (3); Emery-Dreifuss muscular
 10 dystrophy, AR, 604929 (3); Charcot-Marie-Tooth disease, type 2B1, 605588 (3); Muscular dystrophy, limb-girdle, type 1B, 159001 (3); Mandibuloacral dysplasia, 248370 (3); Hutchinson-Gilford progeria, 176670 (3)
 15 ; LDHCP syndrome, 608056; Myopathy, early-onset, and progeroid features, 176670 (3); Restrictive dermopathy, lethal, 275210 (3); Cardiomyopathy, dilated, with quadriceps myopathy, 607920 (3)
 Transcript Sequence (SEQ ID NO: 27):
 20 Protein Sequence (SEQ ID NO: 75):
 SNP Information
 25 Context (SEQ ID NO: 123):
 GCAGCGTCACCAAAAAGCGCAAAGTGGAGTCCACTGAGAGCCGCAGCAGCTTCTCA
 CAGCACGCACGCACTAGCGGGCGCGTGGCCGTGGAGGAGGTGGA
 Y
 GAGGAGGGCAAGTTTGTCCGGCTGCGCAACAAGTCCAATGAGGACCAGTCCATGGG
 30 CAATTGGCAGATCAAGCGCCAGAATGGAGATGATCCCTTGCTGA
 Celera SNP ID: hCV630374
 Public SNP ID: rs505058
 SNP in Transcript Sequence SEQ ID NO: 27
 SNP Position Transcript: 2071
 35 SNP Source: Applera
 Population (Allele,Count): caucasian (C,2|T,36) african american (C,22|T,16) total (C,24|T,52)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 75, at position 446,(D,GAT) (D,GAC)
 40 SNP Source: dbSNP; HGBASE
 Population (Allele,Count): caucasian (T,107|C,9)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 75, at position 446,(D,GAT) (D,GAC)
 45 Gene Number: 8
 Celera Gene: hCG1997343 - 63000132525707
 Celera Transcript: hCT2278871 - 63000132525739
 Public Transcript Accession: NM_005572
 Celera Protein: hCP1859534 - 197000069372668
 50 Public Protein Accession: NP_005563
 Gene Symbol: LMNA

Protein Name: lamin A/C

Celera Genomic Axis: GA_x5YUV32VSLS (5284831..5367979)

Chromosome: 1

OMIM NUMBER: 150330

5 OMIM Information: Emery-Dreifuss muscular dystrophy, AD, 181350 (3);

Cardiomyopathy, dilated, 1A, 115200 (3); Lipodystrophy, familial partial, 151660 (3); Emery-Dreifuss muscular

dystrophy, AR, 604929 (3); Charcot-Marie-Tooth disease, type 2B1, 605588 (3); Muscular dystrophy, limb-girdle

10 , type 1B, 159001 (3); Mandibuloacral dysplasia, 248370 (3); Hutchinson-Gilford progeria, 176670 (3)

; LDHCP syndrome, 608056; Myopathy, early-onset, and progeroid features, 176670 (3); Restrictive der

15 mopathy, lethal, 275210 (3); Cardiomyopathy, dilated, with quadriceps myopathy, 607920 (3)

Transcript Sequence (SEQ ID NO: 28):

Protein Sequence (SEQ ID NO: 76):

20

SNP Information

Context (SEQ ID NO: 124):

25 GCAGCGTCACCAAAAAGCGCAAAGTGGAGTCCACTGAGAGCCGCAGCAGCTTCTCA
CAGCACGCACGCACTAGCGGGCGCGTGGCCGTGGAGGAGGTGGA

Y

GAGGAGGGCAAGTTTGTCCGGCTGCGCAACAAGTCCAATGAGGACCAGTCCATGGG
CAATTGGCAGATCAAGCGCCAGAATGGAGATGATCCCTTGCTGA

Celera SNP ID: hCV630374

30 Public SNP ID: rs505058

SNP in Transcript Sequence SEQ ID NO: 28

SNP Position Transcript: 1716

SNP Source: Applera

35 Population (Allele,Count): caucasian (C,2|T,36) african american (C,22|T,16) total
(C,24|T,52)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 76, at position 446,(D,GAT) (D,GAC)

SNP Source: dbSNP; HGBASE

Population (Allele,Count): caucasian (T,107|C,9)

40 SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 76, at position 446,(D,GAT) (D,GAC)

Gene Number: 8

Celera Gene: hCG1997343 - 63000132525707

45 Celera Transcript: hCT2278872 - 63000132525725

Public Transcript Accession: NM_005572

Celera Protein: hCP1859533 - 197000069372667

Public Protein Accession: NP_005563

Gene Symbol: LMNA

50 Protein Name: lamin A/C

Celera Genomic Axis: GA_x5YUV32VSLS (5284831..5367979)

Chromosome: 1
 OMIM NUMBER: 150330
 OMIM Information: Emery-Dreifuss muscular dystrophy, AD, 181350 (3);
 Cardiomyopathy, dilated, 1A, 115200 (3); Lipodystrophy, familial partial, 151660 (3); Emery-Dreifuss muscular
 5 dystrophy, AR, 604929 (3); Charcot-Marie-Tooth disease, type 2B1, 605588 (3); Muscular dystrophy, limb-girdle,
 type 1B, 159001 (3); Mandibuloacral dysplasia, 248370 (3); Hutchinson-Gilford progeria,
 10 176670 (3); LDHCP syndrome, 608056; Myopathy, early-onset, and progeroid features, 176670 (3);
 Restrictive dermopathy, lethal, 275210 (3); Cardiomyopathy, dilated, with quadriceps myopathy, 607920 (3)

15 Transcript Sequence (SEQ ID NO: 29):
 Protein Sequence (SEQ ID NO: 77):
 SNP Information

20 Context (SEQ ID NO: 125):
 GCAGCGTCACCAAAAAGCGCAAAGTGGAGTCCACTGAGAGCCGCGAGCAGCTTCTCA
 CAGCACGCACGCACTAGCGGGCGCGTGGCCGTGGAGGAGGTGGA
 Y

25 GAGGAGGGCAAGTTTGTCCGGCTGCGCAACAAGTCCAATGAGGACCAGTCCATGGG
 CAATTGGCAGATCAAGCGCCAGAATGGAGATGATCCCTTGCTGA
 Celera SNP ID: hCV630374
 Public SNP ID: rs505058
 SNP in Transcript Sequence SEQ ID NO: 29
 30 SNP Position Transcript: 1544
 SNP Source: Applera
 Population (Allele,Count): caucasian (C,2|T,36) african american (C,22|T,16) total
 (C,24|T,52)
 SNP Type: Silent Mutation
 35 Protein Coding: SEQ ID NO: 77, at position 446,(D,GAT) (D,GAC)
 SNP Source: dbSNP; HGBASE
 Population (Allele,Count): caucasian (T,107|C,9)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 77, at position 446,(D,GAT) (D,GAC)

40 Gene Number: 8
 Celera Gene: hCG1997343 - 63000132525707
 Celera Transcript: hCT2278873 - 63000132525793
 Public Transcript Accession:
 45 Celera Protein: hCP1859539 - 197000069372673
 Public Protein Accession:
 Gene Symbol: LMNA
 Protein Name: lamin A/C
 Celera Genomic Axis: GA_x5YUV32VSLS (5284831..5367979)
 50 Chromosome: 1
 OMIM NUMBER: 150330

OMIM Information: Emery-Dreifuss muscular dystrophy, AD, 181350 (3);

Cardiomyopathy, dilated, 1A, 115200 (3); Lipodystrophy, familial partial, 151660 (3); Emery-Dreifuss muscular

dystrophy, AR, 604929 (3); Charcot-Marie-Tooth disease, type 2B1, 605588 (3); Muscular dystrophy, limb-girdle

, type 1B, 159001 (3); Mandibuloacral dysplasia, 248370 (3); Hutchinson-Gilford progeria, 176670 (3)

; LDHCP syndrome, 608056; Myopathy, early-onset, and progeroid features, 176670 (3);

Restrictive dermopathy, lethal, 275210 (3); Cardiomyopathy, dilated, with quadriceps myopathy, 607920 (3)

Transcript Sequence (SEQ ID NO: 30):

Protein Sequence (SEQ ID NO: 78):

SNP Information

Context (SEQ ID NO: 126):

GCAGCGTCACCAAAAAGCGCAAAGTGGAGTCCACTGAGAGCCGCGCAGCAGCTTCTCA
CAGCACGCACGCACTAGCGGGCGCGTGGCCGTGGAGGAGGTGGA

Y

GAGGAGGGCAAGTTTGTCCGGCTGCGCAACAAGTCCAATGAGGACCAAGTCCATGGG
CAATTGGCAGATCAAGCGCCAGAATGGAGATGATCCCTTGCTGA

Celera SNP ID: hCV630374

Public SNP ID: rs505058

SNP in Transcript Sequence SEQ ID NO: 30

SNP Position Transcript: 1830

SNP Source: Applera

Population (Allele, Count): caucasian (C, 2 | T, 36) african american (C, 22 | T, 16) total (C, 24 | T, 52)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 78, at position 572, (D, GAT) (D, GAC)

SNP Source: dbSNP; HGBASE

Population (Allele, Count): caucasian (T, 107 | C, 9)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 78, at position 572, (D, GAT) (D, GAC)

Gene Number: 9

Celera Gene: hCG20335 - 65000099361015

Celera Transcript: hCT11411 - 65000099361023

Public Transcript Accession: NM_001646

Celera Protein: hCP38031 - 197000069435191

Public Protein Accession: NP_001637

Gene Symbol: APOC4

Protein Name: apolipoprotein C-IV

Celera Genomic Axis: GA_x5YUV32VTL1 (7646587..7670232)

Chromosome: 19

OMIM NUMBER: 600745

OMIM Information:

Transcript Sequence (SEQ ID NO: 31):

Protein Sequence (SEQ ID NO: 79):

5 SNP Information

Context (SEQ ID NO: 127):

CTCCTCAGAAACAGGCTCCAGGCCCTGCCTGCCCTGTGCCTCTGCGTGCTGGTCCTG
GCCTGCATTGGGGCATGCCAGCCAGAGGCCAGGAAGGAACCC

10 Y

GAGCCCCCACCAAAGCTAAAGATGAGTCGCTGGAGCCTGGTGAGGGGCAGGATGA
AGGAGCTGCTGGAGACAGTGGTGAACAGGACCAGAGACGGGTGG

Celera SNP ID: hCV1841831

Public SNP ID: rs1132899

15 SNP in Transcript Sequence SEQ ID NO: 31

SNP Position Transcript: 147

SNP Source: Applera

Population (Allele,Count): caucasian (C,7|T,11) african american (C,15|T,13) total
(C,22|T,24)

20 SNP Type: Missense Mutation

Protein Coding: SEQ ID NO: 79, at position 36,(L,CTG) (P,CCG)

SNP Source: dbSNP; Celera; HapMap; HGBASE

Population (Allele,Count): caucasian (C,51|T,65)

SNP Type: Missense Mutation

25 Protein Coding: SEQ ID NO: 79, at position 36,(L,CTG) (P,CCG)

Gene Number: 9

Celera Gene: hCG20335 - 65000099361015

Celera Transcript: hCT2277012 - 65000099361016

30 Public Transcript Accession: NM_001646

Celera Protein: hCP1891926 - 197000069435190

Public Protein Accession: NP_001637

Gene Symbol: APOC4

Protein Name: apolipoprotein C-IV

35 Celera Genomic Axis: GA_x5YUV32VTL1 (7646587..7670232)

Chromosome: 19

OMIM NUMBER: 600745

OMIM Information:

40 Transcript Sequence (SEQ ID NO: 32):

Protein Sequence (SEQ ID NO: 80):

SNP Information

45

Context (SEQ ID NO: 128):

CTCCTCAGAAACAGGCTCCAGGCCCTGCCTGCCCTGTGCCTCTGCGTGCTGGTCCTG
GCCTGCATTGGGGCATGCCAGCCAGAGGCCAGGAAGGAACCC

Y

50 GAGCCCCCACCAAAGCTAAAGATGAGTCGCTGGAGCCTGGTGAGGGGCAGGATGA
AGGAGCTGCTGGAGACAGTGGTGAACAGGACCAGAGACGGGTGG

Celera SNP ID: hCV1841831
 Public SNP ID: rs1132899
 SNP in Transcript Sequence SEQ ID NO: 32
 SNP Position Transcript: 147
 5 SNP Source: Applera
 Population (Allele,Count): caucasian (C,7|T,11) african american (C,15|T,13) total (C,22|T,24)
 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 80, at position 36,(L,CTG) (P,CCG)
 10 SNP Source: dbSNP; Celera; HapMap; HGBASE
 Population (Allele,Count): caucasian (C,51|T,65)
 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 80, at position 36,(L,CTG) (P,CCG)
 15 Gene Number: 10
 Celera Gene: hCG22337 - 146000220683873
 Celera Transcript: hCT13430 - 146000220683874
 Public Transcript Accession:
 Celera Protein: hCP39814 - 197000069390979
 20 Public Protein Accession:
 Gene Symbol: SERPINA13
 Protein Name: serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiprotei nase, antitrypsin), member 13
 25 Celera Genomic Axis: GA_x5YUV32W0QP (41908319..41977785)
 Chromosome: 14
 OMIM NUMBER:
 OMIM Information:
 30 Transcript Sequence (SEQ ID NO: 33):
 Protein Sequence (SEQ ID NO: 81):
 SNP Information
 35 Context (SEQ ID NO: 129):
 ACTCTGGGCACTGAGTTCAGTCGGCCCTTCCTGACGAGAAAACAGCAAGGAAGGCT
 CAACTCGGGCTTGCCCCGGGGGGCTCTTCCCCTTCGGCGTGGAAG
 W
 40 TTCAATGCGAGGAAGCAAAGAGCCTAAAATACGAACACCCAGGAAATTAAGCAGGG
 GAAGGGGCCCTTCTGAGACAGTAAAGAAACATACATCCGAAGAC
 Celera SNP ID: hCV2188996
 Public SNP ID: rs11622883
 SNP in Transcript Sequence SEQ ID NO: 33
 45 SNP Position Transcript: 1100
 SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (T,62|A,58)
 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 81, at position 367,(V,GTT) (D,GAT)
 50 Gene Number: 11

Celera Gene: hCG25596 - 146000220733557
 Celera Transcript: hCT16721 - 146000220733558
 Public Transcript Accession:
 Celera Protein: hCP42419 - 197000069409199
 5 Public Protein Accession:
 Gene Symbol:
 Protein Name:
 Celera Genomic Axis: GA_x54KRFTF0F9 (28582642..28603669)
 Chromosome: 6
 10 OMIM NUMBER:
 OMIM Information:

Transcript Sequence (SEQ ID NO: 34):

15 Protein Sequence (SEQ ID NO: 82):

SNP Information

Context (SEQ ID NO: 130):
 20 ATGGAGCTCAGCCCTCCACCTCCACAGCTCAAGAGCAAGATGATGTTCTCATAGTTG
 ATTAGGATGAAGAAGATTCTTCAAATAATGCCGACGTCAGTGA
 R
 GAAGAGAGAAGCCGCAAGAAGAAATTAGATGAGAAAGAGAATCTCAGTGCAAAGA
 GGTCACGTATAGAACAGAAGGAAGAGCTTGATGATGTCATAGCAT
 25 Celera SNP ID: hCV25741966
 Public SNP ID: rs6907175
 SNP in Transcript Sequence SEQ ID NO: 34
 SNP Position Transcript: 236
 SNP Source: Applera
 30 Population (Allele,Count): caucasian (A,24|G,16) african american (A,29|G,7) total
 (A,53|G,23)
 SNP Type: UTR3
 SNP Source: dbSNP; HapMap
 Population (Allele,Count): caucasian (G,59|A,61)
 35 SNP Type: UTR3

Gene Number: 12
 Celera Gene: hCG25621 - 104000116735770
 Celera Transcript: hCT16746 - 104000116735771
 40 Public Transcript Accession: NM_006398
 Celera Protein: hCP42368 - 197000069366772
 Public Protein Accession: NP_006389
 Gene Symbol: UBD
 Protein Name: ubiquitin D
 45 Celera Genomic Axis: GA_x5YUV32W6W6 (2583709..2607910)
 Chromosome: 6
 OMIM NUMBER: 606050
 OMIM Information:

50 Transcript Sequence (SEQ ID NO: 35):

Protein Sequence (SEQ ID NO: 83):

SNP Information

- 5 Context (SEQ ID NO: 131):
 AAGAGACTGGAAGATGGGAAGATGATGGCAGATTACGGCATCAGAAAGGGCAACTT
 ACTCTTCCTGGCATCTTATTGTATTGGAGGGTGACCACCCTGGG
 S
 ATGGGGTGTTGGCAGGGGTCAAAAAGCTTATTTCTTTTAATCTCTTACTCAACGAAC
 10 ACATCTTCTGATGATTTCCTCCAAAATTAATGAGAATGAGATGAG
 Celera SNP ID: hCV588691
 Public SNP ID: rs444013
 SNP in Transcript Sequence SEQ ID NO: 35
 SNP Position Transcript: 537
 15 SNP Source: Applera
 Population (Allele,Count): caucasian (G,10|C,8) african american (G,16|C,6) total (G,26|C,14)
 SNP Type: UTR3
 SNP Source: dbSNP; Celera; HapMap; HGBASE
 Population (Allele,Count): caucasian (G,49|C,55)
 20 SNP Type: UTR3

 Gene Number: 13
 Celera Gene: hCG28649 - 67000129132473
 Celera Transcript: hCT19804 - 67000129132557
 25 Public Transcript Accession:
 Celera Protein: hCP44058 - 197000069364458
 Public Protein Accession:
 Gene Symbol: AGC1
 Protein Name: aggrecan 1 (chondroitin sulfate proteoglycan 1, large aggregating prot
 30 eoglycan, antigen identified by monoclonal antibody A0122)
 Celera Genomic Axis: GA_x5YUV32W6GH (4231607..4322592)
 Chromosome: 15
 OMIM NUMBER: 155760
 OMIM Information:
 35 Transcript Sequence (SEQ ID NO: 36):

 Protein Sequence (SEQ ID NO: 84):

 40 SNP Information

 Context (SEQ ID NO: 132):
 GAAGTTCTAGAGATTTCTGTCTCTGGAGTAGAGGACATCAGTGGGCTTCCTTCTGGA
 GAGGTTGTAGAGACTTCTGCCTCTGGAATAGAGGATGTCAGTG
 45 M
 ACTTCCTTCAGGAGAAGGTCTAGAGACCTCTGCTTCTGGAGTAGAGGACCTCAGCAG
 GCTCCCTTCTGGAGAAGAAGTTCTAGAGATTTCTGCCTCTGGA
 Celera SNP ID: hCV1834284
 Public SNP ID: rs2882676
 50 SNP in Transcript Sequence SEQ ID NO: 36
 SNP Position Transcript: 4101

SNP Source: dbSNP; Celera; HapMap; HGBASE
 Population (Allele,Count): caucasian (A,43|C,77)
 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 84, at position 1242,(E,GAA) (A,GCA)

5

Gene Number: 13
 Celera Gene: hCG28649 - 67000129132473
 Celera Transcript: hCT2269025 - 67000129132516
 Public Transcript Accession:

10

Celera Protein: hCP1851322 - 197000069364456
 Public Protein Accession:

Gene Symbol: AGC1
 Protein Name: aggrecan 1 (chondroitin sulfate proteoglycan 1, large aggregating proteoglycan, antigen identified by monoclonal antibody A0122)

15

Celera Genomic Axis: GA_x5YUV32W6GH (4231607..4322592)
 Chromosome: 15
 OMIM NUMBER: 155760
 OMIM Information:

20 Transcript Sequence (SEQ ID NO: 37):

Protein Sequence (SEQ ID NO: 85):

SNP Information

25

Context (SEQ ID NO: 133):
 GAAGTTCTAGAGATTTCTGTCTCTGGAGTAGAGGACATCAGTGGGCTTCCTTCTGGA
 GAGGTTGTAGAGACTTCTGCCTCTGGAATAGAGGATGTCAGTG
 M

30

ACTTCCTTCAGGAGAAGGTCTAGAGACCTCTGCTTCTGGAGTAGAGGACCTCAGCAG
 GCTCCCTTCTGGAGAAGAAGTTCTAGAGATTTCTGCCTCTGGA

Celera SNP ID: hCV1834284

Public SNP ID: rs2882676

SNP in Transcript Sequence SEQ ID NO: 37

35

SNP Position Transcript: 4101

SNP Source: dbSNP; Celera; HapMap; HGBASE

Population (Allele,Count): caucasian (A,43|C,77)

SNP Type: Missense Mutation

Protein Coding: SEQ ID NO: 85, at position 1242,(E,GAA) (A,GCA)

40

Gene Number: 13
 Celera Gene: hCG28649 - 67000129132473
 Celera Transcript: hCT2269026 - 67000129132494
 Public Transcript Accession:

45

Celera Protein: hCP1851320 - 197000069364454
 Public Protein Accession:

Gene Symbol: AGC1

Protein Name: aggrecan 1 (chondroitin sulfate proteoglycan 1, large aggregating proteoglycan, antigen identified by monoclonal antibody A0122)

50

Celera Genomic Axis: GA_x5YUV32W6GH (4231607..4322592)
 Chromosome: 15

OMIM NUMBER: 155760
OMIM Information:

Transcript Sequence (SEQ ID NO: 38):

5

Protein Sequence (SEQ ID NO: 86):

SNP Information

10 Context (SEQ ID NO: 134):

GAAGTTCTAGAGATTTCTGTCTCTGGAGTAGAGGACATCAGTGGGCTTCCTTCTGGA
GAGGTTGTAGAGACTTCTGCCTCTGGAATAGAGGATGTCAGTG
M

15 ACTTCCTTCAGGAGAAGGTCTAGAGACCTCTGCTTCTGGAGTAGAGGACCTCAGCAG
GCTCCCTTCTGGAGAAGAAGTTCTAGAGATTTCTGCCTCTGGA

Celera SNP ID: hCV1834284

Public SNP ID: rs2882676

SNP in Transcript Sequence SEQ ID NO: 38

SNP Position Transcript: 3784

20 SNP Source: dbSNP; Celera; HapMap; HGBASE

Population (Allele,Count): caucasian (A,43|C,77)

SNP Type: Missense Mutation

Protein Coding: SEQ ID NO: 86, at position 1242,(E,GAA) (A,GCA)

25 Gene Number: 13

Celera Gene: hCG28649 - 67000129132473

Celera Transcript: hCT2269027 - 67000129132536

Public Transcript Accession:

Celera Protein: hCP1851323 - 197000069364457

30 Public Protein Accession:

Gene Symbol: AGC1

Protein Name: aggrecan 1 (chondroitin sulfate proteoglycan 1, large aggregating proteoglycan, antigen identified by monoclonal antibody A0122)

Celera Genomic Axis: GA_x5YUV32W6GH (4231607..4322592)

35 Chromosome: 15

OMIM NUMBER: 155760

OMIM Information:

Transcript Sequence (SEQ ID NO: 39):

40

Protein Sequence (SEQ ID NO: 87):

SNP Information

45 Context (SEQ ID NO: 135):

GAAGTTCTAGAGATTTCTGTCTCTGGAGTAGAGGACATCAGTGGGCTTCCTTCTGGA
GAGGTTGTAGAGACTTCTGCCTCTGGAATAGAGGATGTCAGTG
M

50 ACTTCCTTCAGGAGAAGGTCTAGAGACCTCTGCTTCTGGAGTAGAGGACCTCAGCAG
GCTCCCTTCTGGAGAAGAAGTTCTAGAGATTTCTGCCTCTGGA

Celera SNP ID: hCV1834284

Public SNP ID: rs2882676
 SNP in Transcript Sequence SEQ ID NO: 39
 SNP Position Transcript: 4101
 SNP Source: dbSNP; Celera; HapMap; HGBASE
 5 Population (Allele,Count): caucasian (A,43|C,77)
 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 87, at position 1242,(E,GAA) (A,GCA)

Gene Number: 13
 10 Celera Gene: hCG28649 - 67000129132473
 Celera Transcript: hCT2269028 - 67000129132474
 Public Transcript Accession:
 Celera Protein: hCP1851319 - 197000069364453
 Public Protein Accession:
 15 Gene Symbol: AGC1
 Protein Name: aggrecan 1 (chondroitin sulfate proteoglycan 1, large aggregating proteoglycan, antigen identified by monoclonal antibody A0122)
 Celera Genomic Axis: GA_x5YUV32W6GH (4231607..4322592)
 Chromosome: 15
 20 OMIM NUMBER: 155760
 OMIM Information:

Transcript Sequence (SEQ ID NO: 40):

25 Protein Sequence (SEQ ID NO: 88):

SNP Information

Context (SEQ ID NO: 136):
 30 GAAGTTCTAGAGATTTCTGTCTCTGGAGTAGAGGACATCAGTGGGCTTCCTTCTGGA
 GAGGTTGTAGAGACTTCTGCCTCTGGAATAGAGGATGTCAGTG
 M
 ACTTCCTTCAGGAGAAGGTCTAGAGACCTCTGCTTCTGGAGTAGAGGACCTCAGCAG
 GCTCCCTTCTGGAGAAGAAGTTCTAGAGATTTCTGCCTCTGGA
 35 Celera SNP ID: hCV1834284
 Public SNP ID: rs2882676
 SNP in Transcript Sequence SEQ ID NO: 40
 SNP Position Transcript: 3784
 SNP Source: dbSNP; Celera; HapMap; HGBASE
 40 Population (Allele,Count): caucasian (A,43|C,77)
 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 88, at position 1242,(E,GAA) (A,GCA)

Gene Number: 14
 45 Celera Gene: hCG37868 - 84000314283049
 Celera Transcript: hCT29103 - 84000314283050
 Public Transcript Accession: NM_002591
 Celera Protein: hCP49882 - 197000069371703
 Public Protein Accession: NP_002582
 50 Gene Symbol: PCK1
 Protein Name: phosphoenolpyruvate carboxykinase 1 (soluble)

Printed: 09/14/2007

SPC

PCT/US2007/016254

US 2007/016254

LI, et al.
Date of Deposit: July 16, 2007

Attorney Docket #: CD000009PCT
Express Mail #: ER 622077997 US

Celera Genomic Axis: GA_x5YUV32VYHA (6713662..6739039)
Chromosome: 20
OMIM NUMBER: 261680
OMIM Information: Hypoglycemia due to PCK1 deficiency (1)

5

Transcript Sequence (SEQ ID NO: 41):

Protein Sequence (SEQ ID NO: 89):

10 SNP Information

Context (SEQ ID NO: 137):

TGGGTACGGCGGGAAGTCTCGGGAAGAAGTGCTTTGCTCTCAGGATGGCCAG
CCGGCTGGCCAAGGAGGAAGGGTGGCTGGCAGAGCACATGCTG

15

R
TTCTGGGTATAACCAACCCTGAGGGTGAGAAGAAGTACCTGGCGGCCGCATTTCCCA
GCGCCTGCGGGAAGACCAACCTGGCCATGATGAACCCCAGCCT

Celera SNP ID: hCV2508722

Public SNP ID: rs8192708

20 SNP in Transcript Sequence SEQ ID NO: 41

SNP Position Transcript: 966

SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (A,108|G,12)

SNP Type: Missense Mutation

25 Protein Coding: SEQ ID NO: 89, at position 267,(I,ATT) (V,GTT)

Gene Number: 15

Celera Gene: hCG39378 - 30000023325762

Celera Transcript: hCT2280516 - 30000023325759

30 Public Transcript Accession: NM_004079

Celera Protein: hCP1859259 - 30000023325755

Public Protein Accession: NP_004070

Gene Symbol: CTSS

Protein Name: cathepsin S

35 Celera Genomic Axis: GA_x5YUV32VSL (10613850..10669294)

Chromosome: 1

OMIM NUMBER: 116845

OMIM Information:

40 Transcript Sequence (SEQ ID NO: 42):

Protein Sequence (SEQ ID NO: 90):

45 SNP Information

Context (SEQ ID NO: 138):

TCCTTGTTGCTTTGAAATCTTAGAAGAGAGCCCACTAATTCAAGGACTCTTACTGTAG
GAGCAACTGCTGGTTCTATCACAATGAAACGGCTGGTTTGTG

Y

50 GCTCTTGGTGTGCTCCTCTGCAGTGGCACAGTTGCATAAAGATCCTACCCTGGATCA
CCTGTCATCTCTGGAAGAAAACCTATGGCAAACAATAACAAG

Celera SNP ID: hCV15746640
 Public SNP ID:
 SNP in Transcript Sequence SEQ ID NO: 42
 SNP Position Transcript: 281
 5 SNP Source: CDX_Alzheimer
 Population (Allele,Count): caucasian (T,104|C,16)
 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 90, at position 7,(V,GTG) (A,GCG)

10 Gene Number: 15
 Celera Gene: hCG39378 - 30000023325762
 Celera Transcript: hCT2280518 - 30000023325763
 Public Transcript Accession:
 Celera Protein: hCP1859260 - 30000023325757
 15 Public Protein Accession:
 Gene Symbol: CTSS
 Protein Name: cathepsin S
 Celera Genomic Axis: GA_x5YUV32VSLS (10613850..10669294)
 Chromosome: 1
 20 OMIM NUMBER: 116845
 OMIM Information:

Transcript Sequence (SEQ ID NO: 43):

25 Protein Sequence (SEQ ID NO: 91):

SNP Information

Context (SEQ ID NO: 139):

30 TCCTTGTTGCTTTGAAATCTTAGAAGAGAGCCCACTAATTCAAGGACTCTTACTGTAG
 GAGCAACTGCTGGTTCTATCACAATGAAACGGCTGGTTTGTG
 Y
 GCTCTTGGTGTGCTCCTCTGCAGTGGCACAGTTGCATAAAGATCCTACCCTGGATCA
 CCACTGGCATCTCTGGAAGAAAACCTATGGCAAACAATACAAG

35 Celera SNP ID: hCV15746640
 Public SNP ID:
 SNP in Transcript Sequence SEQ ID NO: 43
 SNP Position Transcript: 281
 SNP Source: CDX_Alzheimer
 40 Population (Allele,Count): caucasian (T,104|C,16)
 SNP Type: Missense Mutation
 Protein Coding: SEQ ID NO: 91, at position 7,(V,GTG) (A,GCG)

Gene Number: 15
 45 Celera Gene: hCG39378 - 30000023325762
 Celera Transcript: hCT30629 - 30000023325788
 Public Transcript Accession: NM_004079
 Celera Protein: hCP49339 - 30000023325753
 Public Protein Accession: NP_004070
 50 Gene Symbol: CTSS
 Protein Name: cathepsin S

Celera Genomic Axis: GA_x5YUV32VSLS (10613850..10669294)

Chromosome: 1

OMIM NUMBER: 116845

OMIM Information:

5

Transcript Sequence (SEQ ID NO: 44):

Protein Sequence (SEQ ID NO: 92):

10 SNP Information

Context (SEQ ID NO: 140):

TCCTTGTTGCTTTGAAATCTTAGAAGAGAGCCCACTAATTCAAGGACTCTTACTGTAG
GAGCAACTGCTGGTTCTATCACAATGAAACGGCTGGTTTGTG

15 Y

GCTCTTGGTGTGCTCCTCTGCAGTGGCACAGTTGCATAAAGATCCTACCCTGGATCA
CCACTGGCATCTCTGGAAGAAAACCTATGGCAAACAATAACAAG

Celera SNP ID: hCV15746640

Public SNP ID:

20 SNP in Transcript Sequence SEQ ID NO: 44

SNP Position Transcript: 281

SNP Source: CDX_Alzheimer

Population (Allele,Count): caucasian (T,104|C,16)

SNP Type: Missense Mutation

25 Protein Coding: SEQ ID NO: 92, at position 7,(V,GTG) (A,GCG)

Gene Number: 16

Celera Gene: hCG42001 - 103000085248952

Celera Transcript: hCT2269221 - 103000085248953

30 Public Transcript Accession: NM_003985

Celera Protein: hCP1891079 - 197000069434343

Public Protein Accession: NP_003976

Gene Symbol: TNK1

Protein Name: tyrosine kinase, non-receptor, 1

35 Celera Genomic Axis: GA_x5YUV32W2JD (8122109..8151222)

Chromosome: 17

OMIM NUMBER: 608076

OMIM Information:

40 Transcript Sequence (SEQ ID NO: 45):

Protein Sequence (SEQ ID NO: 93):

SNP Information

45

Context (SEQ ID NO: 141):

TAGGAACTCCTTCTCCAGACATGCTTCCTGAGGCTGGCTCTCTGTGGCTACTGAAGCT
GCTCCGGGACATCCAGTTGGCCAGTTTACTGGCCCATCCT

W

50 GAGGAGCTTAATGTCACTCGGCCAGAGCACTTCGACTTTGTAAAGCCTGAGGACCTG
GACGGCATTGGCATGGGCCGGCCTGCCAGCGCAGACTGTCCG

Celera SNP ID: hCV8717889
 Public SNP ID: rs1554948
 SNP in Transcript Sequence SEQ ID NO: 45
 SNP Position Transcript: 172
 5 SNP Source: dbSNP; HapMap; HGBASE
 Population (Allele,Count): caucasian (T,57|A,63)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 93, at position 27,(L,CTT) (L,CTA)

10 Context (SEQ ID NO: 142):
 GACCTGGTCTCTCTAGGGCCTACCCTGAGCTCACCATCTGAAGGAGAGTGCCATCAT
 CCTTAGGAACCTCTTCTCCAGACATGCTTCCTGAGGCTGGCTC
 Y
 CTGTGGCTACTGAAGCTGCTCCGGGACATCCAGTTGGCCCAGTTTTACTGGCCCCATC
 15 CTTGAGGAGCTTAATGTCACTCGGCCAGAGCACTTCGACTTTG
 Celera SNP ID: hCV1247824
 Public SNP ID: rs1554947
 SNP in Transcript Sequence SEQ ID NO: 45
 SNP Position Transcript: 112
 20 Related Interrogated SNP: hCV8717889 (Power=.8)
 SNP Source: dbSNP; Celera; HapMap; HGBASE
 Population (Allele,Count): caucasian (C,73|T,43)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 93, at position 7,(S,TCT) (S,TCC)

25 Gene Number: 16
 Celera Gene: hCG42001 - 103000085248952
 Celera Transcript: hCT2269223 - 103000085248986
 Public Transcript Accession:
 30 Celera Protein: hCP1891081 - 197000069434345
 Public Protein Accession:
 Gene Symbol: TNK1
 Protein Name: tyrosine kinase, non-receptor, 1
 Celera Genomic Axis: GA_x5YUV32W2JD (8122109..8151222)
 35 Chromosome: 17
 OMIM NUMBER: 608076
 OMIM Information:

Transcript Sequence (SEQ ID NO: 46):
 40 Protein Sequence (SEQ ID NO: 94):

SNP Information

45 Context (SEQ ID NO: 143):
 TAGGAACCTCTTCTCCAGACATGCTTCCTGAGGCTGGCTCTCTGTGGCTACTGAAGCT
 GCTCCGGGACATCCAGTTGGCCCAGTTTTACTGGCCCCATCCT
 W
 GAGGAGCTTAATGTCACTCGGCCAGAGCACTTCGACTTTGTAAAGCCTGAGGACCTG
 50 GACGGCATTGGCATGGGCCGGCCTGCCAGCGCAGACTGTCCG
 Celera SNP ID: hCV8717889

Public SNP ID: rs1554948
 SNP in Transcript Sequence SEQ ID NO: 46
 SNP Position Transcript: 198
 SNP Source: dbSNP; HapMap; HGBASE
 5 Population (Allele,Count): caucasian (T,57|A,63)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 94, at position 27,(L,CTT) (L,CTA)

Context (SEQ ID NO: 144):

10 GACCTGGTCTCTCTAGGGCCTACCCTGAGCTCACCATCTGAAGGAGAGTGCCATCAT
 CCTTAGGAACCTCTTCTCCAGACATGCTTCCTGAGGCTGGCTC
 Y

CTGTGGCTACTGAAGCTGCTCCGGGACATCCAGTTGGCCCAGTTTACTGGCCCCATC
 CTTGAGGAGCTTAATGTCACTCGGCCAGAGCACTTCGACTTTG

15 Celera SNP ID: hCV1247824
 Public SNP ID: rs1554947
 SNP in Transcript Sequence SEQ ID NO: 46
 SNP Position Transcript: 138
 Related Interrogated SNP: hCV8717889 (Power=.8)
 20 SNP Source: dbSNP; Celera; HapMap; HGBASE
 Population (Allele,Count): caucasian (C,73|T,43)
 SNP Type: Silent Mutation
 Protein Coding: SEQ ID NO: 94, at position 7,(S,TCT) (S,TCC)

25 Gene Number: 16
 Celera Gene: hCG42001 - 103000085248952
 Celera Transcript: hCT33277 - 103000085248969
 Public Transcript Accession: NM_003985
 Celera Protein: hCP51810 - 197000069434344
 30 Public Protein Accession: NP_003976
 Gene Symbol: TNK1
 Protein Name: tyrosine kinase, non-receptor, 1
 Celera Genomic Axis: GA_x5YUV32W2JD (8122109..8151222)
 Chromosome: 17
 35 OMIM NUMBER: 608076
 OMIM Information:

Transcript Sequence (SEQ ID NO: 47):

40 Protein Sequence (SEQ ID NO: 95):

SNP Information

Context (SEQ ID NO: 145):

45 TAGGAACCTCCTTCTCCAGACATGCTTCCTGAGGCTGGCTCTCTGTGGCTACTGAAGCT
 GCTCCGGGACATCCAGTTGGCCCAGTTTACTGGCCCCATCCT
 W

GAGGAGCTTAATGTCACTCGGCCAGAGCACTTCGACTTTGTAAAGCCTGAGGACCTG
 GACGGCATTGGCATGGGCCGGCCTGCCAGCGCAGACTGTCCG

50 Celera SNP ID: hCV8717889
 Public SNP ID: rs1554948

SNP in Transcript Sequence SEQ ID NO: 47

SNP Position Transcript: 198

SNP Source: dbSNP; HapMap; HGBASE

Population (Allele,Count): caucasian (T,57|A,63)

5 SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 95, at position 27,(L,CTT) (L,CTA)

Context (SEQ ID NO: 146):

10 GACCTGGTCTCTCTAGGGCCTACCCTGAGCTCACCATCTGAAGGAGAGTGCCATCAT
CCTTAGGAACCTCTTCTCCAGACATGCTTCCTGAGGCTGGCTC
Y

CTGTGGCTACTGAAGCTGCTCCGGGACATCCAGTTGGCCCAGTTTTACTGGCCCATC
CTTGAGGAGCTTAATGTCACTCGGCCAGAGCACTTCGACTTTG

Celera SNP ID: hCV1247824

15 Public SNP ID: rs1554947

SNP in Transcript Sequence SEQ ID NO: 47

SNP Position Transcript: 138

Related Interrogated SNP: hCV8717889 (Power=.8)

SNP Source: dbSNP; Celera; HapMap; HGBASE

20 Population (Allele,Count): caucasian (C,73|T,43)

SNP Type: Silent Mutation

Protein Coding: SEQ ID NO: 95, at position 7,(S,TCT) (S,TCC)

Gene Number: 17

25 Celera Gene: hCG2044039 - 30000063654100

Celera Transcript: hCT2351144 - 30000063654103

Public Transcript Accession: NM_203307

Celera Protein: hCP1916374 - 30000063654101

Public Protein Accession: NP_976052

30 Gene Symbol: MGC35402

Protein Name: hypothetical protein MGC35402

Celera Genomic Axis: GA_x5YUV32VY4T (2771485..2810693)

Chromosome: 19

OMIM NUMBER:

35 OMIM Information:

Transcript Sequence (SEQ ID NO: 48):

Protein Sequence (SEQ ID NO: 96):

40

SNP Information

Context (SEQ ID NO: 147):

45 ATGTGAGGTATGTGACAAGGTCCTTAATCAGAATTGATTCCTTGTATGCCATAATAG
ATGTCACACTGGTAAGAAACCTTACAGTTGTTATGAGTGTGGC
R

AGACCTTCAGTCAGACGTCATCCTTTACATGCTATCGTAGACTTCATACTGGAGGGA
AACCTTACAAGTGTAAGTGTGAGCATAACAAGACCTTTGGTTGAAA

Celera SNP ID: hCV29207239

50 Public SNP ID: rs8110693

SNP in Transcript Sequence SEQ ID NO: 48

SNP Position Transcript: 745
Related Interrogated SNP: hCV3084827 (Power=.7)
SNP Source: dbSNP
Population (Allele,Count): caucasian (A,113|G,1)
5 SNP Type: UTR3

Table 2

	Gene Number:	1
	Celera Gene:	hCG1640936 - 67000129056428
5	Gene Symbol:	MYH13
	Protein Name:	myosin, heavy polypeptide 13, skeletal muscle
	Celera Genomic Axis:	GA_x5YUV32W2JD (5134352..5226503)
	Chromosome:	17
	OMIM NUMBER:	603487
10	OMIM Information:	
	Genomic Sequence (SEQ ID NO: 148):	
	SNP Information	
15	Context	(SEQ ID NO: 177):
	GGGTGAAAGTTGGCTTTTAGGCTGAGACACAGAGGTGGTAAAAGAAATATCTACTTT	
	TTCAATTTCTCTTCTATTTGCTGCTTGTCATTTTCTAGATCCA	
	Y	
20	AATGGATTCTGTTGGACATTTTCAGATCTCCTTCCAGCTTCCTCTTCGCCCTTTCCAAG	
	TCCGCCCGCAGTTTCTTCTCCTGCTCTAAGGAACCCTCAAGC	
	Celera SNP ID:	hCV2179030
	Public SNP ID:	rs2074877
	SNP in Genomic Sequence:	SEQ ID NO: 148
25	SNP Position Genomic:	29532
	SNP Source:	Applera
	Population (Allele,Count):	caucasian (C,14 T,2) african american (C,2 T,4) total (C,16 T,6)
	SNP Type:	MISSENSE MUTATION; ESS; TRANSCRIPTION FACTOR
	BINDING SITE; HUMAN-MOUSE	
30	SYNTENIC REGION	
	SNP Source:	dbSNP; Celera; HapMap; HGBASE
	Population (Allele,Count):	caucasian (T,46 C,74)
	SNP Type:	MISSENSE MUTATION; ESS; TRANSCRIPTION FACTOR
	BINDING SITE; HUMAN-MOUSE	
35	SYNTENIC REGION	
	Gene Number:	2
	Celera Gene:	hCG1641280 - 65000099415023
	Gene Symbol:	GALP
40	Protein Name:	galanin-like peptide precursor
	Celera Genomic Axis:	GA_x5YUV32VY4T (6028380..6058135)
	Chromosome:	19
	OMIM NUMBER:	
	OMIM Information:	
45	Genomic Sequence (SEQ ID NO: 149):	
	SNP Information	
50	Context	(SEQ ID NO: 178):

GCTCTCTCCCTCCCTACCCAGTCCTCCACCTTCCCCAAATGGGTGACCAAGACGGAA
AGAGGGAGACAGCCCTTGAGATCCTAGACCTGTGGAAGGCCAT

S

GGTGAGTGAGCGGGGAGTTAGACGTCTTCTCCTGCGTCCTCCCCTGCGGGGCTCTCTC
CTGGCTTTGGAAACTGGTAGATGTGAGCTCCAGCCCTCACCCCT

Celera SNP ID: hCV1936853

Public SNP ID: rs3745833

SNP in Genomic Sequence: SEQ ID NO: 149

SNP Position Genomic: 16231

10 SNP Source: Applera

Population (Allele,Count): caucasian (C,27|G,13) african american (C,23|G,15) total
(C,50|G,28)

SNP Type: MISSENSE MUTATION; HUMAN-MOUSE SYNTENIC REGION

SNP Source: dbSNP; Celera; HapMap; HGBASE

15 Population (Allele,Count): caucasian (G,62|C,58)

SNP Type: MISSENSE MUTATION; HUMAN-MOUSE SYNTENIC REGION

Gene Number: 3

Celera Gene: hCG16992 - 145000147109979

20 Gene Symbol: TRAK2

Protein Name: amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidat
e 3

Celera Genomic Axis: GA_x5YUV32VWPT (6898299..6992749)

Chromosome: 2

25 OMIM NUMBER:

OMIM Information:

Genomic Sequence (SEQ ID NO: 150):

30 SNP Information

Context (SEQ ID NO: 179):

AAACCTAATTATTTTCATACTGCAGACATTTGATCAAAGGCTTACTTGATCAAAGGCT
TGTCCCAATTGCTCCTCCAGGGATTCTGTTCTGCTCAGATAAGA

35 Y

ATGGTTCCGCTTTAAGAGAGCTTGTCCAATTCGAGCAGCGAGTTCCAGATCACGATC
CCTCTGTAAAATAACCAAATACACCTTTTCTTTGAGTATAAAC

Celera SNP ID: hCV1226613

Public SNP ID: rs13022344

40 SNP in Genomic Sequence: SEQ ID NO: 150

SNP Position Genomic: 32229

SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (C,50|T,70)

SNP Type: MISSENSE MUTATION; TRANSCRIPTION FACTOR BINDING

45 SITE; HUMAN-MOUSE SYNT
ENIC REGION

Gene Number: 4

Celera Gene: hCG1996194 - 65000099360912

50 Gene Symbol: TOMM40

Protein Name: translocase of outer mitochondrial membrane 40 homolog (yeast)

Celera Genomic Axis: GA_x5YUV32VTL1 (7595592..7628038)

Chromosome: 19

OMIM NUMBER: 608061

OMIM Information:

5

Genomic Sequence (SEQ ID NO: 151):

SNP Information

10 Context (SEQ ID NO: 180):

GTTGGAGTGGAGTGTGACAGCGTTTCTCTTCTCCAGAGCTGTTTCCCATTTCAGATGGA
GGGTGTCAAGCTCACAGTCAACAAAGGGTTGAGTAACCATTT

Y

CAGGTGAGCCTTCCTGGTGTCTTACCCACCAGAGATCGTCCCCGCCGTCCCCCTCCC
15 TGCATCTGCACACTCGGCCCAATTACTCCTCCCTCAAGAGCT

Celera SNP ID: hCV3084827

Public SNP ID: rs157581

SNP in Genomic Sequence: SEQ ID NO: 151

SNP Position Genomic: 11236

20 SNP Source: dbSNP; Celera; HGBASE

Population (Allele,Count): caucasian (C,28|T,92)

SNP Type: TRANSCRIPTION FACTOR BINDING SITE; HUMAN-MOUSE

SYNENIC REGION; SILENT

MUTATION

25

Context (SEQ ID NO: 181):

CCAGGCACAGGAAAGGACAGGGTCAGGAAAGGAGGACTCTGGGCGGCAGCCTCCA
CATTCCCCTTCCACGCTTGCCCCCAGAATGGAGGAGGGTGTCTG

K

30 ATTACTGGGCGAGGTGTCCTCCCTTCCTGGGGACTGTGGGGGGGTGGTCAAAAGACCT
CTATGCCCCACCTCCTTCCTCCCTCTGCCCTGCTGTGCCTGGG

Celera SNP ID: hCV905013

Public SNP ID: rs405509

SNP in Genomic Sequence: SEQ ID NO: 151

35 SNP Position Genomic: 24337

SNP Source: dbSNP; HapMap; ABI_Val; HGBASE; CDX_Heart

Population (Allele,Count): caucasian (T,60|G,60)

SNP Type: INTERGENIC; UNKNOWN

40 Gene Number: 5

Celera Gene: hCG1778763 - 106000122211945

Gene Symbol: FAM63A

Protein Name: hypothetical protein FLJ11280

Celera Genomic Axis: GA_x5YUV32VSL5 (10371159..10402785)

45 Chromosome: 1

OMIM NUMBER:

OMIM Information:

Genomic Sequence (SEQ ID NO: 152):

50

SNP Information

Context (SEQ ID NO: 182):

AGCTAGCTTCAGGCAGAGGGGACTCAAGGTTGTCCCCACACTGGCTAGGCAGCAAA
GCTTGGTCTGCTGGCTCCCGTTCTCTAGCCTCCCCATCAGCATC

5 W

CTTGCACTCTGTGTCCTGAGGGTGCTCATCTGGGCCTGCCAGAACCTCATGGTTTTTCAG
GGATGACTGCTTCTGCAGTCCCGGCCTTACCAGGGGGCTGGAT

Celera SNP ID: hCV22274641

Public SNP ID:

10 SNP in Genomic Sequence: SEQ ID NO: 152

SNP Position Genomic: 15756

SNP Source: Applera

Population (Allele,Count): caucasian (T,27|A,1) african american (T27,|A,7) total (T,54|A,8)

SNP Type: MISSENSE MUTATION; ESS; HUMAN-MOUSE SYNTENIC

15 REGION; INTRON

Gene Number: 6

Celera Gene: hCG2010706 - 207000007279826

Gene Symbol: BCR

20 Protein Name: breakpoint cluster region

Celera Genomic Axis: GA_x5YUV32VU5C (1592626..1750151)

Chromosome: 22

OMIM NUMBER: 151410

OMIM Information: Leukemia, chronic myeloid, 608232 (3); Leukemia, acute

25 lymphocytic/(3)

Genomic Sequence (SEQ ID NO: 153):

SNP Information

30

Context (SEQ ID NO: 183):

CAGGTTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
GTGTGAAACTCCAGACTGTCCACAGCATTCCGCTGACCATCAA

Y

35 AAGGAAGGTGGGCCCCCCCCGTTTCCGTGTACAGGGCACCTGCAGGGAGGGCAGGCA

GCTAGCCTGAAGGCTGATCCCCCCTTCCTGTTAGCACTTTTGAT

Celera SNP ID: hCV2447571

Public SNP ID: rs9608099

SNP in Genomic Sequence: SEQ ID NO: 153

40 SNP Position Genomic: 119887

SNP Source: Applera

Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total
(C,22|T,38)

SNP Type: HUMAN-MOUSE SYNTENIC REGION; UTR5; SILENT

45 MUTATION; INTRON

SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (C,36|T,84)

SNP Type: HUMAN-MOUSE SYNTENIC REGION; UTR5; SILENT

MUTATION; INTRON

50

Gene Number: 7

Celera Gene: hCG1811794 - 84000313720751
 Gene Symbol: PGBD1
 Protein Name: piggyBac transposable element derived 1
 Celera Genomic Axis: GA_x5YUV32W6W6 (1310121..1351328)
 5 Chromosome: 6
 OMIM NUMBER:
 OMIM Information:

Genomic Sequence (SEQ ID NO: 154):

SNP Information

Context (SEQ ID NO: 184):

TTGTCTTTTGTAGACATTGGTGAAGACTGAGGAAGAAACAGCCCAGGCCGTTGCTGCA
 15 GAGAAGTGGTCACATCTGAGTCTGACTCGGAGGAACCTCTGTG

R

GAAGTCAGCTCAGGAGACAGTTATGAGCCTCAGTCCGATGAGTAAGGCCAGGCCTCT
 GGCTGGACCCCTGTCTGGACTCTCATTTTCCTGTCACTCACCC

Celera SNP ID: hCV25942411

20 Public SNP ID: rs3800324

SNP in Genomic Sequence: SEQ ID NO: 154

SNP Position Genomic: 25326

SNP Source: Applera

Population (Allele,Count): caucasian (A,0|G,34) african american (A,2|G,34) total (A,2|G,68)

25 SNP Type: MISSENSE MUTATION

SNP Source: dbSNP; HapMap; HGBASE

Population (Allele,Count): caucasian (G,109|A,7)

SNP Type: MISSENSE MUTATION

30 Gene Number: 8

Celera Gene: hCG1997343 - 63000132525707

Gene Symbol: LMNA

Protein Name: lamin A/C

Celera Genomic Axis: GA_x5YUV32VSLS (5284831..5367979)

35 Chromosome: 1

OMIM NUMBER: 150330

OMIM Information: Emery-Dreifuss muscular dystrophy, AD, 181350 (3);

Cardiomyopathy, dilated, 1A, 115200 (3); Lipodystrophy, familial partial, 151660 (3); Emery-Dreifuss muscular

40 dystrophy, AR, 604929 (3); Charcot-Marie-Tooth disease, type 2B1, 605588 (3); Muscular dystrophy, limb-girdle

, type 1B, 159001 (3); Mandibuloacral dysplasia, 248370 (3); Hutchinson-Gilford progeria, 176670 (3)

45 ; LDHCP syndrome, 608056; Myopathy, early-onset, and progeroid features, 176670 (3); Restrictive der

mopathy, lethal, 275210 (3); Cardiomyopathy, dilated, with quadriceps myopathy, 607920 (3)

Genomic Sequence (SEQ ID NO: 155):

SNP Information

SNP Information

Context (SEQ ID NO: 185):

GCAGCGTCACCAAAAAGCGCAAACCTGGAGTCCACTGAGAGCCGCAGCAGCTTCTCA
CAGCACGCACGCACTAGCGGGCGCGTGGCCGTGGAGGAGGTGGA

5 Y

GAGGAGGGCAAGTTTGTCCGGCTGCGCAACAAGTCCAATGAGGTAGGCTCCTGCTC
AGGGTCTAAGGGGATACAGCTGCATCAGGGAGAGAGTGGCAAGA

Celera SNP ID: hCV630374

Public SNP ID: rs505058

10 SNP in Genomic Sequence: SEQ ID NO: 155

SNP Position Genomic: 69461

SNP Source: Applera

Population (Allele,Count): caucasian (C,2|T,36) african american (C,22|T,16) total
(C,24|T,52)

15 SNP Type: HUMAN-MOUSE SYNTENIC REGION; SILENT MUTATION

SNP Source: dbSNP; HGBASE

Population (Allele,Count): caucasian (T,107|C,9)

SNP Type: HUMAN-MOUSE SYNTENIC REGION; SILENT MUTATION

20 Gene Number: 9

Celera Gene: hCG20335 - 65000099361015

Gene Symbol: APOC4

Protein Name: apolipoprotein C-IV

Celera Genomic Axis: GA_x5YUV32VTL1 (7646587..7670232)

25 Chromosome: 19

OMIM NUMBER: 600745

OMIM Information:

Genomic Sequence (SEQ ID NO: 156):

30

SNP Information

Context (SEQ ID NO: 186):

GCATGGAGGGAAAGGGAGAAGGGGATTCTAGGGTCCCAGCCTACCCAAGTTGCCCT
35 CTGGTTCCACCTAGCATGCCAGCCAGAGGCCAGGAAGGAACCC

Y

GAGCCCCCACCAAAGCTAAAGATGAGTCGCTGGAGCCTGGTGAGGGGCAGGATGA
AGGAGCTGCTGGAGACAGTGGTGAACAGGACCAGAGACGGGTGG

Celera SNP ID: hCV1841831

40 Public SNP ID: rs1132899

SNP in Genomic Sequence: SEQ ID NO: 156

SNP Position Genomic: 12542

SNP Source: Applera

Population (Allele,Count): caucasian (C,7|T,11) african american (C,15|T,13) total
45 (C,22|T,24)

SNP Type: MISSENSE MUTATION; TRANSCRIPTION FACTOR BINDING
SITE

SNP Source: dbSNP; Celera; HapMap; HGBASE

Population (Allele,Count): caucasian (C,51|T,65)

50 SNP Type: MISSENSE MUTATION; TRANSCRIPTION FACTOR BINDING
SITE

Gene Number: 10
Celera Gene: hCG22337 - 146000220683873
Gene Symbol: SERPINA13
5 Protein Name: serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiprotei
nase, antitrypsin), member 13
Celera Genomic Axis: GA_x5YUV32W0QP (41908319..41977785)
Chromosome: 14
OMIM NUMBER:
10 OMIM Information:

Genomic Sequence (SEQ ID NO: 157):

SNP Information

15 Context (SEQ ID NO: 187):
CATAATCTCCGATTTTATTTTCTGTTTTGCAGACGAGAAAACAGCAAGGAAGGCTC
AACTCGGGCTTGCCCGGGGGGCTCTTCCCCTTCGGCGTGGAAG
W
20 TTCAATGCGAGGAAGCAAAGAGCCTAAAATACGAACACCCAGGAAATTAAGCAGGG
GAAGGGGCCCTTCTGAGACAGTAAAGAAACATACATCCGAAGAC
Celera SNP ID: hCV2188996
Public SNP ID: rs11622883
SNP in Genomic Sequence: SEQ ID NO: 157
25 SNP Position Genomic: 58090
SNP Source: dbSNP; Celera; HapMap
Population (Allele, Count): caucasian (T, 62 | A, 58)
SNP Type: MISSENSE MUTATION; ESS

30 Gene Number: 11
Celera Gene: hCG25596 - 146000220733557
Gene Symbol:
Protein Name:
Celera Genomic Axis: GA_x54KRFTF0F9 (28582642..28603669)
35 Chromosome: 6
OMIM NUMBER:
OMIM Information:

Genomic Sequence (SEQ ID NO: 158):

40 SNP Information

Context (SEQ ID NO: 188):
45 ATGGAGCTCAGCCCTCCACCTCCACAGCTCAAGAGCAAGATGATGTTCTCATAGTTG
ATTAGGATGAAGAAGATTCTTCAAATAATGCCGACGTCAGTGA
R
GAAGAGAGAAGCCGCAAGAAGAAATTAGATGAGAAAGAGAATCTCAGTGCAAAGA
GGTCACGTATAGAACAGAAGGAAGAGCTTGATGATGTCATAGCAT
Celera SNP ID: hCV25741966
50 Public SNP ID: rs6907175
SNP in Genomic Sequence: SEQ ID NO: 158

- SNP Position Genomic: 10236
 SNP Source: Applera
 Population (Allele,Count): caucasian (A,24|G,16) african american (A,29|G,7) total (A,53|G,23)
- 5 SNP Type: UTR3; INTRON
 SNP Source: dbSNP; HapMap
 Population (Allele,Count): caucasian (G,59|A,61)
 SNP Type: UTR3; INTRON
- 10 Gene Number: 12
 Celera Gene: hCG25621 - 104000116735770
 Gene Symbol: UBD
 Protein Name: ubiquitin D
 Celera Genomic Axis: GA_x5YUV32W6W6 (2583709..2607910)
- 15 Chromosome: 6
 OMIM NUMBER: 606050
 OMIM Information:
- Genomic Sequence (SEQ ID NO: 159):
- 20 SNP Information
- Context (SEQ ID NO: 189):
 CTCATCTCATTCTCATTAATTTTGGGAAATCATCAGAAGATGTGTTTCGTTGAGTAAGA
 25 GATTAAAAGAAATAAGCTTTTTGACCCCTGCCAACACCCCAT
 S
 CCCAGGGTGGTCACCCTCCAATAACAATAAGATGCCAGGAAGAGTAAGTTGCCCTTTC
 TGATGCCGTAATCTGCCATCATCTTCCCATCTTCCAGTCTCTT
- 30 Celera SNP ID: hCV588691
 Public SNP ID: rs444013
 SNP in Genomic Sequence: SEQ ID NO: 159
 SNP Position Genomic: 10341
 SNP Source: Applera
 Population (Allele,Count): caucasian (C,10|G,8) african american (C,16|G,6) total (C,26|G,14)
- 35 SNP Type: UTR3
 SNP Source: dbSNP; Celera; HapMap; HGBASE
 Population (Allele,Count): caucasian (C,49|G,55)
 SNP Type: UTR3
- 40 Gene Number: 13
 Celera Gene: hCG28649 - 67000129132473
 Gene Symbol: AGC1
 Protein Name: aggrecan 1 (chondroitin sulfate proteoglycan 1, large aggregating proteoglycan, antigen identified by monoclonal antibody A0122)
- 45 Celera Genomic Axis: GA_x5YUV32W6GH (4231607..4322592)
 Chromosome: 15
 OMIM NUMBER: 155760
 OMIM Information:
- 50 Genomic Sequence (SEQ ID NO: 160):

SNP Information

Context (SEQ ID NO: 190):

5 GAAGTTCTAGAGATTTCTGTCTCTGGAGTAGAGGACATCAGTGGGCTTCCTTCTGGA
 GAGGTTGTAGAGACTTCTGCCTCTGGAATAGAGGATGTCAGTG
 M
 ACTTCCTTCAGGAGAAGGTCTAGAGACCTCTGCTTCTGGAGTAGAGGACCTCAGCAG
 GCTCCCTTCTGGAGAAGAAGTTCTAGAGATTTCTGCCTCTGGA
 10 Celera SNP ID: hCV1834284
 Public SNP ID: rs2882676
 SNP in Genomic Sequence: SEQ ID NO: 160
 SNP Position Genomic: 62869
 SNP Source: dbSNP; Celera; HapMap; HGBASE
 Population (Allele,Count): caucasian (A,43|C,77)
 15 SNP Type: NONSENSE MUTATION; MISSENSE MUTATION;
 TRANSCRIPTION FACTOR BINDING SIT
 E; HUMAN-MOUSE SYNTENIC REGION

Gene Number: 14

20 Celera Gene: hCG37868 - 84000314283049
 Gene Symbol: PCK1
 Protein Name: phosphoenolpyruvate carboxykinase 1 (soluble)
 Celera Genomic Axis: GA_x5YUV32VYHA (6713662..6739039)
 Chromosome: 20
 25 OMIM NUMBER: 261680
 OMIM Information: Hypoglycemia due to PCK1 deficiency (1)

Genomic Sequence (SEQ ID NO: 161):

30 SNP Information

Context (SEQ ID NO: 191):

GAAATAGATCTTGAGGGCCTACATTCCAACCTCTGGGCTGAAGTACCAACCTCGGGG
 35 AGAAGGAAACAAAGATCACAATAAAGAATCTTGTCCCCAACAG
 R
 TTCTGGGTATAACCAACCCTGAGGGTGAGAAGAAGTACCTGGCGGCCGCATTTCCCA
 GCGCCTGCGGGAAGACCAACCTGGCCATGATGAACCCCAGCCT
 Celera SNP ID: hCV2508722
 Public SNP ID: rs8192708
 40 SNP in Genomic Sequence: SEQ ID NO: 161
 SNP Position Genomic: 12489
 SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (A,108|G,12)
 SNP Type: MISSENSE MUTATION; ESS; HUMAN-MOUSE SYNTENIC
 45 REGION; INTRON

Gene Number: 15

Celera Gene: hCG39378 - 30000023325762
 Gene Symbol: CTSS
 50 Protein Name: cathepsin S
 Celera Genomic Axis: GA_x5YUV32VSL (10613850..10669294)

Chromosome: 1
OMIM NUMBER: 116845
OMIM Information:

5 Genomic Sequence (SEQ ID NO: 162):

SNP Information

Context (SEQ ID NO: 192):

10 CTTGTATTGTTTGCCATAGGTTTTCTTCCAGAGATGCCAGTGGTGATCCAGGGTAGGA
TCTTTATGCAACTGTGCCACTGCAGAGGAGCACACCAAGAGC
R
CACAAACCAGCCGTTTCATTCTGTGTAAAGGAAAGGTAACATAGGAAGTGTTGCTATT
AGCTGCATATGTTGTGACCACATGTTTTTCATAAGAGAAAATAA

15 Celera SNP ID: hCV15746640

Public SNP ID:

SNP in Genomic Sequence: SEQ ID NO: 162

SNP Position Genomic: 44231

SNP Source: CDX_Alzheimer

20 Population (Allele,Count): caucasian (A,104|G,16)

SNP Type: MISSENSE MUTATION; HUMAN-MOUSE SYNTENIC REGION;
INTRON

Gene Number: 16

25 Celera Gene: hCG42001 - 103000085248952

Gene Symbol: TNK1

Protein Name: tyrosine kinase, non-receptor, 1

Celera Genomic Axis: GA_x5YUV32W2JD (8122109..8151222)

Chromosome: 17

30 OMIM NUMBER: 608076

OMIM Information:

Genomic Sequence (SEQ ID NO: 163):

35 SNP Information

Context (SEQ ID NO: 193):

40 TAGGAACTCCTTCTCCAGACATGCTTCCTGAGGCTGGCTCTCTGTGGCTACTGAAGCT
GCTCCGGGACATCCAGTTGGCCAGTTTTACTGGCCCATCCT
W
GAGGAGCTTAATGTCACTCGGCCAGAGCACTTCGACTTTGTAAAGCCTGAGGACCTG
GACGGCATTGGCATGGGCCGGCCTGGTGAGGGACCCCTGCCCC

Celera SNP ID: hCV8717889

Public SNP ID: rs1554948

45 SNP in Genomic Sequence: SEQ ID NO: 163

SNP Position Genomic: 11921

SNP Source: dbSNP; HapMap; HGBASE

Population (Allele,Count): caucasian (T,57|A,63)

50 SNP Type: NONSENSE MUTATION; SILENT MUTATION

Context (SEQ ID NO: 194):

- GGTGTGGTTGTCATATCTGAGCATCTCCAGTGTGTGTGGTAGGCAAACATGTCTCAT
GTTGGAGGAAGGGAGATGTCTCATGTTGGAGGAAGGGAGATGT
Y
GTCTCAGGTGTGTATGAGTGAGTTTGTGCCACTGCGTAGCTCAGGTCTAAAAGGGCA
5 GTGTTTCTGGGTGTCCGGGTGTGTGATGTGCCGTGATATTTCA
Celera SNP ID: hCV1247823
Public SNP ID: rs2269762
SNP in Genomic Sequence: SEQ ID NO: 163
SNP Position Genomic: 11425
10 Related Interrogated SNP: hCV8717889 (Power=.8)
SNP Source: dbSNP; Celera; HapMap; ABI_Val; HGBASE
Population (Allele,Count): caucasian (C,79|T,41)
SNP Type: INTRON
- 15 Context (SEQ ID NO: 195):
CTGCCTCCGGTGGAGGAGAGGCAATGACTGTGGACCTCTTTTGGCTTGAGGGGTCCC
TATGCCCCCATAGCTCAGGTATGCTGTGTGCACACTGTACCTC
R
ATGTCTGTGTGTTGTGTTTCATGTGGACCTTTGTGTCTTAGCCGTGTGGGGCAGTGTAT
20 CTAGGTGTTTGGAGGCAGTGAAGTCCCATGTCCCATGGATGT
Celera SNP ID: hCV1247820
Public SNP ID: rs12948090
SNP in Genomic Sequence: SEQ ID NO: 163
SNP Position Genomic: 10697
25 Related Interrogated SNP: hCV8717889 (Power=.9)
SNP Source: dbSNP; Celera; HapMap
Population (Allele,Count): caucasian (G,53|A,57)
SNP Type: INTRON
- 30 Context (SEQ ID NO: 196):
GACCTGGTCTCTCTAGGGCCTACCCTGAGCTCACCATCTGAAGGAGAGTGCCATCAT
CCTTAGGAACCTCCTTCTCCAGACATGCTTCCTGAGGCTGGCTC
Y
CTGTGGCTACTGAAGCTGCTCCGGGACATCCAGTTGGCCCAGTTTACTGGCCCATC
35 CTTGAGGAGCTTAATGTCACTCGGCCAGAGCACTTCGACTTTG
Celera SNP ID: hCV1247824
Public SNP ID: rs1554947
SNP in Genomic Sequence: SEQ ID NO: 163
SNP Position Genomic: 11861
40 Related Interrogated SNP: hCV8717889 (Power=.8)
SNP Source: dbSNP; Celera; HapMap; HGBASE
Population (Allele,Count): caucasian (C,73|T,43)
SNP Type: MISSENSE MUTATION; SILENT MUTATION
- 45 Context (SEQ ID NO: 197):
GGAGGCAGAGGTTGCAGTGAGCTGATATCATGTCACTGCACTCCAGCCTGGACTACA
GAGCAGGACTCCATCTAAACAAACAAACAAAAACCACAATGA
R
ATATCACTATGGCACTGACTAGAAATGGCTAAAATTAGAAAGTGTTGCTGAAGATGTA
50 GAGCAACTGGAACCTTCATACTCTGCTGGGAAGACTTTGAAAT
Celera SNP ID: hCV1477271

Public SNP ID: rs11655156
 SNP in Genomic Sequence: SEQ ID NO: 163
 SNP Position Genomic: 3458
 Related Interrogated SNP: hCV8717889 (Power=.6)
 5 SNP Source: dbSNP; Celera; HapMap; ABI_Val
 Population (Allele,Count): caucasian (G,88|A,32)
 SNP Type: INTERGENIC; UNKNOWN; REPEATS

Context (SEQ ID NO: 198):

10 CCAAAAAAAAAAAAAAAAAAGAAAGAAAGAAAAAAAAAGAAACAAAAAGAAAGCTGGGCA
 CGGTGGCACACACCTGTGGTGTAGTCCCAGTTCCTCAGGAGGTTGA
 K
 TTGAGGTGAGAGAATCATTGTGTTGCCAGGTTATTCTCAAACCTCCTGGCCTCAAGCG
 ATCCTCCCGCCTTGGCCTCCCAAATGATGGGATTACAGGCAT

15 Celera SNP ID: hCV26960066
 Public SNP ID: rs4325594
 SNP in Genomic Sequence: SEQ ID NO: 163
 SNP Position Genomic: 1428
 Related Interrogated SNP: hCV8717889 (Power=.6)
 20 SNP Source: Celera; HapMap; ABI_Val; HGBASE
 Population (Allele,Count): caucasian (T,88|G,32)
 SNP Type: INTERGENIC; UNKNOWN; REPEATS

Context (SEQ ID NO: 199):

25 TATTATATGCCACACTGTCCTAGGTACTTAGAACACAAGTATGAACAAGCCAGGCC
 CAGCCCCTACCCTGTGGAGCTACAGAACCTCTAACAGGGATCG
 M
 AGATGTGCAGTAACAATTATAGTGTGATGAGCGATCCTGGTTGGAGGGCAAAAGAA
 TCGGCAAGAAAACAATACCATAGAACAGCATTATGAAGATTTTC

30 Celera SNP ID: hCV30677672
 Public SNP ID: rs7219773
 SNP in Genomic Sequence: SEQ ID NO: 163
 SNP Position Genomic: 8737
 Related Interrogated SNP: hCV8717889 (Power=.9)
 35 SNP Source: dbSNP; HapMap
 Population (Allele,Count): caucasian (C,55|A,65)
 SNP Type: INTERGENIC; UNKNOWN

Gene Number: 17
 40 Celera Gene: hCG2044039 - 30000063654100
 Gene Symbol: MGC35402
 Protein Name: hypothetical protein MGC35402
 Celera Genomic Axis: GA_x5YUV32VY4T (2771485..2810693)
 Chromosome: 19

45 OMIM NUMBER:
 OMIM Information:

Genomic Sequence (SEQ ID NO: 164):

50 SNP Information

Context (SEQ ID NO: 200):

TTTCAACCAAAGGTCTTGTTATGCTCACTACACTTGTAAGGTTTCCCTCCAGTATGAA
GTCTACGATAGCATGTAAAGGATGACGTCTGACTGAAGGTCT

Y

5 GCCACACTCATAACAACACTGTAAGGTTTCTTACCAGTGTGACATCTATTATGGCATA
AAGGAATCAATTCTGATTAAAGACCTTGTCACATACCTCACAT

Celera SNP ID: hCV29207239

Public SNP ID: rs8110693

SNP in Genomic Sequence: SEQ ID NO: 164

10 SNP Position Genomic: 15337

Related Interrogated SNP: hCV3084827 (Power=.7)

SNP Source: dbSNP

Population (Allele,Count): caucasian (T,113|C,1)

SNP Type: TRANSCRIPTION FACTOR BINDING SITE; UTR3; INTRON

15

Gene Number: 18

Celera Gene: hCG16801

Gene Symbol: NMBR

Protein Name: neuromedin B receptor

20 Celera Genomic Axis: GA_x54KRFTF0F9 (28557584..28608029)

Chromosome: 6

OMIM NUMBER: 162341

OMIM Information:

25 Genomic Sequence (SEQ ID NO: 165):

SNP Information

Context (SEQ ID NO: 201):

30 ATGGAGCTCAGCCCTCCACCTCCACAGCTCAAGAGCAAGATGATGTTCTCATAGTTG
ATTAGGATGAAGAAGATTCTTCAAATAATGCCGACGTCAGTGA

R

GAAGAGAGAAGCCGCAAGAAGAAATTAGATGAGAAAGAGAATCTCAGTGCAAAGA
GGTCACGTATAGAACAGAAGGAAGAGCTTGATGATGTCATAGCAT

35 Celera SNP ID: hCV25741966

Public SNP ID: rs6907175

SNP in Genomic Sequence: SEQ ID NO: 165

SNP Position Genomic: 14596

SNP Source: Applera

40 Population (Allele,Count): caucasian (A,24|G,16) african american (A,29|G,7) total
(A,53|G,23)

SNP Type: UTR3; INTRON

SNP Source: dbSNP; HapMap

Population (Allele,Count): caucasian (G,59|A,61)

45 SNP Type: UTR3; INTRON

Gene Number: 19

Celera Gene: hCG16926

Gene Symbol: THEM5

50 Protein Name: hypothetical protein FLJ37964

Celera Genomic Axis: GA_x5YUV32VSLs (9526929..9561424)

Chromosome: 1
 OMIM NUMBER:
 OMIM Information:

5 Genomic Sequence (SEQ ID NO: 166):

SNP Information

Context (SEQ ID NO: 202):

10 CTTGTATTGTTTGCCATAGGTTTTCTTCCAGAGATGCCAGTGGTGATCCAGGGTAGGA
 TCTTTATGCAACTGTGCCACTGCAGAGGAGCACACCAAGAGC
 R
 CACAAACCAGCCGTTTCATTCTGTGTAAGGAAAGGTAACATAGGAAGTGTTGCTATT
 AGCTGCATATGTTGTGACCACATGTTTTTCATAAGAGAAAATAA
 15 Celera SNP ID: hCV15746640
 Public SNP ID:
 SNP in Genomic Sequence: SEQ ID NO: 166
 SNP Position Genomic: 1063641
 SNP Source: CDX_Alzheimer
 20 Population (Allele,Count): caucasian (A,104|G,16)
 SNP Type: MISSENSE MUTATION; HUMAN-MOUSE SYNTENIC REGION;
 INTRON

Context (SEQ ID NO: 203):

25 AGCTAGCTTCAGGCAGAGGGGACTCAAGGTTGTCCCCACACTGGCTAGGCAGCAAA
 GCTTGGTCTGCTGGCTCCCGTTCTCTAGCCTCCCCATCAGCATC
 W
 CTTGCATCTGTGTCCTGAGGGTGCTCATCTGGGCCTGCCAGAACCTCATGGTTTTTCAG
 GGATGACTGCTTCTGCAGTCCCGGCCTTACCAGGGGCTGGAT
 30 Celera SNP ID: hCV22274641
 Public SNP ID:
 SNP in Genomic Sequence: SEQ ID NO: 166
 SNP Position Genomic: 825607
 SNP Source: Applera
 35 Population (Allele,Count): caucasian (T,27|A,1) african american (T27,|A,7) total (T,54|A,8)
 SNP Type: MISSENSE MUTATION; ESS; HUMAN-MOUSE SYNTENIC
 REGION; INTRON

Gene Number: 20

40 Celera Gene: hCG1773556
 Gene Symbol:
 Protein Name:
 Celera Genomic Axis: GA_x5YUV32VY4T (2220042..2256373)
 Chromosome: 19
 45 OMIM NUMBER:
 OMIM Information:

Genomic Sequence (SEQ ID NO: 167):

50 SNP Information

Context (SEQ ID NO: 204):

GGGATTATTTTTTATTTTTTATTTTTTTTAGACAGAGTCTTGCTCTGTACAGGCTGTA
GTGAAGTGGCACAATCTCAGCTCACTGCAATCTCCGCCTCC

Y

5 GGTTC AAGTGATTCTTCTAGTTCAGGGATTTTTTATGGAGGCTTCATCACATAGGTGT
GATTGATATTATCTCAGTCTCCAGCCCCTCTTCCCTTCCTGG

Celera SNP ID: hCV3066768

Public SNP ID: rs162129

SNP in Genomic Sequence: SEQ ID NO: 167

10 SNP Position Genomic: 18010

Related Interrogated SNP: hCV905013 (Power=.51)

SNP Source: dbSNP; Celera; HapMap; HGBASE

Population (Allele,Count): caucasian (T,44|C,76)

SNP Type: INTRON; REPEATS

15

Gene Number: 21

Celera Gene: hCG2010846

Gene Symbol: LOC51233

20 Protein Name: hypothetical protein LOC51233

Celera Genomic Axis: GA_x5YUV32VU5C (2050760..2089798)

Chromosome: 22

OMIM NUMBER:

OMIM Information:

25

Genomic Sequence (SEQ ID NO: 168):

SNP Information

30 Context (SEQ ID NO: 205):

CAGGTTTCAGAAAGCTTCTCCCTGACATCCGTGGAGCTGCAGATGCTGACCAACTCGT
GTGTGAAACTCCAGACTGTCCACAGCATTCGCTGACCATCAA

Y

AAGGAAGGTGGGCCCCCCCCGTTTCCGTGTACAGGGCACCTGCAGGGAGGGCAGGCA

35 GCTAGCCTGAAGGCTGATCCCCCCTTCCTGTTAGCACTTTTGAT

Celera SNP ID: hCV2447571

Public SNP ID: rs9608099

SNP in Genomic Sequence: SEQ ID NO: 168

SNP Position Genomic: 338247

40 SNP Source: Applera

Population (Allele,Count): caucasian (C,6|T,26) african american (C,16|T,12) total
(C,22|T,38)

SNP Type: HUMAN-MOUSE SYNTENIC REGION; UTR5; SILENT
MUTATION; INTRON

45 SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (C,36|T,84)

SNP Type: HUMAN-MOUSE SYNTENIC REGION; UTR5; SILENT
MUTATION; INTRON

50 Gene Number: 22

Celera Gene: hCG20334

Gene Symbol: APOC2
 Protein Name: apolipoprotein C-II
 Celera Genomic Axis: GA_x5YUV32VTL1 (7650331..7673907)
 Chromosome: 19
 5 OMIM NUMBER: 608083
 OMIM Information: Hyperlipoproteinemia, type Ib, 207750 (3)

Genomic Sequence (SEQ ID NO: 169):

10 SNP Information

Context (SEQ ID NO: 206):

GCATGGAGGGAAAGGGAGAAAGGGGATTCTAGGGTCCCAGCCTACCCAAGTTGCCCT
 CTGGTTCCACCTAGCATGCCAGCCAGAGGCCAGGAAGGAACCC

15 Y

GAGCCCCCACCAAAGCTAAAGATGAGTCGCTGGAGCCTGGTGAGGGGCAGGATGA
 AGGAGCTGCTGGAGACAGTGGTGAACAGGACCAGAGACGGGTGG

Celera SNP ID: hCV1841831

Public SNP ID: rs1132899

20 SNP in Genomic Sequence: SEQ ID NO: 169

SNP Position Genomic: 8798

SNP Source: Applera

Population (Allele,Count): caucasian (C,7|T,11) african american (C,15|T,13) total
 (C,22|T,24)

25 SNP Type: MISSENSE MUTATION; TRANSCRIPTION FACTOR BINDING
 SITE

SNP Source: dbSNP; Celera; HapMap; HGBASE

Population (Allele,Count): caucasian (C,51|T,65)

30 SNP Type: MISSENSE MUTATION; TRANSCRIPTION FACTOR BINDING
 SITE

Gene Number: 23

Celera Gene: hCG22130

Gene Symbol: PVRL2

35 Protein Name: poliovirus receptor-related 2 (herpesvirus entry mediator B)

Celera Genomic Axis: GA_x5YUV32VTL1 (7550655..7613594)

Chromosome: 19

OMIM NUMBER: 600798

OMIM Information:

40

Genomic Sequence (SEQ ID NO: 170):

SNP Information

45 Context (SEQ ID NO: 207):

GTTGGAGTGGAGTGTGACAGCGTTTCTCTTCTCCAGAGCTGTTTCCCATTCAGATGGA
 GGGTGTCAAGCTCACAGTCAACAAAGGGTTGAGTAACCATTT

Y

CAGGTGAGCCTTCCTGGTGTCTTACCCACCAGAGATCGTCCCCGCCGTCCCCCTCCC

50 TGCATCTGCACACTCGGCCCAATTACTCCTCCCTCAAGAGCT

Celera SNP ID: hCV3084827

Public SNP ID: rs157581
 SNP in Genomic Sequence: SEQ ID NO: 170
 SNP Position Genomic: 56173
 SNP Source: dbSNP; Celera; HGBASE
 5 Population (Allele,Count): caucasian (C,28|T,92)
 SNP Type: TRANSCRIPTION FACTOR BINDING SITE; HUMAN-MOUSE
 SYNTENIC REGION; SILENT
 MUTATION

10 Gene Number: 24
 Celera Gene: hCG41189
 Gene Symbol:
 Protein Name:
 Celera Genomic Axis: GA_x54KRFTF114 (2412552..2582896)
 15 Chromosome: 10
 OMIM NUMBER:
 OMIM Information:

Genomic Sequence (SEQ ID NO: 171):

20

SNP Information

Context (SEQ ID NO: 208):

25 TCACATTCCATGCCAAATGAGGCAAGCAGTCGGGCAGCCTGTGAAAGGCCAGATAT
 CCGGCTATCATATGCAGGCCAGTATCCAGAGCGGGTCATACAAA
 Y
 AGAGCCTGACACGACCACATAGACTCAATGCTTCCTGCACGCTGGAATAGCCTACAG
 CTCGGGAAGGATACATATTAAATATATTTGAGTGTACTGGAG
 Celera SNP ID: hCV10057645
 30 Public SNP ID: rs11016976
 SNP in Genomic Sequence: SEQ ID NO: 171
 SNP Position Genomic: 48464
 SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (T,104|C,16)
 35 SNP Type: TRANSCRIPTION FACTOR BINDING SITE; HUMAN-MOUSE
 SYNTENIC REGION; INTRON

Gene Number: 25
 Celera Gene: hCG2036867
 40 Gene Symbol: ZNF578
 Protein Name:
 Celera Genomic Axis: GA_x5YUV32VY4T (2343316..2384971)
 Chromosome: 19
 OMIM NUMBER:
 45 OMIM Information:

Genomic Sequence (SEQ ID NO: 172):

SNP Information

50

Context (SEQ ID NO: 209):

CTTTTGCTTAGTGATTTGGGGCTCCATATTCATTTTCCTTTACACAGCCCAACCCTC
ACACCACCTGTGGTCTATCTTGCCTGTTCTGTGTGATTACTG
Y

TGTGCCTTTGTCATCCACTGTCTGGATCCTAGTGAGAACAAGGCCCAAGGGGAGGG
5 CAGAGCCAAAACATGGGGCTGTGCTGGGCTCGCCCCGTCTGAG

Celera SNP ID: hDV70782148

Public SNP ID: rs16983943

SNP in Genomic Sequence: SEQ ID NO: 172

SNP Position Genomic: 35463

10 Related Interrogated SNP: hCV3084827 (Power=.9)

SNP Source: dbSNP; HapMap

Population (Allele,Count): caucasian (T,102|C,6)

SNP Type: INTERGENIC; UNKNOWN

15 Context (SEQ ID NO: 210):

ACAAGAAAACCTATGTTTCAGTTTGTGTAGGGTCAACCCTGTAAGAGATATGGTTTT
CATCAATCTTATTTAGGGAGCTTGTGAGGTTGAATGAGGTTTC

M

ATGGTTTAGTTCTACTGTCATTTTCCTTCAGACACAGTGGCTTTCAACTTGTTAGTTTCT

20 ATGTTGTTTTTTTTTTTGTAGTCAGAGTCTCACTCTGTTTCC

Celera SNP ID: hCV31444359

Public SNP ID: rs7252479

SNP in Genomic Sequence: SEQ ID NO: 172

SNP Position Genomic: 23010

25 Related Interrogated SNP: hCV3084827 (Power=.9)

SNP Source: dbSNP; HapMap

Population (Allele,Count): caucasian (C,114|A,6)

SNP Type: INTRON

30 Gene Number: 26

Celera Gene: hCG2036874

Gene Symbol: ZNF610

Protein Name: zinc finger protein 610

Celera Genomic Axis: GA_x5YUV32VY4T (2186456..2237914)

35 Chromosome: 19

OMIM NUMBER:

OMIM Information:

Genomic Sequence (SEQ ID NO: 173):

40

SNP Information

Context (SEQ ID NO: 211):

AGGTTGCCTCATTAGAGCAGATCACACTCGTACCACCCAGGACATCTCAAGGTCTTC
45 AGAGCTCTGTGTCAGGAACCGGGGAAGAGACCAATGTATGTAT

K

TATTCTTATTCCACAAGGCCCTCAGCCATTCTTCAGTTGTGCTTCATCCACACTTTA
CTGGTGTGTGTCACAGGGGAGTAGGTGGCTCTTCTGAGATGC

Celera SNP ID: hCV2931396

50 Public SNP ID: rs2560983

SNP in Genomic Sequence: SEQ ID NO: 173

SNP Position Genomic: 29843
 Related Interrogated SNP: hCV3084827 (Power=.6)
 SNP Source: dbSNP; Celera; HapMap; HGBASE
 Population (Allele,Count): caucasian (G,115|T,3)
 5 SNP Type: TRANSCRIPTION FACTOR BINDING SITE; INTRON

Gene Number: 27
 Celera Gene: hCG2044059
 Gene Symbol: ZNF528
 10 Protein Name: zinc finger protein 528
 Celera Genomic Axis: GA_x5YUV32VY4T (2248450..2289007)
 Chromosome: 19
 OMIM NUMBER:
 OMIM Information:

15 Genomic Sequence (SEQ ID NO: 174):

SNP Information

20 Context (SEQ ID NO: 212):
 TGTGTATTTGGAAGAAGGGGTCAGGGGGCTCCTTCTAGTGAACAAGGGCCCTGAGCT
 TTAGAGCCCTTTGTATTTTATTGGGTAAAGGAGATGGTGGCAG
 K
 GGGAGGGGGTGGTTATTGGTCAGCGGCTTGATTTACAACAGGTTTGCAAGACTGCAT
 25 TCTTTGAACAATAGACTGTAGATGTCCCAGTAGATAATTTC
 Celera SNP ID: hCV27899309
 Public SNP ID: rs4801914
 SNP in Genomic Sequence: SEQ ID NO: 174
 SNP Position Genomic: 26622
 30 Related Interrogated SNP: hCV3084827 (Power=.6)
 SNP Source: dbSNP; HGBASE
 Population (Allele,Count): caucasian (G,16|T,100)
 SNP Type: TRANSCRIPTION FACTOR BINDING SITE; INTRON; REPEATS

35 Context (SEQ ID NO: 213):
 GAAGAATTTTCTTAGTACAGAACAAAATGGAGTCTCCTGTGTCTACTTCTTTCTACA
 CAGAAACAATCCGATCTCTCTTTCTTTCCCCACAAAAGTTG
 M
 CTGAGATACGTTTTATCCTAGTTCAATGAAAATTATTGTACAAAAGTGTTTATTATAA
 40 AATCAACATGATTTGTTTATTGTTTATTGATTATATAATTAT
 Celera SNP ID: hDV71600369
 Public SNP ID: rs16983796
 SNP in Genomic Sequence: SEQ ID NO: 174
 SNP Position Genomic: 34074
 45 Related Interrogated SNP: hCV3084827 (Power=.9)
 SNP Source: dbSNP; HapMap
 Population (Allele,Count): caucasian (C,115|A,5)
 SNP Type: INTERGENIC; UNKNOWN

50 Gene Number: 28
 Celera Gene: Chr17:7275229..7315229

Gene Symbol:

Protein Name:

Celera Genomic Axis: GA_x5YUV32W2JD (8136313..8176313)

Chromosome: 17

5 OMIM NUMBER:

OMIM Information:

Genomic Sequence (SEQ ID NO: 175):

10 SNP Information

Context (SEQ ID NO: 214):

TAGGAACTCCTTCTCCAGACATGCTTCCTGAGGCTGGCTCTCTGTGGCTACTGAAGCT
GCTCCGGGACATCCAGTTGGCCAGTTTTACTGGCCCATCCT

15 W

GAGGAGCTTAATGTCACTCGGCCAGAGCACTTCGACTTTGTAAAGCCTGAGGACCTG
GACGGCATTGGCATGGGCCGGCCTGGTGAGGGACCCCTGCCCC

Celera SNP ID: hCV8717889

Public SNP ID: rs1554948

20 SNP in Genomic Sequence: SEQ ID NO: 175

SNP Position Genomic: 37012

SNP Source: dbSNP; HapMap; HGBASE

Population (Allele,Count): caucasian (T,57|A,63)

SNP Type: NONSENSE MUTATION; SILENT MUTATION

25

Context (SEQ ID NO: 215):

GGTGTGGTTGTCATATCTGAGCATCTCCAGTGTGTGTGGTAGGCAAACATGTCTCAT
GTTGGAGGAAGGGAGATGTCTCATGTTGGAGGAAGGGAGATGT

Y

30 GTCTCAGGTGTGTATGAGTGAGTTTGTGCCACTGCGTAGCTCAGGTCTAAAAGGGCA

GTGTTTCTGGGTGTCCGGGTGTGTGATGTGCCGTGATATTTCA

Celera SNP ID: hCV1247823

Public SNP ID: rs2269762

SNP in Genomic Sequence: SEQ ID NO: 175

35 SNP Position Genomic: 36516

Related Interrogated SNP: hCV8717889 (Power=.8)

SNP Source: dbSNP; Celera; HapMap; ABI_Val; HGBASE

Population (Allele,Count): caucasian (C,79|T,41)

SNP Type: INTRON

40

Context (SEQ ID NO: 216):

CTGCCTCCGGTGGAGGAGAGGCAATGACTGTGGACCTCTTTTGGCTTGAGGGGTCCC
TATGCCCCCATAGCTCAGGTATGCTGTGTGCACACTGTACCTC

R

45 ATGTCTGTGTGTTGTGTTTCATGTGGACCTTTGTGTCTTAGCCGTGTGGGGCAGTGTAT

CTAGGTGTTTGGAGGCAGTGAAGTCCCATGTCCCATGGATGT

Celera SNP ID: hCV1247820

Public SNP ID: rs12948090

SNP in Genomic Sequence: SEQ ID NO: 175

50 SNP Position Genomic: 35788

Related Interrogated SNP: hCV8717889 (Power=.9)

SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (G,53|A,57)
 SNP Type: INTRON

5 Context (SEQ ID NO: 217):
 GACCTGGTCTCTCTAGGGCCTACCCTGAGCTCACCATCTGAAGGAGAGTGCCATCAT
 CCTTAGGAACTCCTTCTCCAGACATGCTTCCTGAGGCTGGCTC
 Y

10 CTGTGGCTACTGAAGCTGCTCCGGGACATCCAGTTGGCCCAGTTTTACTGGCCCCATC
 CTTGAGGAGCTTAATGTCACCTCGGCCAGAGCACTTCGACTTTG
 Celera SNP ID: hCV1247824
 Public SNP ID: rs1554947

SNP in Genomic Sequence: SEQ ID NO: 175
 SNP Position Genomic: 36952
 15 Related Interrogated SNP: hCV8717889 (Power=.8)
 SNP Source: dbSNP; Celera; HapMap; HGBASE
 Population (Allele,Count): caucasian (C,73|T,43)
 SNP Type: MISSENSE MUTATION; SILENT MUTATION

20 Context (SEQ ID NO: 218):
 GGAGGCAGAGGTTGCAGTGAGCTGATATCATGTCACCTGCACTCCAGCCTGGACTACA
 GAGCAGGACTCCATCTAAACAAACAAACAAAAAACCACAATGA
 R

25 ATATCACTATGGCACTGACTAGAAATGGCTAAAATTAGAAAGTGTTGCTGAAGATGTA
 GAGCAACTGGAACCTTCATACTCTGCTGGGAAGACTTTGAAAT
 Celera SNP ID: hCV1477271
 Public SNP ID: rs11655156

SNP in Genomic Sequence: SEQ ID NO: 175
 SNP Position Genomic: 28549
 30 Related Interrogated SNP: hCV8717889 (Power=.6)
 SNP Source: dbSNP; Celera; HapMap; ABI_Val
 Population (Allele,Count): caucasian (G,88|A,32)
 SNP Type: INTERGENIC; UNKNOWN; REPEATS

35 Context (SEQ ID NO: 219):
 GGGCCAGGGAGTACCAAAGTGGGAGGGCCTCATGGATGGGTTTCTAGAAGAAGAGTGC
 TTATTGGTGAGGTTTGGAGATTGCCATTCCCATCACTTCCGGCA
 Y

40 AGGAGCTCTTAATTACAAACCTCTCCTTTGCTTCACTCAATTCAGATGACTCTGTCTC
 ATCCATTTCGTTTCTAGGAAATGTCACCTGAGCACGTAGCATGTAA
 Celera SNP ID: hCV26960061
 Public SNP ID: rs4796409

SNP in Genomic Sequence: SEQ ID NO: 175
 SNP Position Genomic: 20001
 45 Related Interrogated SNP: hCV8717889 (Power=.7)
 SNP Source: dbSNP; Celera; HapMap; ABI_Val; HGBASE
 Population (Allele,Count): caucasian (C,42|T,78)
 SNP Type: TRANSCRIPTION FACTOR BINDING SITE; INTERGENIC;
 UNKNOWN

50 Context (SEQ ID NO: 220):

CCAAAAAAAAAAAAAAAAAGAAAGAAAGAAAAAAAAAGAAAACAAAAAGAAAGCTGGGCA
CGGTGGCACACACCTGTGGTGTAGTCCCAGTTCCTCAGGAGGTTGA

K

TTGAGGTGAGAGAATCATTGTGTTGCCAGGTTATTCTCAAACCTCCTGGCCTCAAGCG
5 ATCCTCCCGCCTTGGCCTCCCAAATGATGGGATTACAGGCAT

Celera SNP ID: hCV26960066

Public SNP ID: rs4325594

SNP in Genomic Sequence: SEQ ID NO: 175

SNP Position Genomic: 26519

10 Related Interrogated SNP: hCV8717889 (Power=.6)

SNP Source: Celera; HapMap; ABI_Val; HGBASE

Population (Allele,Count): caucasian (T,88|G,32)

SNP Type: INTERGENIC; UNKNOWN; REPEATS

15 Context (SEQ ID NO: 221):

TATTATATGCCACACTGTCTAGGTACTTAGAACACAAGTATGAACAAGCCAGGCC
CAGCCCCTACCCTGTGGAGCTACAGAACCTCTAACAGGGATCG

M

AGATGTGCAGTAACAATTATAGTGTGATGAGCGATCCTGGTTGGAGGGCAAAAGAA
20 TCGGCAAGAAAACAATACCATAGAACAGCATTATGAAGATTTTC

Celera SNP ID: hCV30677672

Public SNP ID: rs7219773

SNP in Genomic Sequence: SEQ ID NO: 175

SNP Position Genomic: 33828

25 Related Interrogated SNP: hCV8717889 (Power=.9)

SNP Source: dbSNP; HapMap

Population (Allele,Count): caucasian (C,55|A,65)

SNP Type: INTERGENIC; UNKNOWN

30 Gene Number: 29

Celera Gene: Chr7:25698180..25784576

Gene Symbol:

Protein Name:

Celera Genomic Axis: GA_x5YUV32W3JR (18869045..18955441)

35 Chromosome: 7

OMIM NUMBER:

OMIM Information:

Genomic Sequence (SEQ ID NO: 176):

40

SNP Information

Context (SEQ ID NO: 222):

GTAAAGAACTTCTCCGTAAGGCCTGGTTTTCCTCCTTTTTCAACGTGAGCCTCAGCA
45 TCTTCTGCTCTTTGCTGCCCCACCAAATCTCCCTCTGTGGCA

Y

CAACCTGCCATCACACACATTGTCAGAGGTGGCATTGGGCAACGCAACTTCAGAGTT
CCTTGTCTCTGCTGATCTCTGAGAATCCATGGGGCACATTCCAC

Celera SNP ID: hCV11827509

50 Public SNP ID: rs1859849

SNP in Genomic Sequence: SEQ ID NO: 176

SNP Position Genomic: 9511
SNP Source: dbSNP; Celera; HapMap; HGBASE
Population (Allele,Count): caucasian (T,89|C,25)
SNP Type: INTERGENIC; UNKNOWN

5

Context (SEQ ID NO: 223):
CAATCACTACTACCAAAGTTTCTTTCTCTACACCTTTTAGGGGATAGGACAAGTTTGT
TCTGAGAGCAAAGTTTAAATCTGAGGGAAAGACAATCACTA
Y

10

GGATTTCAAGAGGACTAAATGTCTTAGCAGCTTGCAGTTAAATGTCTTAGCTTGCAG
TTTCCCCTGAGCAGCTGCAGCTCTTGTGGTTGCCCCAGAGGTT

Celera SNP ID: hCV15738

Public SNP ID: rs740345

SNP in Genomic Sequence: SEQ ID NO: 176

15

SNP Position Genomic: 9126

Related Interrogated SNP: hCV11827509 (Power=.51)

SNP Source: dbSNP; Celera; HGBASE

Population (Allele,Count): caucasian (T,94|C,26)

SNP Type: INTERGENIC; UNKNOWN

20

Context (SEQ ID NO: 224):
GCCAGCGGTAGCCAGTTAACTCATGCACAATCCCTAGCTAGTGTCCCTGCCCCCATC
ACTCCACTCACCTGTACAGACAAACCCAGGTCCCCTCCAAACT
Y

25

GGCTTCTTTGACTCTAATGCTAAGTGATGTAGCCATCACAGCACATACTGTTCTTAGA
CCTCGCCAGCCTCGCCTGTGACAGCAGGATTAGACAGGGAA

Celera SNP ID: hCV345686

Public SNP ID: rs9942690

SNP in Genomic Sequence: SEQ ID NO: 176

30

SNP Position Genomic: 1009

Related Interrogated SNP: hCV11827509 (Power=.51)

SNP Source: dbSNP; Celera; HapMap

Population (Allele,Count): caucasian (C,92|T,28)

SNP Type: HUMAN-MOUSE SYNTENIC REGION; INTERGENIC; UNKNOWN

35

Context (SEQ ID NO: 225):
CTGCGTAGTGCCCGGGTCGGTGTGAGGGGCACCAGGAAAGGGGATTTATGTCTCTTT
GCCTTTACCCCTAATTCTCGCCTTGACCATATCCAGTAAAATC
M

40

GCAAGTAAAATCTAGCAAGCTCCAGTGAAATCAGGTTAAAATAATCACATGGGTAG
CTTTCAGGAGGCTTAAGGTAAGTGGGGCAAGTTTGTACGCATAG

Celera SNP ID: hCV346553

Public SNP ID: rs4722492

SNP in Genomic Sequence: SEQ ID NO: 176

45

SNP Position Genomic: 8407

Related Interrogated SNP: hCV11827509 (Power=.51)

SNP Source: dbSNP; Celera; HGBASE

Population (Allele,Count): caucasian (A,94|C,26)

SNP Type: INTERGENIC; UNKNOWN

50

Context (SEQ ID NO: 226):

- CATGTTTGGCCCCATATGGACTACACAGCTCAGAGAAATGAAGTGGCACATTTATCA
AGTCTCAAGTTTGGGGTCCCATATAGATTTATTTGAGCCACAC
Y
GGCCAGAGGTGACTGAGCCTCCACAGTCCTTAACGTGCTTATGTTTTTTATTGTCCTC
5 TCCAGCTGTATAAGTGCCTCTGCCCTTTGGAACACAGTATTA
Celera SNP ID: hCV346555
Public SNP ID: rs11772927
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 7835
10 Related Interrogated SNP: hCV11827509 (Power=.51)
SNP Source: dbSNP; Celera; HapMap
Population (Allele,Count): caucasian (C,83|T,19)
SNP Type: INTERGENIC; UNKNOWN
- 15 Context (SEQ ID NO: 227):
GTATTGTCTTCCCGCTCTGCAGTCACGGGCAGCTGAACATTTTGCTTGGTACGTCCAC
AGAAGTTTCTTCTTGCATGCAAGATTGGGGTTTTAATACC
Y
GCTGCGTTTGGAAAAATTCTAAAATCAGAAATGATGTGGTCTGAAGTGTTGACATGT
20 TTGGCCCCATATGGACTACACAGCTCAGAGAAATGAAGTGGCA
Celera SNP ID: hCV346556
Public SNP ID: rs11762724
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 7682
25 Related Interrogated SNP: hCV11827509 (Power=.51)
SNP Source: dbSNP; Celera; HapMap
Population (Allele,Count): caucasian (T,94|C,26)
SNP Type: HUMAN-MOUSE SYNTENIC REGION; INTERGENIC; UNKNOWN
- 30 Context (SEQ ID NO: 228):
TGCAGCAGTAAACAGTTCCTTTAAAACTGACTGAAGGATTTGGTGCTACTTTTGAAG
GCAATTTATAAAGCTATTCATGTTCTCTAAAGGATCTGCCCT
K
CAACTAACTTCCAGAAAGGTTGTCTTGATCTGTTAATCCTGACAGAGACACCTTGT
35 TCTCTTTGGAGAGGAAAGAGGAGGGAAAAAACAGAAACATAAG
Celera SNP ID: hCV349359
Public SNP ID: rs11983371
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 37505
40 Related Interrogated SNP: hCV11827509 (Power=.51)
SNP Source: dbSNP; Celera; HapMap
Population (Allele,Count): caucasian (G,92|T,26)
SNP Type: HUMAN-MOUSE SYNTENIC REGION; INTERGENIC; UNKNOWN
- 45 Context (SEQ ID NO: 229):
ACTGCCACAGCAGAGGCAGGATGTGAACAGGCCTGGATGAGTAGGATACAGATTGG
AAAATAAAGGGGAAAAAACATTCTAAATAGGATAAAGGGATA
Y
GAATAAAAGAACAACCTCTGGAATGGGTGTGTTACATTAGGGATGGTTCAGAGACAG
50 AATAGAGGAGTCTCTTGGAAGCAGAGTTTACACAGGGCAGATAA
Celera SNP ID: hCV2576008

Public SNP ID: rs10807830
 SNP in Genomic Sequence: SEQ ID NO: 176
 SNP Position Genomic: 5184
 Related Interrogated SNP: hCV11827509 (Power=.51)
 5 SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (C,94|T,26)
 SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 230):

10 AGCTTTATTCATGACTGCCAACACTTGCAAGCAACCAAGATGTCCTTCAATATGTAA
 ATAGATCAAACCAATACAACCATGTGATGGCTGTACTTATTCA
 S
 TGATGAAAAGAGCTATCAAGCCACCAAGATACCTGGAGGAAACAAATGCATACTGC
 TAAGTGAAAGAAGTTCATCTTAGAAGAAGCTAC

15 Celera SNP ID: hCV2576009
 Public SNP ID: rs11762312
 SNP in Genomic Sequence: SEQ ID NO: 176
 SNP Position Genomic: 5505
 Related Interrogated SNP: hCV11827509 (Power=.51)
 20 SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (G,94|C,26)
 SNP Type: INTERGENIC; UNKNOWN; REPEATS

Context (SEQ ID NO: 231):

25 GCCTCTGCCCTTTGGAACACAGTATTAACGGCACTGACAATTCTGCTCATTTCCTC
 TTTTATCAGTAATTCAGCCTTATGAATCCAATAAATATCAG
 M
 CTAGGCTTGAATTATACATTGGTGAATTCTTTATCTGTCCTTTTGCTAGTTTCCACAG
 CTTTGGAGCAAATTATCATACTAGTGCGACATACTTGGCA

30 Celera SNP ID: hCV2576011
 Public SNP ID: rs758012
 SNP in Genomic Sequence: SEQ ID NO: 176
 SNP Position Genomic: 8009
 Related Interrogated SNP: hCV11827509 (Power=.51)
 35 SNP Source: dbSNP; Celera; HapMap; ABI_Val; HGBASE
 Population (Allele,Count): caucasian (A,92|C,26)
 SNP Type: HUMAN-MOUSE SYntenic REGION; INTERGENIC; UNKNOWN

Context (SEQ ID NO: 232):

40 CAGGAAGACCTTCCTGATTCTTCCAGTAGGAATTGATCTCTCTCTCTCTCTCTCTCT
 CCCGAAGTCTACAGGCCCTTGTTGTCTCTGCTCCTTTCGC
 K

TGATTCTCTCACATTGCTTTGTGTTATGTTTGCTTTCCAAACAGGCTGTGCTCCCAA
 GGAGTGAGTTGTATTTTGGATTTTCTGTACTACAAATTTTCT
 45 Celera SNP ID: hCV2576021
 Public SNP ID: rs1544480
 SNP in Genomic Sequence: SEQ ID NO: 176
 SNP Position Genomic: 17966
 Related Interrogated SNP: hCV11827509 (Power=.51)
 50 SNP Source: dbSNP; Celera; HapMap
 Population (Allele,Count): caucasian (G,91|T,23)

SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 233):

5 AAACAAACCGAAGGTCCAAGGGGACTTCCAGCAAGGTCTGAAATTGAGAATTTGGC
TTAAAAGACTGCAAGGGATGAAGAAAAGACCGGGAGAAGGCTAA
M
TATTTGGTTAAGACAATATCATATTACAGAGTTATGAATTATGCAAATGAAACCACG
GGATTCAACCTCATGAGATTTCTTGTACCTACACAAAAAAC
10 Celera SNP ID: hCV2576027
Public SNP ID: rs886836
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 19189
Related Interrogated SNP: hCV11827509 (Power=.51)
SNP Source: dbSNP; Celera; HapMap
15 Population (Allele,Count): caucasian (C,94|A,26)
SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 234):

20 TTCTCCACATTCCAACATGGACAGTGCTTAGTACTGGCTTATCCAACCCTGACTGTGG
GCTCTCTGTCAAGCACTCTGCTATGGGTACTGGGGTGTCTTCT
R
AGGTAACACGTGGAGGAAAGCCACTTGACTTTAGACACTGTTTTGTGATATCAGCTC
CTAAGCCTGCAGCGGTATCTGTGAAAGCTGTAAAATGGAATA
25 Celera SNP ID: hCV2576036
Public SNP ID: rs10951092
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 30065
Related Interrogated SNP: hCV11827509 (Power=.51)
SNP Source: dbSNP; Celera; HapMap
30 Population (Allele,Count): caucasian (G,94|A,26)
SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 235):

35 GGGGGTTTGCCAGAACCTTTGGTTAAGGAAGCTTGCCCTCATTTCAACAACCCCCAG
GCACCTCTGCATGGGCCGAGGCCAGGCCCTGCCTCTGTGGTTG
K
TCTCCTCCCACGACCAGAGGAAATTAGACCTGTTCTTTGCTGTCAGCTTAAATCCTTA
GGAGGGGGGCTAGAAATTAAGATTGTTCTTCACTCTACCAATG
Celera SNP ID: hCV2576041
40 Public SNP ID: rs10951098
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 38829
Related Interrogated SNP: hCV11827509 (Power=.51)
SNP Source: dbSNP; Celera; HapMap
45 Population (Allele,Count): caucasian (G,94|T,26)
SNP Type: TRANSCRIPTION FACTOR BINDING SITE; HUMAN-MOUSE
SYNENIC REGION; INTERG
ENIC; UNKNOWN

50 Context (SEQ ID NO: 236):

- AAATTCCTGGCCAAGTACATTGAACATCAGAAAAGATCACCCAGGTAAAATGTAGTT
GGGTTATATAATTAGTTTATTTAAATCAAGAGCAAGGAATACA
Y
GTTTCAGAAAAGCATAAAACATCAGACAAATCCAAATTCAAAAAAAAAAAAAACAACAA
5 CAAATTCAGAAATATGCTACAAAATAAGCAGCCAATACTCTTTA
Celera SNP ID: hCV2576042
Public SNP ID: rs2391151
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 39097
10 Related Interrogated SNP: hCV11827509 (Power=.51)
SNP Source: dbSNP; Celera; HapMap; ABI_Val; HGBASE
Population (Allele,Count): caucasian (T,94|C,26)
SNP Type: INTERGENIC; UNKNOWN
- 15 Context (SEQ ID NO: 237):
AAAGCATAAAACATCAGACAAATCCAAATTCAAAAAAAAAAAAAACAACAAATTC
AGAAATATGCTACAAAATAAGCAGCCAATACTCTTTAAATTTGT
R
AAAAACAAAGACTAAGGAACCGTTACAGATCGGAAGAAAGGAAGGAAAGACAATA
20 GCTACATGCAATGTGAGATCACAGGCTGGATCCTGGACCAGATCA
Celera SNP ID: hCV2576044
Public SNP ID: rs2391152
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 39205
25 Related Interrogated SNP: hCV11827509 (Power=.51)
SNP Source: dbSNP; Celera; HapMap; HGBASE
Population (Allele,Count): caucasian (A,92|G,24)
SNP Type: INTERGENIC; UNKNOWN; REPEATS
- 30 Context (SEQ ID NO: 238):
AGGCAGCCTTCAAACCTCAGGTCTGTCTGACTTCAGAACACATTCTCTTTCAATACATA
TAACACGCTTCCCCATTTCTTCTTGATGGAAAAAAGCAGCCA
R
TGGTTTAGATTGTCATTCCAGTTCCCTATATTACTACCTCTTGAATACTGAAAAATTG
35 GTTCCATATTAAGAGTGGTGAAACTTTGGTGGCCAAATAGA
Celera SNP ID: hCV2576045
Public SNP ID: rs6953294
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 43713
40 Related Interrogated SNP: hCV11827509 (Power=.51)
SNP Source: dbSNP; Celera; HapMap; ABI_Val
Population (Allele,Count): caucasian (G,93|A,25)
SNP Type: INTERGENIC; UNKNOWN
- 45 Context (SEQ ID NO: 239):
AATTTTCTGCTCCACTCTGCAGAGCTGCGTAGTGCCCGGGTCGGTGTGAGGGGCACC
AGGAAAGGGGATTTATGTCTCTTTGCCTTACCCCTAATTCTC
R
CCTTGACCATATCCAGTAAAATCAGCAAGTAAAATCTAGCAAGCTCCAGTGAAATCA
50 GGTAAATAATCACATGGGTAGCTTTTCAGGAGGCTTAAGGTA
Celera SNP ID: hCV7600851

Public SNP ID: rs1106876
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 8383
Related Interrogated SNP: hCV11827509 (Power=.51)
5 SNP Source: dbSNP; Celera; HapMap; HGBASE
Population (Allele,Count): caucasian (G,94|A,26)
SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 240):

10 CAATCTGACACACATTTATAGTGAATTTTGTGTGCCAACAAAGATAAAAAGGAAAA
GGATACGTGGTTGTCCTCAAGGAAGTTATTACAATTATCTCTCA
K
CTGACTGGGACCTCTGTTTTCTGAATATTGTTCCCTTCTATCCTCTGCCTACTTCAAAA
AGGACTGAAGTGGTACAGGATTTTCAAAGTACTCTTTAAG
15 Celera SNP ID: hCV7600852
Public SNP ID: rs1558077
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 14389
Related Interrogated SNP: hCV11827509 (Power=.51)
20 SNP Source: dbSNP; Celera; HapMap; HGBASE
Population (Allele,Count): caucasian (T,87|G,25)
SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 241):

25 CACTGACAAGGGATCCTCTCATAATCTGCATTTACCTCCAATTGTAGCAATATTTTCA
CTGAGTTATAACTGCTGCCTCCCTGTCTGTCTTCCTGCCTGC
R
TGAGGGATAGAATTGTGTTCTGCTTAGTTGAAATTCCAGTAGCTAATCTAATCTCTGG
CTGAAAGAAGGTTACCAAAAATTAAATAATTTGGGAGGGATA
30 Celera SNP ID: hCV7600855
Public SNP ID: rs1558075
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 36350
Related Interrogated SNP: hCV11827509 (Power=.51)
35 SNP Source: dbSNP; HapMap; ABI_Val; HGBASE
Population (Allele,Count): caucasian (A,94|G,26)
SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 242):

40 TGCGTTTTGGTTCATGGGAAATGATCTGTGGGCTACAAGACAAATTTTATTTAATGGG
AGTTCTGAGGCAAATAAAAACAAGCTGATTTGTGGCAAATGAG
Y
GAAGAAGGAACAGTCACCCACTTGGGTGATTCTTCTTACAGGATCAC
GAGGTGTTTTGATTATTTTAGTTTCAGTTTTAAATTTTGGTTTT
45 Celera SNP ID: hCV9949982
Public SNP ID: rs12667142
SNP in Genomic Sequence: SEQ ID NO: 176
SNP Position Genomic: 1969
Related Interrogated SNP: hCV11827509 (Power=.51)
50 SNP Source: dbSNP; Celera
Population (Allele,Count): caucasian (T,92|C,28)

SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 243):

CAAACCCAGAAAAGGTAATGCATTTTGTGCTAGGACCTTATGGCAGCTATGATGTCA
5 TTAGACAATAGGAATTTTCCAGCTCCATTATAGTTTTGTCGGC

R

CACTGTCATAAGTGTGGTCTGTGGTTGACCGGAACACAGCTATGCAGCACGTGACTG
TATCCCTCCTCCCATGCTCACTTCTCTTTTCTACCTGCTTCTG

Celera SNP ID: hCV11170193

10 Public SNP ID: rs2058024

SNP in Genomic Sequence: SEQ ID NO: 176

SNP Position Genomic: 66396

Related Interrogated SNP: hCV11827509 (Power=.51)

SNP Source: dbSNP; Celera; HapMap

15 Population (Allele,Count): caucasian (A,75|G,19)

SNP Type: INTERGENIC; UNKNOWN; REPEATS

Context (SEQ ID NO: 244):

ATTTTATTCAGGAGTTTCTACTTTAATCCTCACGTCTTTGCCCTCTTGTTTTGCCTGTA
20 AAGAACTTCTCCGTAAGGCCTGGTTTTCCCTCCTTTTTCAA

Y

GTGAGCCTCAGCATCTTCTGCTCTTTGCTGCCCCACCAAATCTCCCTCTGTGGCACCA
ACCTGCCATCACACACATTGTCAGAGGTGGCATTGGGCAACG

Celera SNP ID: hCV11170195

25 Public SNP ID: rs1859848

SNP in Genomic Sequence: SEQ ID NO: 176

SNP Position Genomic: 9455

Related Interrogated SNP: hCV11827509 (Power=.51)

SNP Source: dbSNP; Celera; HGBASE

30 Population (Allele,Count): caucasian (C,94|T,26)

SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 245):

TCTATACATCTAGGTTATAGGGCCCTGGTCTTGTGCACTCTCACAAATCCCCAGCAAT
35 ATTCCATGTTCTGGTCCATCCCTTCCACTAACAGGCGTCACC

Y

ACCTGGTGTTTCCCTTTCCTTAGGAGTCAGTTCTCTGCTAGATGATAGTTAATTGTAT
AATTTTAAATCTCAAACATCTCAGTACACATCAGTGTACA

Celera SNP ID: hCV15854511

40 Public SNP ID: rs2214976

SNP in Genomic Sequence: SEQ ID NO: 176

SNP Position Genomic: 13020

Related Interrogated SNP: hCV11827509 (Power=.51)

SNP Source: dbSNP; HapMap; ABI_Val; HGBASE

45 Population (Allele,Count): caucasian (C,94|T,26)

SNP Type: TRANSCRIPTION FACTOR BINDING SITE; INTERGENIC;
UNKNOWN

Context (SEQ ID NO: 246):

50 GCAGCTCTCAGCAGTGTGTCCAGGATGTACTAATATTTACCAGAAGAGGCGGCCTCT
GGTGTCAGTACTGAGTCCCCACTCACCTCATGTGACTGCTGCTTCC

M

AGCCTGCTTCACAGAGTTGGCCCAGAGGAGGGTGTGGTCTGAGTTTGCATCAGAGAT
CTGGTTTTACCTTAACCCCTCAAAAATCACAATTATATTTAGC

Celera SNP ID: hCV16151469

5 Public SNP ID: rs2190532

SNP in Genomic Sequence: SEQ ID NO: 176

SNP Position Genomic: 6691

Related Interrogated SNP: hCV11827509 (Power=.51)

SNP Source: dbSNP; Celera; HapMap; ABI_Val; HGBASE

10 Population (Allele,Count): caucasian (A,90|C,22)

SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 247):

TTGCAGATGGAAGGAGCCCACCGCTCCTGCAGGGCTTAGCACGTGGCTGGTCACTGC
15 CTCGGTGGGGTCGATCAGGCCCTCTCTCTGCCATTCTGGGA

R

AAGTATCTATCATGGTTCTTTCCACACTGATTTTATTCAGGAGTTTCTACTTTAATCCT
CACGTCTTTGCCCTCTTGTCTTGCCTGTAAAGAACTTCTCC

Celera SNP ID: hCV27263228

20 Public SNP ID: rs2893212

SNP in Genomic Sequence: SEQ ID NO: 176

SNP Position Genomic: 9325

Related Interrogated SNP: hCV11827509 (Power=.51)

SNP Source: dbSNP; Celera; HGBASE

25 Population (Allele,Count): caucasian (G,94|A,26)

SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 248):

TCCCTCATTCATCATTTCTTAACATTATTTAAATATATGTAAGTAAGACTTGTTCTTC
30 TCCCCATGTATTTATTTAAACATGTATTTATATCAGTATGG

W

TTTATAGATATTTGTTTCATTTGGGGGTTATAATCTAATACTGTCATTGTTGATTTTGC
TTCTCAGATTTTCTAGGTTTGGCCATTAGAACTCTTTCAG

Celera SNP ID: hCV27263241

35 Public SNP ID: rs2158932

SNP in Genomic Sequence: SEQ ID NO: 176

SNP Position Genomic: 41035

Related Interrogated SNP: hCV11827509 (Power=.51)

SNP Source: dbSNP; Celera; HapMap; HGBASE

40 Population (Allele,Count): caucasian (A,93|T,25)

SNP Type: INTERGENIC; UNKNOWN; REPEATS

Context (SEQ ID NO: 249):

TCATTGAAGAGTGAAAGAGAGATGGAGGGTGCTTGGGCCTCAGGCTCATGAATGTA
45 AGCCCTGAGGGTTTCTGGGTATTTGCAGCAGTGGATCAGGATCC

W

TGAGATCACTTCCCTCTAGCACATGAGGTCACCTTGCATTTAGGGGCCCCAATACCA
ATGTGTTTGTGGTGTGGTCCTGGTTTGTTTAGAAATGGTTACCA

Celera SNP ID: hCV29401758

50 Public SNP ID: rs6942970

SNP in Genomic Sequence: SEQ ID NO: 176

SNP Position Genomic: 6275
 Related Interrogated SNP: hCV11827509 (Power=.51)
 SNP Source: dbSNP; HapMap
 Population (Allele,Count): caucasian (A,94|T,26)
 5 SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 250):
 TTCCTCTCCCTACTTTCCACAATTAGAACACTATTGGCCAAGGTCCTTTTGGATCAGC
 TGCTATATAATGTTGACAATAATGTCAACTGCACTCTCCTGG

10 R
 ACTAACTAATTAAATTTCAAAGTTTAATCAAAAGAAAGTGTACATGTTCCCTATTAA
 TATGCTGACTTTTTTGGGAAGTTGGGTGAATGTTAGCTGCACTT
 Celera SNP ID: hCV32157909
 Public SNP ID: rs11489283

15 SNP in Genomic Sequence: SEQ ID NO: 176
 SNP Position Genomic: 42497
 Related Interrogated SNP: hCV11827509 (Power=.51)
 SNP Source: dbSNP; HapMap
 Population (Allele,Count): caucasian (A,94|G,24)
 20 SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 251):
 TTGCTTGGTCCTGATATATTAACCTTTTTCTATATTATTGGGAACAATTTGCTAATATT
 TTACAGAGGATTTGAATCTATATTCATGAGAGATAGTCTG

25 K
 GGTTTTCTTTTCTTGTATTGTCTCTAGTTTTGGTACTAGTAAGGGTAGAGTATAGAAA
 TAGAAGAGGAGATGGTGGGGAAATACCTCTAAAGACTGCCTA
 Celera SNP ID: hCV32157908
 Public SNP ID: rs11973825

30 SNP in Genomic Sequence: SEQ ID NO: 176
 SNP Position Genomic: 42096
 Related Interrogated SNP: hCV11827509 (Power=.51)
 SNP Source: dbSNP; HapMap
 Population (Allele,Count): caucasian (G,94|T,26)
 35 SNP Type: INTERGENIC; UNKNOWN; REPEATS

Context (SEQ ID NO: 252):
 ACACAGGGGATTAATGCTGTTTAACCTCACCAATTTAAAATGTAAAGGAATACACGCA
 AAAAGCCTGCCCCTGGCACTGTGCATGCCAAAGGAGTTAAGCA

40 R
 CCCACATGCATGTATTCTAGTTTGGAGACATGACAAGGACACACATAGAAGGCTGT
 GAGGATGCAGAAGAGGGAGAGATCATTCTGGATTGAGATG
 Celera SNP ID: hDV70903693
 Public SNP ID: rs17152187

45 SNP in Genomic Sequence: SEQ ID NO: 176
 SNP Position Genomic: 4646
 Related Interrogated SNP: hCV11827509 (Power=.51)
 SNP Source: dbSNP; HapMap
 Population (Allele,Count): caucasian (G,94|A,26)
 50 SNP Type: INTERGENIC; UNKNOWN

Context (SEQ ID NO: 253):

TTCTGTTGATGTGTGACATTTGAAAGAAGAGTGGAGTCTGGGACTGGGGACGGGAG
GTGGGCAGGCGTCTATGGAGGCTCCTCCATCCAACACTTCTCAA

Y

5 GAGGATTCCAGGAGGCCAGCCTGTAAATGCTTGTATTGTCTTCCCGCTCTGCAGTC
ACGGGCAGCTGAACATTTTGCTTGGTACGTCCACAGAAGTTTT

Celera SNP ID: hDV70903706

Public SNP ID: rs17152208

SNP in Genomic Sequence: SEQ ID NO: 176

10 SNP Position Genomic: 7548

Related Interrogated SNP: hCV11827509 (Power=.51)

SNP Source: dbSNP; HapMap

Population (Allele,Count): caucasian (T,94|C,26)

SNP Type: INTERGENIC; UNKNOWN

15

Context (SEQ ID NO: 254):

TTATGTGCCATGCTTCAGGAAACAAACTCCCTTTCTGTCCTCCTTGCCTTCTCTTTGAC
TCAAGCACACAGCAATTCCTGTTCAGGCCCAAGATGCAGAT

M

20 TAGAATTACCAGCCCTGGGATCCCAGCCCCAACTCAATGTACTACTGCTACTACCCA
AAGAACTCCAGGAGGTTCTCCTGTCTGCTTTAATGTAGCAAA

Celera SNP ID: hDV70903726

Public SNP ID: rs17152235

SNP in Genomic Sequence: SEQ ID NO: 176

25 SNP Position Genomic: 10000

Related Interrogated SNP: hCV11827509 (Power=.51)

SNP Source: dbSNP

Population (Allele,Count): caucasian (A,94|C,26)

SNP Type: INTERGENIC; UNKNOWN

Table 3

5	hCV22274641	SEQ ID NO: 104
	hCV22274641	SEQ ID NO: 105
	hCV15746640	SEQ ID NO: 138
	hCV15746640	SEQ ID NO: 139
	hCV15746640	SEQ ID NO: 140

Table 4

5	hCV22274641	SEQ ID NO:182
	hCV15746640	SEQ ID NO:192
	hCV15746640	SEQ ID NO:202
	hCV22274641	SEQ ID NO:203

Table 5 Primers

Marker	Alleles	Primer 1 (Allele-specific Primer)	Primer 2 (Allele-specific Primer)	Common Primer
hCV10057645	C/T	CAGAGCGGGTCATACAAAC (SEQ ID NO:255)	CAGAGCGGGTCATACAAAT (SEQ ID NO:256)	GAGATTAACTCCAGTACACTCA AATATA (SEQ ID NO:257)
hCV11827509	C/T	AAATCTCCCTCTGTGGCAC (SEQ ID NO:258)	AAATCTCCCTCTGTGGCAT (SEQ ID NO:259)	TCCCCCTCCTCAAGACA (SEQ ID NO:260)
hCV1226613	C/T	ATTCGTTCTGCTCAGATAAG AC (SEQ ID NO:261)	ATTCGTTCTGCTCAGATAAGAT (SEQ ID NO:262)	TGCCATTAAACCTTGTTTATACT C (SEQ ID NO:263)
hCV15746640	A/G	GAAACGGCTGTTTGTGT (SEQ ID NO:264)	AACGGCTGTTTGTGTC (SEQ ID NO:265)	TCCAAATCTACCTTTTCCTTGTAT T (SEQ ID NO:266)
hCV1834284	A/C	GACCTTCTCCTGAAGGAAGT T (SEQ ID NO:267)	ACCTTCTCCTGAAGGAAGTG (SEQ ID NO:268)	TTCTGGAGGAGAAAGTTCTAGAG A (SEQ ID NO:269)
hCV1841831	C/T	CCAGGAAGGAACCCC (SEQ ID NO:270)	CCCAGGAAGGAACCCCT (SEQ ID NO:271)	GGTCCCTGTTCAACCACTGTC (SEQ ID NO:272)
hCV1936853	C/G	ACCTGTGGAAGGCCATC (SEQ ID NO:273)	CCTGTGGAAGGCCATG (SEQ ID NO:274)	GGGCCCCCAAAATAGTA (SEQ ID NO:275)
hCV2179030	C/T	GCTTGTCATTTTCTAGATCCA C (SEQ ID NO:276)	GCTTGTCATTTTCTAGATCCAT (SEQ ID NO:277)	CAGCTTGAGGGTTCCTTAGA (SEQ ID NO:278)
hCV2188996	A/T	GCTTCCTCGCATTTGAAT (SEQ ID NO:279)	GCTTCCTCGCATTTGAAA (SEQ ID NO:280)	TGCTTCAAGGCTCTACATAATC (SEQ ID NO:281)
hCV22274641	A/T	CCTCCCCATCAGCATCA (SEQ ID NO:282)	CCTCCCCATCAGCATCT (SEQ ID NO:283)	CCTTTTGACCATGGAATACCA (SEQ ID NO:284)
hCV2447571	C/T	TTCCGCTGACCATCAAC (SEQ ID NO:285)	ATTCCGCTGACCATCAAT (SEQ ID NO:286)	TCCCATCAAAAGTGCTAACA (SEQ ID NO:287)
hCV2508722	A/G	AGGGTTGGTTATACCCAGAA T (SEQ ID NO:288)	GGGTGGTTATACCCAGAAC (SEQ ID NO:289)	TTCCAACCTCTGGGCTGAAGTAC (SEQ ID NO:290)
hCV25741966	G/A	TGCGGCTTCTCTCTTCC (SEQ ID NO:291)	TGCGGCTTCTCTCTTCT (SEQ ID NO:292)	ACCAAAGGCAGTGATGATG (SEQ ID NO:293)
hCV25942411	A/G	TCGGAGGAACCTCTGTGA (SEQ ID NO:294)	CGGAGGAACCTCTGTGG (SEQ ID NO:295)	GGGTGAGTGACAGGAAAATG (SEQ ID NO:296)
hCV3084827	C/T	AGGAAGGCTCACCTGG (SEQ ID NO:297)	CAGGAAGGCTCACCTGA (SEQ ID NO:298)	TGTGACAGCGTTTCTCTTCT (SEQ ID NO:299)
hCV588691	G/C	TGCCAACACCCCCATG (SEQ ID	CTGCCAACACCCCCATC (SEQ ID	CTGAGACCCAGATTGTGACTT

LI, et al.
Date of Deposit: July 16, 2007

Attorney Docket #: CD000009PCT
Express Mail #: ER 622077997 US

hCV630374	C/T	NO:300) CGTGGAGGAGGTGGAC (SEQ ID NO:303)	NO:301) CGTGGAGGAGGTGGAT (SEQ ID NO:304)	(SEQ ID NO:302) CTGAGGGCAAGGATGTTT (SEQ ID NO:305)
hCV8717889	A/T	CGAGTGACATTAAAGCTCCTC T (SEQ ID NO:306)	CGAGTGACATTAAAGCTCCTCA (SEQ ID NO:307)	AGGAACTCCTTCTCCAGACAT (SEQ ID NO:308)
hCV905013	G/T	ACACCTCGCCCAGTAATC (SEQ ID NO:309)	GACACCTCGCCCAGTAATA (SEQ ID NO:310)	CCCCCTCTCCAGATTACATT (SEQ ID NO:311)

What Is Claimed Is:

1. A reagent for detecting a single nucleotide polymorphism (SNP) in a SNP-containing nucleic acid molecule, that selectively hybridizes to the SNP-containing nucleic acid molecule as compared to hybridization to a nucleic acid molecule which does not contain the SNP at the same nucleotide position, wherein the SNP is in any one of the nucleotide sequences of SEQ ID NOS:97-147 and SEQ ID NOS:177-254.

2. The reagent of Claim 1 which is a probe.

3. The reagent of Claim 2 in which the probe is a polynucleotide.

4. The reagent of Claim 2 in which the probe is a peptide nucleic acid (PNA).

5. The reagent of Claim 2, in which the probe is at least about 8 nucleotides and not greater than about 60 nucleotides in length.

6. The reagent of Claim 2, in which the 3' end of the probe hybridizes to the SNP in the SNP-containing nucleic acid molecule.

7. The reagent of Claim 2, in which the probe is labeled with a reporter dye.

8. The reagent of Claim 7, in which the reporter dye is a fluorescent dye.

9. The reagent of Claim 7, in which the probe is further labeled with a quencher dye.

10. The reagent of Claim 2, in which the probe is further labeled with biotin, hapten, or oligonucleotide.

11. The reagent of Claim 2, in which the probe is bound to a solid support.

12. The reagent of Claim 11, in which the solid support is a bead or a chip.

13. The reagent of Claim 1 which is a primer.

14. The reagent of Claim 13, in which the primer is a polynucleotide.
15. The reagent of Claim 13, in which the primer is a PNA.
- 5 16. The reagent of Claim 14, in which the primer is at least about 8 nucleotides and not greater than 30 nucleotides in length.
- 10 17. The reagent of Claim 13, in which the 3' end of the primer hybridizes to the SNP in the SNP-containing nucleic acid molecule.
- 15 18. A kit for detecting a SNP in a SNP-containing nucleic acid molecule comprising a primer and a probe, wherein the probe selectively hybridizes to the SNP-containing nucleic acid molecule as compared to hybridization to a nucleic acid molecule which does not contain the SNP at the same nucleotide position, wherein the SNP is in any one of the nucleotide sequences of SEQ ID NOS:97-147 and SEQ ID NOS:177-254.
19. The kit of Claim 18 further comprising a buffer and an enzyme.
- 20 20. The kit of Claim 18, in which its components are in one or more containers.
21. The kit of Claim 18, comprising two or more probes that detect SNPs in distinct SNP-containing nucleic acid molecules.
- 25 22. A reagent for detecting a variant protein encoded by a SNP-containing nucleic acid molecule, said reagent selectively binds to the variant protein as compared to a protein encoded by another nucleic acid molecule, wherein the SNP-containing nucleic acid molecule contains a SNP in any one of the nucleotide sequences of SEQ ID NOS:97-147 and SEQ ID NOS:177-254.
- 30 23. The reagent of Claim 22, which is an antibody, an antibody fragment, an aptamer, a peptide, a ligand, or a small molecule compound.
24. The reagent of Claim 23, which is labeled with a reporter dye or an imaging agent.

25. A kit comprising the reagent of Claim 22 and a buffer.

26. An amplified polynucleotide containing a SNP which is in any one of the nucleotide sequences of SEQ ID NOS:97-147 and SEQ ID NOS:177-254.

5

27. The amplified polynucleotide of Claim 26 which is at least about 20 nucleotides and not greater than about 500 nucleotides in length.

28. An isolated nucleic acid molecule comprising at least about 12 contiguous nucleotides and including at least one SNP identified in Table 1 or 2, wherein said 12 contiguous nucleotides are selected from within a nucleotide sequence in the transcript sequences of SEQ ID NOS:1-48, the genomic sequences of SEQ ID NOS:148-176, the transcript-based SNP context sequences of SEQ ID NOS:97-147, the genomic-based SNP context sequences of SEQ ID NOS:177-254, or complementary sequences thereof.

15

29. An isolated nucleic acid molecule that includes at least one SNP identified in Table 1 or 2, wherein said nucleic acid molecule comprises a nucleotide sequence in the transcript sequences of SEQ ID NOS:1-48, the genomic sequences of SEQ ID NOS:148-176, the transcript-based SNP context sequences of SEQ ID NOS:97-147, the genomic-based SNP context sequences of SEQ ID NOS:177-254, or complementary sequences thereof.

20

30. A method of detecting a SNP in a nucleic acid molecule in a test sample, comprising determining the nucleotide at a specified SNP position in the nucleic acid molecule, wherein the SNP is in any one of the nucleotide sequences of SEQ ID NOS:97-147 and SEQ ID NOS:177-254, and thereby detecting the SNP.

25

31. A method of identifying an individual having an altered risk of developing Alzheimer's Disease, comprising detecting the presence or absence of one or more SNP(s) in said individual's nucleic acids, wherein the SNP(s) is in any one of the nucleotide sequences of SEQ ID NOS:97-147 and SEQ ID NOS:177-254, and the presence or absence of the SNP(s) is correlated with an altered risk of developing Alzheimer's Disease.

30

32. The method of Claim 31, comprising contacting the nucleic acids of the individual with a detection reagent, and determining which nucleotide is present or absent at a specified SNP position.

5 33. The method of Claim 32, in which the nucleic acids are prepared from blood cells of the individual.

34. The method of Claim 32, in which the detection reagent is a polynucleotide probe.

10 35. The method of Claim 34, in which the 3' end of the probe hybridizes to the SNP in the nucleic acid.

36. The method of Claim 34, in which the probe is labeled with a reporter dye.

15 37. A method of diagnosing Alzheimer's Disease in an individual, comprising detecting the presence or absence of one or more SNP(s) in said individual's nucleic acids, wherein the SNP(s) is in any one of the nucleotide sequences of SEQ ID NOS:97-147 and SEQ ID NOS:177-254, and the presence or absence of the SNP(s) is correlated with the Alzheimer's Disease.

20 38. A method of identifying an individual having an altered risk of developing Alzheimer's Disease, comprising detecting the presence or absence of a polymorphism in said individual's nucleic acids, said polymorphism is in linkage disequilibrium with a SNP in any one of the nucleotide sequences of SEQ ID NOS:97-147 and SEQ ID NOS:177-254, and the presence or absence of the SNP is correlated with an altered risk of developing Alzheimer's Disease.

25

39. A method of treating Alzheimer's Disease in an individual, comprising administering into said individual an effective amount of a composition that modulates the function, activity, or expression of a gene, transcript, or variant protein identified in Table 1 or 2.

30 40. A method of human identification, comprising genotyping 10 or more SNPs in nucleic acids in a biological sample, wherein the SNPs are selected from Table 1 or 2 that are at degenerate codon positions.

41. A computer system for studying SNP linkage, said system comprising storage means containing the SNP information provided in Table 1 or 2 or a subset thereof.

42. Computer readable medium having stored therein the SNP relational information
5 provided in Table 1 or 2.

43. An isolated polypeptide consisting of an amino acid sequence selected from the group consisting of SEQ ID NOS:49-96.

10 44. An isolated polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NOS:49-96.

45. An antibody that selectively binds to a polypeptide of Claim 43.

15 46. An isolated nucleic acid molecule consisting of a nucleotide sequence selected from the group consisting of:

- (a) a nucleotide sequence that encodes an amino acid sequence of SEQ ID NOS:49-96;
- (b) a nucleotide sequence selected from the group consisting of SEQ ID NOS:1-48;
- 20 (c) a nucleotide sequence selected from the group consisting of SEQ ID NOS:148-176;
- (d) a nucleotide sequence selected from the group consisting of SEQ ID NOS:97-147;
- (e) a nucleotide sequence selected from the group consisting of SEQ ID NOS:177-254; and
- 25 (f) a nucleotide sequence that is completely complementary to a nucleotide sequence of (a)-(e).

47. An isolated nucleic acid molecule comprising a nucleotide sequence selected from the group consisting of:

- 30 (a) a nucleotide sequence that encodes an amino acid sequence of SEQ ID NOS:49-96;
- (b) a nucleotide sequence selected from the group consisting of SEQ ID NOS:1-48;
- (c) a nucleotide sequence selected from the group consisting of SEQ ID NOS:148-176;

- (d) a nucleotide sequence selected from the group consisting of SEQ ID NOS:97-147;
(e) a nucleotide sequence selected from the group consisting of SEQ ID NOS:177-
254; and
(f) a nucleotide sequence that is completely complementary to a nucleotide sequence
5 of (a)-(e).

48. A nucleic acid array comprising the nucleic acid molecule of Claim 46.

49. A transgenic non-human animal comprising the nucleic acid molecule of Claim
10 47.

50. A vector comprising the nucleic acid molecule of Claim 47.

51. The vector of Claim 50, in which said vector is selected from the group consisting
15 of a plasmid, a virus, and a bacteriophage.

52. The vector of Claim 50, in which said nucleic acid molecule is inserted into said
vector in proper orientation and correct reading frame such that a polypeptide comprising an
amino acid sequence selected from the group consisting of SEQ ID NOS:49-96 is expressed by a
20 cell transformed with said vector.

53. The vector of Claim 52, wherein said nucleic acid molecule is operatively linked
to a promoter sequence.

25 54. A host cell containing the vector of Claim 50.

55. A method for producing a polypeptide, comprising culturing the host cell of Claim
54 under conditions sufficient for the production of said polypeptide, and recovering said
polypeptide.

30 56. A method for detecting the presence of the polypeptide of Claim 43 in a test
sample, comprising contacting said sample with a detection reagent that specifically allows
detection of the presence of the polypeptide in the sample and then detecting the presence of the
peptide.

57. A method for identifying a modulator of the polypeptide of Claim 43, comprising contacting said polypeptide with an agent and determining if said agent has modulated the function or activity of said polypeptide.

5

58. The method of Claim 57, in which said agent is added to a host cell that expresses said polypeptide.

59. A method for identifying an agent that binds to the polypeptide of Claim 43, comprising contacting the polypeptide with an agent and determining whether a complex is formed between the agent and the polypeptide.

10

60. A pharmaceutical composition comprising an agent identified by the method of Claim 59 and a pharmaceutically acceptable carrier therefor.

15

61. A method for treating Alzheimer's Disease, comprising administering to an individual in need of such treatment a pharmaceutically effective amount of an agent identified by the method of Claim 59.

20

62. An isolated nucleic acid molecule comprising at least about 12 contiguous nucleotides and including at least one SNP identified in Table 3 or 4.

63. An amplified polynucleotide containing at least one SNP identified in Table 3 or 4.

25

64. A SNP detection reagent, which selectively hybridizes to a SNP in a nucleic acid molecule, wherein the SNP is identified in Table 3 or 4.

65. A method of administering a therapeutic agent, the method comprising genotyping one or more SNPs disclosed in Tables 1-2 in the nucleic acid of at least one individual, and either administering the therapeutic agent to the individual based on the results of said genotyping having indicated that said individual is likely to respond to said therapeutic agent, or not administering the therapeutic agent to the individual based on the results of said genotyping having indicated that said individual is not likely to respond to said therapeutic agent.

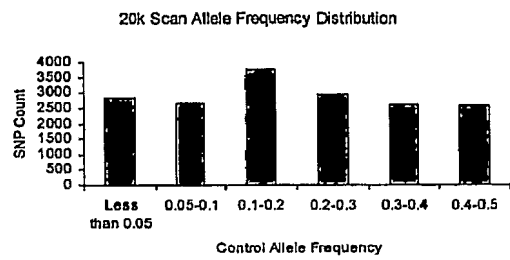
30

66. A method for conducting a clinical trial in which a therapeutic agent is evaluated, the method comprising genotyping one or more SNPs disclosed in Tables 1-2 in the nucleic acid of one or more individuals, and carrying out a course of action based on the results of said
5 genotyping, wherein said course of action comprises including certain of said individuals in the clinical trial based on the results of said genotyping having indicated that said individuals are likely to respond to said therapeutic agent and/or excluding certain of said individuals from participating in the clinical trial based on the results of said genotyping having indicated that said individuals are not likely to respond to said therapeutic agent.

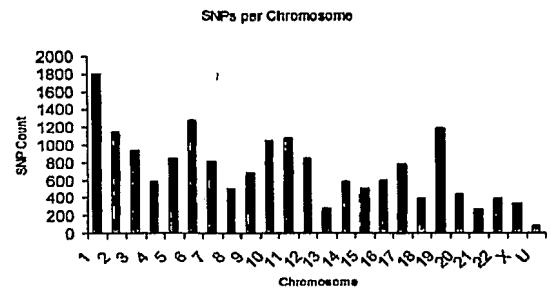
10

67. The method of claim 66, wherein certain of said individuals are excluded from participating in the clinical trial based on the results of said genotyping having indicated that said individuals are at risk for experiencing toxic effects from said therapeutic agent.

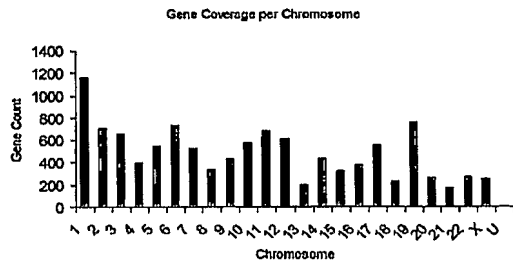
A.



B.



C.



D.

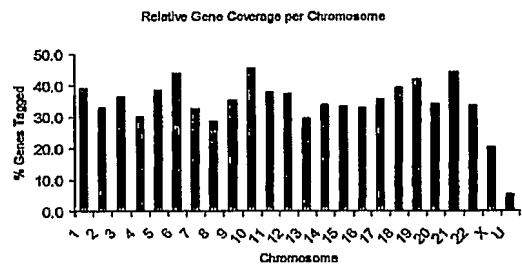
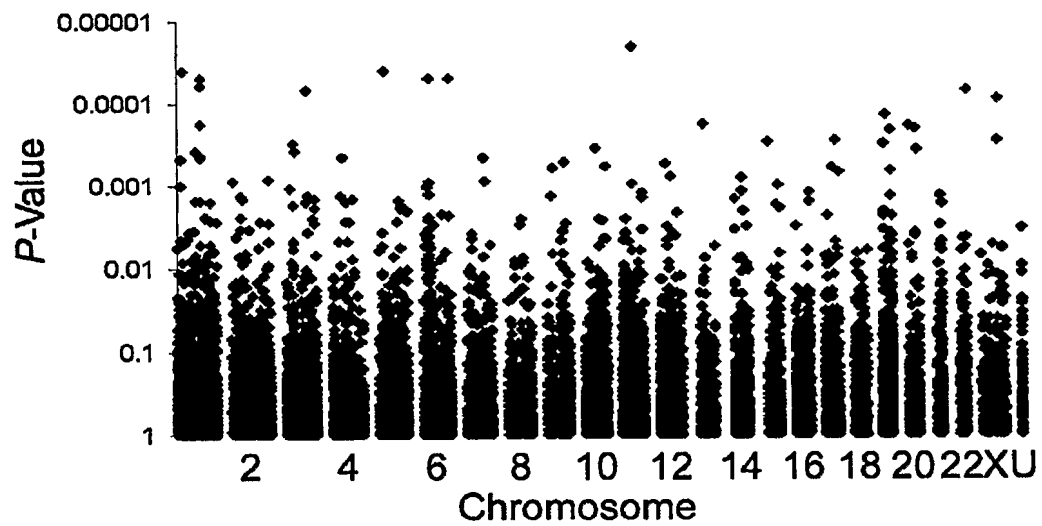


Figure 1

A.



B.

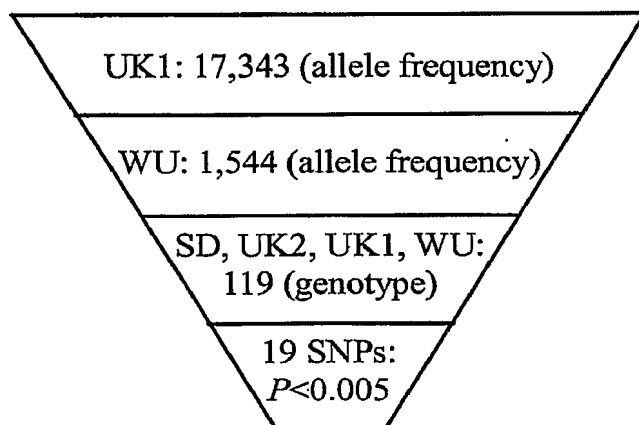


Figure 2