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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 June 2007 (28.06.2007)

PCT

(10) International Publication Number
WO 2007/071996 A2

(51) International Patent Classification:
C07K 14/19 (2006.01)

(21) International Application Number:
PCT/GB2006/004816

(22) International Filing Date:
20 December 2006 (20.12.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/751,872 20 December 2005 (20.12.2005) US

(71) Applicant *(for all designated States except US)*: **UNIVERSITY OF EAST ANGLIA** [GB/GB]; Norwich, Norfolk NR4 7TJ (GB).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **PALMER, Tracy** [GB/GB]; University Of East Anglia, Norwich, Norfolk NR4 7TJ (GB). **WIDDICK, David, Andrew** [GB/GB]; University Of East Anglia, Norwich, Norfolk NE4 7TJ (GB).

(74) Agents: **KREMER, Simon** et al.; Mewburn Ellis LLP, York House, 23 Kingsway, London Greater London WC2B 6HP (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, 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, 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).

Declaration under Rule 4.17:
— *of inventorship (Rule 4.17(iv))*

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.

WO 2007/071996 A2

(54) **Title:** TWIN-ARGININE TRANSLOCATION (TAT) STREPTOMYCES SIGNAL SEQUENCES

(57) **Abstract:** Described herein are novel Tat signal polypeptides and methods for using the Tat signal polypeptides for producing heterologous polypeptides. A novel reporter assay for testing the biological activity of the secreted proteins is also described.

TWIN-ARGININE TRANSLOCATION (TAT) STREPTOMYCES SIGNAL SEQUENCES

FIELD OF THE INVENTION

This invention relates to Twin-arginine translocation (Tat) signal peptides, which have been identified in the cell wall associated fraction of the gram-positive soil bacterium *Streptomyces coelicolor*. The invention further relates to fusion polypeptides comprising a TAT signal peptide and a heterologous polypeptide.

BACKGROUND

In prokaryotes two pathways for protein translocation across the cytoplasmic membrane have been recognized. In most bacteria the general secretory (Sec) pathway is the best-characterized route for protein export. Proteins exported by this pathway are translocated across the membrane in an unfolded state through a membrane-embedded translocon to which they are targeted by cleavable N-terminal signal peptides (Mori et al., (2001) *Trends in Microbiology* 9:494 – 500). More recently a second general export pathway has been described, which is designated the twin-arginine translocation (Tat) pathway and reference is made to US 2002/0110860; WO 03/079007; Berks, B. C. (1996) *Mol. Microbiol.* 22:393-404 and Tjalsma et al., (2000) *Microbiol. & Molecul. Bio. Reviews* 64:515 - 547. Unlike the Sec system, the Tat system is involved in the transport of pre-folded protein substrates (Thomas et al., (2001) *Mol. Microbiol* 39:47 – 53).

Proteins are targeted to the Tat pathway by possession of N-terminal tripartite signal peptides. The signal peptides include a conserved twin-arginine motif in the N-region of Tat signal peptide. The motif has been defined as R-R-x-Φ-Φ, where Φ represents a hydrophobic amino acid. In *E. coli* the Tat pathway comprises the three-membrane protein TatA, TatB and TatC. A fourth protein TatE forms a minor component of the Tat machinery and has a similar function to TatA. Studies by several groups suggest that the major role of the Tat pathway is in the translocation of redox proteins that integrate their cofactors in the cytoplasm. Other more recent studies indicate that the Tat pathway may play a broader role in protein secretion (Ochsner et al., (2001) Proc. Natl. Acad. Sci. USA 99:8312 – 8317). Because of the ability to secrete pre-folded protein substrates, the Tat pathway represents a significant mechanism for secreting a high level of heterologous proteins.

Estimates of Tat substrates in organisms other than *Bacillus subtilits* and *E. coli* have been based predominantly in *in silico* analysis of genome sequences using programs

trained to recognize specific features of tat targeting sequences. While these programs are useful tools for identifying candidate Tat substrates encoded within bacterial genomes experimental verification of the *in silico* predicted sequences other than phenotype analysis has been lacking.

Many of these predicted Tat substrates are in microorganisms of the Streptomyces genes such as *S. coelicolor* and *S. avermitilis*. *Streptomyces* are gram-positive spore forming microorganisms which produce a range of diverse and important secondary metabolites, including many commercially available antibiotics. *Streptomyces* are important in the field of biotechnology because they are prolific protein secretors.

Prior in silico predictions estimate between 145 and 189 proteins from S. coelicolor may be Tat dependent. The present invention demonstrates the verification and importance of the Tat secretory pathway in *Streptomyces* and further is directed to Tat signal peptides which comprise a motif outside of the previous identified Tat signal peptide motif. Further the Tat signal peptides according to the invention maybe useful in the secretion of heterologous proteins in *Streptomyces*.

SUMMARY OF THE INVENTION

Provided herein are novel Tat signal peptides and methods of using the novel Tat signal peptides for producing heterologous polypeptides in a bacterial host cell as described in the appended claims.

In a first aspect, the invention is directed to novel Twin Arginine Translocation (TAT) signal peptides.

In one aspect, the invention is directed to an isolated signal peptide comprising the TAT signal sequence motif $(X^{-1})RR(X^{+2})(X^{+3})(X^{+4})$, wherein R is arginine, X^{-1} is amino acid M, H, A, P, K, R, N, T, G, S, D, Q or E; X^{+2} is amino acid A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is Q, I, L, V, M or F, and wherein said motif is not within the first 35N' terminal residues of the amino acid sequence of the polypeptide.

In another aspect, the invention is directed to an isolated TAT signal peptide comprising the sequence motif $(X^{-1})RR(X^{+2})(X^{+3})(X^{+4})$, wherein R is arginine, X^{-1} is amino acid H, A, P, K, R, N, T, G, S, D, Q E or L; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is T, G or A, and wherein the motif is within the first 35N' terminal residues of the amino acid sequence of the polypeptide. In some embodiments, the TAT signal peptide comprises a sequence motif $(X^{-1})RR(X^{+2})(X^{+3})(X^{+4})$ that is within the first 35N' terminal residues, wherein when X^{-1} is H then X^{+4} is A. In other

embodiments, the TAT signal peptide comprises a sequence motif $(X^{-1})RR(X^{+2})(X^{+3})(X^{+4})$ that is within the first 35N' terminal residues, wherein when X^{-1} is L then X^{+4} is G.

In a further aspect, the invention is directed to an isolated TAT signal peptide comprising the sequence motif $(X^{-1})RR(X^{+2})(X^{+3})$ (X^{+4}), wherein RR represents two adjacent arginine residue and X designates positions restrict to other selected amino acids: X^{-1} is M, H, A, P, K, R, N, T, G, S, D, Q, or E; X^{+2} is a polar amino acid residue; and X^{+3} and X^{+4} are non-polar amino acid residues, and wherein the motif is not within the first 35N' terminal residues of the amino acid sequence of the polypeptide. In some embodiments, the TAT signal polypeptide is a variant of the polypeptide having the TAT motif that is not within the first 35N' terminal residues of the amino acid sequence of the signal peptide.

Other embodiments of the invention comprise TAT signal peptides comprising the amino acid sequences of TAT signal peptides of proteins SCO2286 (SEQ ID NO: 218), SCO3790long (SEQ ID NO: 227), and SCO6580long (SEQ ID NO: 241), SCO1590 (SEQ ID NO: 211), SCO1824 (SEQ ID NO: 213), SCO6580short (SEQ ID NO: 182), or SCO3790short (SEQ ID NO: 122).

In another aspect, the invention is directed to an isolated polynucleotide sequence comprising a polynucleotide sequence encoding a TAT signal peptide of the invention.

In some embodiments, the isolated polynucleotide sequence is a nucleic acid molecule comprising a first nucleotide sequence encoding a TAT signal sequence encompassed by the invention operably linked to a second nucleotide sequence encoding a heterologous polypeptide. In other embodiments, the invention is directed to a nucleic acid molecule comprising a first nucleotide sequence encoding a TAT signal sequence encompassed by the invention operably linked to a second nucleotide sequence encoding a homologous polypeptide.

In a further aspect, the invention is directed to an expression vector comprising a first nucleotide sequence encoding a TAT signal sequence encompassed by the invention operably linked to a second nucleotide sequence encoding a heterologous polypeptide. In one embodiment of this aspect, the invention is directed to a bacterial host cell host cell that is genetically transformed with the recombinant expression vector encompassed by the invention.

In another aspect, the invention is directed to a fusion polypeptide comprising a TAT secretion signal peptide encompassed by the invention and a heterologous polypeptide. In an embodiment of this aspect, the fusion polypeptide comprises a TAT signal peptide that is naturally expressed in *Streptomyces*. In another embodiment of this aspect, the heterologous polypeptide is an enzyme, growth factor or hormone. In yet

another embodiment, the enzyme may be a protease, a carbohydrase, such as amylases, cellulases, xylanases, and lipases; an isomerase, such as racemases, epimerases, tautomerases, or mutases; a transferase; a glucoamylase; a kinase, an amidase, an esterase, or an oxidase. In other embodiments of this aspect, the heterologous polypeptide is not naturally associated with any secretion signal peptide.

In yet another aspect, the invention is directed to a method for producing a heterologous polypeptide comprising culturing host cells in culture medium under conditions suitable for producing said polypeptide, said host cells comprising an expression vector comprising a first nucleotide sequence encoding a TAT signal peptide encompassed by the invention operatively linked to a second nucleotide sequence encoding a heterologous polypeptide, and producing said heterologous polypeptide. In some embodiments, the method uses a TAT signal peptide of the invention that comprises the sequence motif $(X^{-1})RR(X^{+2})(X^{+3})(X^{+4})$, wherein R is arginine, X^{-1} is amino acid M, H, A, P, K, R, N, T, G, S, D, Q or E; X^{+2} is amino acid A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is Q, I, L, V, M or F, and wherein said motif is not within the first 35N' terminal residues of the amino acid sequence of the polypeptide.

In other embodiments, the heterologous polypeptide that is produced by the method of the invention includes a TAT signal peptide that comprises the sequence motif $(X^{-1})RR(X^{+2})(X^{+3})(X^{+4})$, wherein R is arginine, X^{-1} is amino acid H, A, P, K, R, N, T, G, S, D, Q E or L; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is T, G or A, and wherein the motif is within the first 35N' terminal residues of the amino acid sequence of the polypeptide. In some embodiments of this aspect, the step of producing the heterologous polypeptide comprises recovering the polypeptide from the culture medium.

In other embodiments of this aspect, the host cell is a prokaryotic cell. In other embodiments, the host cell is a *Streptomyces* bacterial cell. In yet other embodiments, the host cell is a *S. coelicolor* or an *S. lividans* bacterial cell. In further embodiments of this aspect, the method of the invention produces a heterologous polypeptide that can be an enzyme a growth factor or a hormone.

In yet another aspect, the invention is directed to a method for secreting a heterologous protein from a host microorganism comprising operably ligating a nucleotide acid sequence encoding the heterologous protein to a TAT signal sequence encompassed by the invention and inserting the ligated pair into an expression vector in a host microorganism, expressing the heterologous protein under the control of the TAT signal and secreting the heterologous protein from the microorganism by the TAT expression

pathway. In some embodiments of this aspect, the expressed heterologous protein is secreted in its correctly –folded active form. In some embodiments of this aspect, the host organism is a *Streptomyces* strain, for example a *S. lividans* strain.

In another aspect, the invention provides for a method of identifying TAT signal peptides of polypeptides secreted in a microorganism comprising identifying a TAT signal peptide of a secreted polypeptide and validating the ability of said signal peptide to secrete a biological functional polypeptide. In some aspects, testing the validity of the TAT signal peptide comprises expressing a fusion polypeptide in a host microorganism, comprising the TAT signal sequence operably linked to a heterologous polypeptide and testing the biological activity of the expressed heterologous polypeptide. In some embodiments, the heterologous polypeptide is agarase.

Additional aspects and embodiments of the invention are set forth in the detailed discussion, examples and figures which follow, which are intended for illustrative purposes only and are not intended in any way to limit the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 A-B. The *tatC* mutant of *S. coelicolor* has pleiotropic phenotypes. Colonies of the *S. coelicolor* wildtype (left hand plate) and *ΔtatC* strain TP4 (right hand plate) cultured on either MS (A), or R5 (B), media.

FIG. 2 A-B illustrates a 2-dimensional gel analysis of extracellular protein fractions isolated from the *S. coelicolor* M145 wild type (A) and a *ΔtatC* mutant derivative (B). Strains were cultured on R5, and proteins associated with the cell wall were separated in the first dimension by isoelectric focusing (pH gradient 4 – 7) and in the second dimension by SDS PAGE. Protein spots that are circled represent proteins that migrated at specific positions in the extracellular fractions from the wild type strain but were absent from the corresponding position in extracellular fractions of the *ΔtatC* strain. These proteins spots were identified by in gel tryptic digest followed by mass spectrometry and identities of the proteins are indicated by the SCO number.

FIG. 3 A-D illustrates a 2-dimensional gel analysis of extracellular fractions isolated from strains grown on Complete medium (CM) (A and B) and Mannitol Soya (MS) (C and D). Proteins from the wild type S coelicolor M145 strain are shown in A and C, and proteins from the *tatC* mutant strains are shown in panels B and D. Protein spots that are circled represent proteins that migrated at specific positions in the extracellular fractions from the wild type strain but were absent from the corresponding position in extracellular fractions of the *ΔtatC* strain. These proteins spots were identified by in gel tryptic digest

followed by mass spectrometry and identities of the proteins are indicated by the SCO number.

FIG. 4 A-C Agarase is a Tat substrate. In A) the agarase signal peptide with the consecutive arginine residues of the twin-arginine motif are highlighted in bold and underlined. In B) S. coelicolor strain M145 (WT) and TP1 (*ΔtatC*::Apra^R) were grown on MM-C minimal medium for 5 days and stained with Lugol solution. In C) strains M145 (Wt) and TP4 (*ΔtatC*) harboring either pIJ6902 or pIJ6902-dagA in single copy were grown on MM containing glucose, apramycin and thiostrepton for 72 hours and prior to staining with Lugol solution.

FIG. 5 illustrates the export of agarase activity mediated by *S. coelicolor* signal peptides some of which are encompassed by the instant invention. The signal peptide is fused to the mature sequence of agarase. The y-axis shows the signal peptides from a range of cell wall associated *S. coelicolor* proteins (listed in Tables 3 and 4) gives the % agarase activity of the secreted protein for each fusion protein when compared with agarase including it native signal peptide (set at 100%). Each construct carries the same native agarase promoter and ribosome-binding site and therefore, the activity is a measure of the efficientcy of export directed by each particular signal peptide. The assay was verified using negative controls, signal peptides that do not posses twin-arginines in the signal peptide and were strongly detected in the extracellular fractions of both the wild type (M145) and *ΔtatC* mutant strain by MudPIT analysis. None of these signal peptides mediated extracellular agarase activity. The signal peptides from the following proteins were also tested and were found to be negative in this assay: SCO0432, SCO0474, SCO0494, SCO0930, SCP1230, SCO1396, SCO1824, ACO1948, ACO1968, SCO2226, SCO2383, SCO2446, SCO2591, SCO2637, SCO2786, SACO2821, SCO3456, SCO4010, SCO4142, SCO4152, SCO4884, SCO4885, SCO5074, SCO5113, SCO5447, SCO5461, SCO6009, SCO6644, SCO6738, and SCO7399. The symbol * annotates that two versions of the annotated signal peptides (designated long and short), representing two alternative start sites, were tested (see Tables 3 and 4).

FIG. 6 illustrates the Tat-dependent export of agarase mediated by *S. coelicolor* signal peptides. For each plate, the signal peptide of each designated SCO protein, fused in frame with the mature sequence of agarase, was expressed in two *S. lividans tat+* strains (1326, lower left on each plate, or 10-164, lower right of each plate), or in the 10-164 isogenic *tatC* strain (top of each plate). Strains were cultured on minimal medium containing 0.5% glucose for 5 days and stained with lugol solution. (*Top Left*) plate (DagA) corresponds to agarase expressed from an identical construct with its native signal peptide.

Note that although expression of agarase in *S. coelicolor* is highly inhibited by glucose, heterologous expression of agarase in *S. lividans*, which is *dagA*-, is not (Servin-Gonzalez L, Jensen MR, White J, Bibb M (1994) *Microbiology* 140:2555-2565).

DETAILED DESCRIPTION OF THE INVENTION

The invention will now be described in detail by way of reference only using the following definitions and examples. All patents and publications, including all sequences disclosed within such patents and publications, referred to herein are expressly incorporated by reference.

Unless defined otherwise herein, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Singleton, *et al.,* DICTIONARY OF MICROBIOLOGY AND MOLECULAR BIOLOGY, 2D ED., John Wiley and Sons, New York (1994), and Hale & Marham, THE HARPER COLLINS DICTIONARY OF BIOLOGY, Harper Perennial, NY (1991) provide one of skill with a general dictionary of many of the terms used in this invention. Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, the preferred methods and materials are described. Numeric ranges are inclusive of the numbers defining the range. Unless otherwise indicated, nucleic acids are written left to right in 5' to 3' orientation; amino acid sequences are written left to right in amino to carboxy orientation, respectively. Practitioners are particularly directed to Sambrook *et al.,* 1989, and Ausubel FM *et al.,* 1993, for definitions and terms of the art. It is to be understood that this invention is not limited to the particular methodology, protocols, and reagents described, as these may vary.

The headings provided herein are not limitations of the various aspects or embodiments of the invention which can be had by reference to the specification as a whole. Accordingly, the terms defined immediately below are more fully defined by reference to the specification as a whole.

The term "polypeptide" as used herein refers to a compound made up of a single chain of amino acid residues linked by peptide bonds. The term "protein" as used herein may be synonymous with the term "polypeptide" or may refer, in addition, to a complex of two or more polypeptides.

The term “fusion polypeptide” or”Tat fusion polypeptide” as used herein refers to a Tat signal peptide linked to the protein of interest. It is understood that a protein of interest refers to a heterologous protein that is operably linked to the Tat signal peptide of the

invention.

A "signal peptide" as used herein refers to an amino-terminal extension on a protein to be secreted. Nearly all secreted proteins use an amino-terminal protein extension which plays a crucial role in the targeting to and translocation of precursor proteins across the membrane and which is proteolytically removed by a signal peptidase during or immediately following membrane transfer.

A "Tat signal peptide" refers to a N-terminally extended sequence which includes two consecutive arginine residues and which functions in the secretion of proteins in prefolded confirmation. A "Tat signal peptide" may be interchangeably referred to as "Tat peptide" or "Tat polypeptide".

A "tagged Tat fusion polypeptide" herein refers to a fusion polypeptide, which comprises a Tat signal peptide and a heterologous peptide, to which a tag sequence can be linked and used to identify transformants and/or to facilitate the purification of recombinant Tat fusion polypeptides.

As used herein, a "protein of interest" or "polypeptide of interest" refers to the protein to be expressed and secreted by the host cell. The protein of interest may be any protein which up until now has been considered for expression in prokaryotes. The protein of interest may be either homologous or heterologous to the host. In the first case over expression should be read as expression above normal levels in said host. In the latter case basically any expression is of course over expression.

As used herein, the term "heterologous protein" refers to a protein or polypeptide that does not naturally occur in a host cell. Examples of heterologous proteins include enzymes such as hydrolases including proteases, cellulases, amylases, other carbohydrases, and lipases; isomerases such as racemases, epimerases, tautomerases, or mutases; transferases, kinases and phophatases. The heterologous gene may encode therapeutically significant proteins or peptides, such as growth factors, cytokines, ligands, receptors and inhibitors, as well as vaccines and antibodies. The gene may encode commercially important industrial proteins or peptides, such as proteases, carbohydrases such as amylases and glucoamylases, cellulases, oxidases and lipases. The gene of interest may be a naturally occurring gene, a mutated gene or a synthetic gene.

The term "homologous protein" refers to a protein or polypeptide native or naturally occurring in a host cell. The invention includes host cells producing the homologous protein via recombinant DNA technology. The present invention encompasses a host cell having a deletion or interruption of the nucleic acid encoding the naturally occurring homologous protein, such as a protease, and having nucleic acid encoding the homologous protein re-

introduced in a recombinant form. In another embodiment, the host cell produces the homologous protein.

The term "polynucleotide" or "nucleic acid molecule" includes RNA, DNA and cDNA molecules. As used herein, the term refers to a polymeric form of nucleotides of any length, either ribonucleotides or deoxyribonucleotides. This term refers only to the primary structure of the molecule and thus includes double- and single-stranded DNA and RNA. It also includes known types of modifications, for example, labels which are known in the art, methylation, "caps", substitution of one or more of the naturally occurring nucleotides with an analog, internucleotide modifications such as, for example, those with uncharged linkages (e.g., methyl phosphonates, phosphotriesters, phosphoamidates, carbamates, etc.) and with charged linkages (e.g., phosphorothioates, phosphorodithioates, etc.), those containing pendant moieties, such as, for example proteins (including for e.g., nucleases, toxins, antibodies, signal peptides, poly-L-lysine, etc.), those with intercalators (e.g., acridine, psoralen, etc.), those containing chelates (e.g., metals, radioactive metals, boron, oxidative metals, etc.), those containing alkylators, those with modified linkages (e.g., alpha anomeric nucleic acids, etc.), as well as unmodified forms of the polynucleotide. Generally, nucleic acid segments provided by this invention may be assembled from fragments of the genome and short oligonucleotide linkers, or from a series of oligonucleotides, or from individual nucleotides, to provide a synthetic nucleic acid which is capable of being expressed in a recombinant transcriptional unit comprising regulatory elements derived from a microbial or viral operon, or a eukaryotic gene.

A "heterologous" nucleic acid construct or sequence has a portion of the sequence which is not native to the cell in which it is expressed. Heterologous, with respect to a control sequence refers to a control sequence (*i.e.* promoter or enhancer) that does not function in nature to regulate the same gene the expression of which it is currently regulating. Generally, heterologous nucleic acid sequences are not endogenous to the cell or part of the genome in which they are present, and have been added to the cell, by infection, transfection, microinjection, electroporation, or the like. A "heterologous" nucleic acid construct may contain a control sequence/DNA coding sequence combination that is the same as, or different from a control sequence/DNA coding sequence combination found in the native cell.

As used herein, the term "vector" refers to a nucleic acid construct designed for transfer between different host cells. An "expression vector" refers to a vector that has the ability to incorporate and express heterologous DNA fragments in a foreign cell. Many

prokaryotic and eukaryotic expression vectors are commercially available. Selection of appropriate expression vectors is within the knowledge of those having skill in the art.

Accordingly, an "expression cassette" or "expression vector" is a nucleic acid construct generated recombinantly or synthetically, with a series of specified nucleic acid elements that permit transcription of a particular nucleic acid in a target cell. The recombinant expression cassette can be incorporated into a plasmid, chromosome, mitochondrial DNA, plastid DNA, virus, or nucleic acid fragment. Typically, the recombinant expression cassette portion of an expression vector includes, among other sequences, a nucleic acid sequence to be transcribed and a promoter.

As used herein, the term "plasmid" refers to a circular double-stranded (ds) DNA construct used as a cloning vector, and which forms an extrachromosomal self-replicating genetic element in many bacteria and some eukaryotes.

As used herein, the term "selectable marker-encoding nucleotide sequence" refers to a nucleotide sequence which is capable of expression in mammalian cells and where expression of the selectable marker confers to cells containing the expressed gene the ability to grow in the presence of a corresponding selective agent.

As used herein, the term "promoter" refers to a nucleic acid sequence that functions to direct transcription of a downstream gene. The promoter will generally be appropriate to the host cell in which the target gene is being expressed. The promoter together with other transcriptional and translational regulatory nucleic acid sequences (also termed "control sequences") are necessary to express a given gene. In general, the transcriptional and translational regulatory sequences include, but are not limited to, promoter sequences, ribosomal binding sites, transcriptional start and stop sequences, translational start and stop sequences, and enhancer or activator sequences.

"Chimeric gene" or "heterologous nucleic acid construct", as defined herein refers to a non-native gene (*i.e.*, one that has been introduced into a host) that may be composed of parts of different genes, including regulatory elements. A chimeric gene construct for transformation of a host cell is typically composed of a transcriptional regulatory region (promoter) operably linked to a heterologous protein coding sequence, or, in a selectable marker chimeric gene, to a selectable marker gene encoding a protein conferring antibiotic resistance to transformed cells. A typical chimeric gene of the present invention, for transformation into a host cell, includes a transcriptional regulatory region that is constitutive or inducible, a signal peptide coding sequence, a protein coding sequence, and a terminator sequence. A chimeric gene construct may also include a second DNA sequence encoding a signal peptide if secretion of the target protein is desired.

A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For example, DNA encoding a secretory leader is operably linked to DNA for a polypeptide if it is expressed as a preprotein that participates in the secretion of the polypeptide; a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the sequence; or a ribosome binding site is operably linked to a coding sequence if it is positioned so as to facilitate translation. Generally, "operably linked" means that the DNA sequences being linked are contiguous, and, in the case of a secretory leader, contiguous and in reading phase. However, enhancers do not have to be contiguous. Linking is accomplished by ligation at convenient restriction sites. If such sites do not exist, the synthetic oligonucleotide adaptors or linkers are used in accordance with conventional practice.

As used herein, the term "gene" means the segment of DNA involved in producing a polypeptide chain, that may or may not include regions preceding and following the coding region, *e.g.* 5' untranslated (5' UTR) or "leader" sequences and 3' UTR or "trailer" sequences, as well as intervening sequences (introns) between individual coding segments (exons).

As used herein, "recombinant" includes reference to a cell or vector, that has been modified by the introduction of a heterologous nucleic acid sequence or that the cell is derived from a cell so modified. Thus, for example, recombinant cells express genes that are not found in identical form within the native (non-recombinant) form of the cell or express native genes that are otherwise abnormally expressed, under expressed or not expressed at all as a result of deliberate human intervention.

As used herein, the terms "transformed", "stably transformed" or "transgenic" with reference to a cell means the cell has a non-native (heterologous) nucleic acid sequence integrated into its genome or as an episomal plasmid that is maintained through two or more generations.

As used herein, the term "expression" refers to the process by which a polypeptide is produced based on the nucleic acid sequence of a gene. The process includes both transcription and translation.

The term "introduced" in the context of inserting a nucleic acid sequence into a cell, means "transfection", or "transformation" or "transduction" and includes reference to the incorporation of a nucleic acid sequence into a eukaryotic or prokaryotic cell where the nucleic acid sequence may be incorporated into the genome of the cell (for example, chromosome, plasmid, plastid, or mitochondrial DNA), converted into an autonomous replicon, or transiently expressed (for example, transfected mRNA).

The terms "isolated" or "purified" as used herein refer to a nucleic acid or polypeptide that is removed from at least one component with which it is naturally associated.

As used herein, "substantially equivalent" can refer both to nucleotide and amino acid sequences, for example a mutant sequence, that varies from a reference sequence by one or more substitutions, deletions, or additions, the net effect of which does not result in an adverse functional dissimilarity between the reference and subject sequences. Typically, such a substantially equivalent sequence varies from one of those listed herein by no more than about 20% (i.e., the number of individual residue substitutions, additions, and/or deletions in a substantially equivalent sequence, as compared to the corresponding reference sequence, divided by the total number of residues in the substantially equivalent sequence is about 0.2 or less). Such a sequence is said to have 80% sequence identity to the listed sequence. In one embodiment, a substantially equivalent, e.g., mutant, sequence of the invention varies from a listed sequence by no more than 10% (90% sequence identity); in a variation of this embodiment, by no more than 5% (95% sequence identity); and in a further variation of this embodiment, by no more than 2% (98% sequence identity). Substantially equivalent, e.g., mutant, amino acid sequences according to the invention generally have at least 95% sequence identity with a listed amino acid sequence, whereas substantially equivalent nucleotide sequence of the invention can have lower percent sequence identities, taking into account, for example, the redundancy or degeneracy of the genetic code. For the purposes of the present invention, sequences having substantially equivalent biological activity and substantially equivalent expression characteristics are considered substantially equivalent. As used herein, the term “activity” or “biological activity” refers to an activity associated with a particular protein, such as enzymatic activity. Biological activity refers to any activity that would normally be attributed to that protein by one skilled in the art.

The recognized single letter codes for amino acid residues are used consistently herein wherein alanine (A); arginine (R); asparagine (N); aspartic acid (D); cysteine (C); glutamic acid (E); glutamine (Q); glycine (G); histidine (H); isoleucine (I); leucine (L); lysine (K); methionine (M); phenylalanine (F); proline (P); serine (S); threonine (T); tryptophan (W); tyrosine (Y) and valine (V).

The term “TATFIND”, and “TATFIND1.2” refer to a Tat substrate recognition program developed to detect putative Tat substrates in bacterial genomes. In general, the program is based on the position and sequences of the Tat motif (WO 03/079007). The motif as disclosed in WO 03/079007 is within the first 35 amino acids of a protein

sequence, $(X^{-1})R^0R^{+1}(X^{+2})(X^{+3})$ (X^{+4}), wherein X^{-1} is M, H, A, P, K, R, N, T, G, S, D, Q or E; R^0R^{+1} represent the twin-arginines; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T (positively charged residues being excluded from this position) and X^{+4} is Q, I, L, V, M or F.

The term "TatP" refers to a Tat substrate recognition program that recognizes the same Tat motif as TATFIND. The TatP program partially uses a neural network as well as a rule based classification used in the TATFIND program and reference is made to Bendtsen, J.D. et al., (2005), BMC Bioinformatics 6:167 – 175.

The present invention provides novel gram-positive microorganism secretion factors and methods that can be used in microorganisms to provide protein secretion and the production of proteins in secreted form.

Tat Nucleic Acids and Amino Acids

The invention is based on the discovery of novel Tat signal peptides.

The invention comprises isolated Tat signal peptides that include, but are not limited to, a polypeptide comprising the amino acid sequence set forth as SEQ ID NO: 17, 20, 23 26 29 32 35, 38, 41, 44, 47, 50,53, 56, 59, 62, 65, 68, 71, 74, 77, 80, 83, 86, 89, 92, 95, 98, 101, 104, 107, 110, 113 116, 119, 122, 125, 128 131, 134, 137, 140, 143, 146, 149, 152, 155, 158, 161, 164, 167, 170, 173, 176, 179, 182, 185, 188, 191, 194, 197, 200, 203, and 204-253. The Tat signal peptides include polypeptides comprising an amino acid sequence encoded by any one nucleotide sequence generated by amplifying a nucleic acid using polymerase chain reaction (PCR) primer pairs having SEQ ID NO: 15 and 16, 18 and 19, 21 and 22, 24 and 25, 27 and 28, 30 and 31, 33 and 34, 36 and 37, 39 and 40, 42 and 43, 45 and 46, 48 and 49, 51 and 52, 54 and 55, 57 and 58, 60 and 61, 63 and 64, 66 and 67, 69 and 70, 72 and 73, 75 and 76, 78 and 79, 81 and 82, 84 and 85, 87 and 88, 90 and 91, 93 and 94, 96 and 97, 99 and 100, 102 and 103, 105 and 106, 108 and 109, 111 and 112, 114 and 115, 117 and 118, 120 and 121, 123 and 124, 126 and 127, 129 and 130, 132 and 133, 135 and 136, 138 and 139, 141 and 142, 144 and 145, 147 and 148, 150 and 151, 153 and 154, 156 and 157, 159 and 160, 162 and 163, 165 and 166, 168 and 169, 171 and 172, 174 and 175, 177 and 178, 180 and 181, 183 and 184, 186 and 187, 189 and 190, 192 and 193, 195 and 196, 198 and 199, and 201 and 202.

The present invention is directed to certain novel Tat signal peptides that comprise sequences which include the motif $(X^{-1})RR(X^{+2})(X^{+3})$ (X^{+4}), wherein RR represents two adjacent arginine residue and X designates positions restrict to other selected amino acids: X^{-1} is M, H, A, P, K, R, N, T, G, S, D, Q, or E; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is

I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is Q, I, L, V, M or F and wherein the motif is not within the first 35N' terminal residues of the amino acid sequence of the polypeptide.

Tat signal sequences of the invention include the motif $(X^{-1})RR(X^{+2})(X^{+3})$ (X^{+4}), wherein RR represents two adjacent arginine residue and X designates positions restrict to other selected amino acids: X^{-1} is M, H, A, P, K, R, N, T, G, S, D, Q, or E; X^{+2} is a polar amino acid residue; and X^{+3} and X^{+4} are non-polar amino acid residues, and wherein the motif is not within the first 35N' terminal residues of the amino acid sequence of the polypeptide.

In the present context, non-polar amino acids are alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, and methionine; and polar amino acids include glycine, serine, threonine, cysteine, tyrosine, asparagine, glutamine, arginine, lysine, histidine, aspartic acid and glutamic acid.

Some preferred Tat signal peptides encompassed by the invention include, but are not limited to the Tat signal peptides of proteins SCO2286 - MTPANHQAPTSAPSPAPSQSSHAPELRAAARSLG**RR**RFLTVTGAAAALAFAVNLPAAGTA SAA (SEQ ID NO: 218) and SCO3790 long - MRKLLPLIGTPSGSHPGGRSAMTCRFRCGDACFHE**V**PNTSSNEYVGDVIAGALS**RR**SMM RAAAVVTVAAAGAGAVGVAGAPSAQAA (SEQ ID NO: 227).

Another Tat signal peptide comprises the Tat signal peptide of protein SCO6580 long - MTPFTDSSRTDA GTDPSADGPGESLRRALGVN**RR**RFLSTCTAVAAGAVAAPVFGASPALAH (SEQ ID NO: 241).

In addition, the invention contemplates isolated variants of the Tat signal peptides of the invention.

In some embodiments, the variant of a Tat signal peptide is encoded by an alternative start site that leads to the translation of a shorter Tat signal peptide. Variant Tat signal peptides include, but are not limited to the Tat signal peptide of protein SCO3790, designated SCO3790short (SEQ ID NO: 122), and the Tat signal peptide of protein SCO6580, designated SCO6580 short - (SEQ ID NO: 182).

The invention also contemplates Tat signal peptides that comprise the motif $(X^{-1})RR(X^{+2})(X^{+3})$ (X^{+4}), wherein RR represents two adjacent arginine residue and X designates positions restrict to other selected amino acids: X^{-1} is H, A, P, K, R, N, T, G, S, D, Q E or L; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is T, G or A and wherein the motif is within the first 35N' terminal residues of the amino acid sequence of the polypeptide. In some embodiments when X^{-1} is S then X^{+4}

will be T; in other embodiments when X^{-1} is H then X^{+4} will be A, and in other embodiments when X^{-1} is L then X^{+4} will be G. The primary amino acid sequence of some of the Tat signal peptides dictates that the amino acid residues at positions X^{+3} and X^{+4} are non-polar amino acid residues. Thus, the invention comprises Tat signal sequences that include the motif $(X^{-1})RR(X^{+2})(X^{+3})$ (X^{+4}), wherein RR represents two adjacent arginine residue and X designates positions restrict to other selected amino acids: X^{-1} is H, A, P, K, R, N, T, G, S, D, Q E or L; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; and X^{+3} and X^{+4} are non-polar amino acid residues.

Preferred Tat signal peptides include, but are not limited to the Tat signal sequence of protein SCO1590 - MGGVSRRAFTVAALSAFTLVPEASAA (SEQ ID NO: 211) and Tat signal sequence of protein SCO1824 - MTAPLSR HRRALAIPAGLAVAASLAFLP GTPAAATPAAEAA (SEQ ID NO: 213).

The invention also provides biologically active variants of any of the amino acid sequences set forth as SEQ ID NO: 17, 20, 23 26 29 32 35, 38, 41, 44, 47, 50,53, 56, 59, 62, 65, 68, 71, 74, 77, 80, 83, 86, 89, 92, 95, 98, 101, 104, 107, 110, 113 116, 119, 122, 125, 128 131, 134, 137, 140, 143, 146, 149, 152, 155, 158, 161, 164, 167, 170, 173, 176, 179, 182, 185, 188, 191, 194, 197, 200, 203, and 204-253, and substantial equivalents thereof. Substantial equivalent Tat signal peptides have at least about 80%, or 85%, more typically at least about 90%, 91%, 92%, 93%, or 94% and even more typically at least about 95%, 96%, 97%, 98% or 99% amino acid identity, and that retain biological activity. The biological activity of a signal peptide refers to the ability of the signal peptide to translocate the polypeptide to the extracellular space. Fragments of the Tat peptides of the present invention which are capable of exhibiting biological activity are also encompassed by the present invention. The term "variant" when made in reference to a polypeptide refers to any polypeptide differing from naturally occurring polypeptides by amino acid insertions, deletions, and substitutions, or combinations thereof, which can be created using, e g., recombinant DNA techniques. A variant of a polypeptide can also refer to a shortened version of a polypeptide that is generated by use of an alternative start codon as a translation initiation site. The alternative codon can be an in-frame or an out-of-frame start codon. Examples of variants of polypeptides that are translated from alternative start codons include but are not limited to the TAT signal polypeptides of SEQ ID NO: 122 and SEQ ID NO: 182.

Guidance in determining which amino acid residues may be replaced, added or deleted without abolishing activities of interest, may be found by comparing the sequence of the particular polypeptide with that of homologous peptides and minimizing the number

of amino acid sequence changes made in regions of high homology (conserved regions) or by replacing amino acids with consensus sequence. Preferably, amino acid "substitutions" are the result of replacing one amino acid with another amino acid having similar structural and/or chemical properties, i.e., conservative amino acid replacements. "Conservative" amino acid substitutions may be made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues involved. For example, nonpolar (hydrophobic) amino acids include alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, and methionine; polar neutral amino acids include glycine, serine, threonine, cysteine, tyrosine, asparagine, and glutamine; positively charged (basic) amino acids include arginine, lysine, and histidine; and negatively charged (acidic) amino acids include aspartic acid and glutamic acid. "Insertions" or "deletions" are typically in the range of about 1 to 5 amino acids. The variation allowed may be experimentally determined by systematically making insertions, deletions, or substitutions of amino acids in a polypeptide molecule using recombinant DNA techniques and assaying the resulting recombinant variants for activity.

Where alteration of function is desired, insertions, deletions or non-conservative alterations can be engineered to produce altered polypeptides. Such alterations can, for example, alter one or more of the biological functions or biochemical characteristics of the polypeptides of the invention. For example, such alterations may change polypeptide characteristics such as ligand-binding affinities, interchain affinities, or degradation/turnover rate. Further, such alterations can be selected so as to generate polypeptides that are better suited for expression, scale up and the like in the host cells chosen for expression.

The present invention also provides isolated Tat peptides encoded by the nucleic acids/polynculeotides of the present invention or by degenerate variants of the nucleic acids of the invention. By “degenerate variants” is intended nucleotide fragments that differ from a nucleic acid of the invention by nucleotide sequence but, due to the degeneracy of the genetic code, encode an identical polypeptide sequence.

The invention provides for fusion or chimeric polypeptides. As used herein, a fusion protein or fusion polypeptide comprises a Tat signal peptide operatively linked to a polypeptide/protein of interest. Within the fusion protein, the term “operatively linked” is intended to indicate that the Tat signal polypeptide and the polypeptide of interest are fused in-frame with one another. The Tat signal peptides are fused to the N-terminal end of the peptide of interest. Polypeptides of interest include homologous or heterologous polypeptides. Polypeptides of interest include full-length polypeptides that are naturally synthesized with a signal peptide, the mature form of the full-length polypeptides, and

polypetides that naturally lack a signal peptide. The fusion polypeptides thus comprise a Tat signal sequence that includes the motif $(X^{-1})RR(X^{+2})(X^{+3})$ (X^{+4}), wherein RR represents two adjacent arginine residue and X designates positions restrict to other selected amino acids: X^{-1} is M, H, A, P, K, R, N, T, G, S, D, Q, or E; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is Q, I, L, V, M or F and wherein the motif is not within the first 35N' terminal residues of the amino acid sequence of the polypeptide. In other embodiments Tat signal sequences comprised by the fusion polypepitde can include the motif $(X^{-1})RR(X^{+2})(X^{+3})$ (X^{+4}), wherein RR represents two adjacent arginine residue and X designates positions restrict to other selected amino acids: X^{-1} is M, H, A, P, K, R, N, T, G, S, D, Q, or E; X^{+2} is a polar amino acid residue; and X^{+3} and X^{+4} are non-polar amino acid residues, and wherein the motif is not within the first 35N' terminal residues of the amino acid sequence of the polypeptide.

As recited herein, in the present context, non-polar amino acids are alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, and methionine; and polar amino acids include glycine, serine, threonine, cysteine, tyrosine, asparagine, glutamine, arginine, lysine, histidine, aspartic acid and glutamic acid.

Some preferred fusion polypepitdes comprise Tat signal peptides that include, but are not limited to the Tat signal peptides of proteins SCO2286 - MTPANHQAPTSAPSPAPSQSSHAPELRAAARSLG**RR**RFLTVTGAAAALAFAVNLPAAGTA SAA (SEQ ID NO: 218) and SCO3790 long - MRKLLPLIGTPSGSHPGGRSAMTCRFRCGDACFHE**V**PNTSSNEYVGDVIAGALS**RR**SMM RAAAVVTVAAAGAGAVGVAGAPSAQAA (SEQ ID NO: 227).

Another Tat signal peptide comprises the Tat signal peptide of protein SCO6580 long - MTPFTDSSRTDA GTDPSADGPGESLRRALGVN**RR**RFLSTCTAVAAGAVAAPVFGASPALAH (SEQ ID NO: 241).

In addition, the fusion polypeptides of the invention comprise variants of the novel Tat signal peptides. In some embodiments, the fusion polypeptide comprises a variant of a Tat signal peptide that is encoded by an alternative start site that leads to the translation of a shorter Tat signal peptide. Variant Tat signal peptides include, but are not limited to the Tat signal peptide of protein SCO3790, designated SCO3790short (SEQ ID NO: 122), and the Tat signal peptide of protein SCO6580, designated SCO6580 short - (SEQ ID NO: 182).

The invention also contemplates fusion polypepitdes that comprise Tat signal peptides that include the motif $(X^{-1})RR(X^{+2})(X^{+3})$ (X^{+4}), wherein RR represents two adjacent

arginine residue and X designates positions restrict to other selected amino acids: X^{-1} is H, A, P, K, R, N, T, G, S, D, Q E or L; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is T, G or A and wherein the motif is within the first 35N' terminal residues of the amino acid sequence of the polypeptide. In some embodiments when X^{-1} is S then X^{+4} will be T; in other embodiments when X^{-1} is H then X^{+4} will be A, and in other embodiments when X^{-1} is L then X^{+4} will be G. The primary amino acid sequence of some of the Tat signal peptides dictates that the amino acid residues at positions X^{+3} and X^{+4} are non-polar amino acid residues. Thus, the invention comprises Tat signal sequences that include the motif $(X^{-1})RR(X^{+2})(X^{+3})$ (X^{+4}), wherein RR represents two adjacent arginine residue and X designates positions restrict to other selected amino acids: X^{-1} is H, A, P, K, R, N, T, G, S, D, Q E or L; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; and X^{+3} and X^{+4} are non-polar amino acid residues.

Preferred fusion polypepitdes comprise Tat signal peptides that include, but are not limited to the Tat signal sequence of protein SCO1590 - MGGVSRRAFTVAALSAFTLVPEASAA (SEQ ID NO: 211) and Tat signal sequence of protein SCO1824 - MTAPLSR HRRALAIPAGLAVAASLAFLP GTPAAATPAAEAA (SEQ ID NO: 213).

In some embodiments, the fusion polypeptide comprises a Tat signal peptide that is the secretory leader sequence of polypeptides that are naturally expressed by *Streptomyces* that is operably linked to a heterologous polypeptide or protein of interest. In some embodiments, the fusion polypeptide comprises a Tat signal peptide and a heterologous polypeptide such as an enzyme, a growth factor or a hormone. Enzymes include but are not limited to protease, a carbohydrase, such as amylases, cellulases, xylanases, and lipases; an isomerase, such as racemases, epimerases, tautomerases, or mutases; a transferase; a glucoamylase; a kinase, an amidase, an esterase, or an oxidase. Thus, the protein of interest may be an enzyme such as a carbohydrase, such as an α-amylase, an alkaline α-amylase, a β-amylase, a cellulase; a dextranase, an α-glucosidase, an α-galactosidase, a glucoamylase, a hemicellulase, a pentosanase, a xylanase, an invertase, a lactase, a naringanase, a pectinase or a pullulanase; a protease such as an acid protease, an alkali protease, bromelain, ficin, a neutral protease, papain, pepsin, a peptidase, rennet, rennin, chymosin, subtilisin, thermolysin, an aspartic proteinase, or trypsin; a lipase or esterase, such as a triglyceridase, a phospholipase, a pregastric esterase, a phosphatase, a phytase, an amidase, an iminoacylase, a glutaminase, a lysozyme, or a penicillin acylase; an isomerase such as glucose isomerase; an oxidoreductases, e.g., an amino acid oxidase, a catalase, a chloroperoxidase, a glucose

oxidase, a hydroxysteroid dehydrogenase or a peroxidase; a lyase such as a acetolactate decarboxylase, an aspartic β-decarboxylase, a fumarese or a histadase; a transferase such as cyclodextrin glycosyltranferase; or a ligase, for example. In particular embodiments, the protein may be an aminopeptidase, a carboxypeptidase, a chitinase, a cutinase, a deoxyribonuclease, an α-galactosidase, a β-galactosidase, a β-glucosidase, a laccase, a mannosidase, a mutanase, a pectinolytic enzyme, a polyphenoloxidase, ribonuclease or transglutaminase, for example. The enzyme may be a wild-type enzyme or a variant of a wild-type enzyme. The enzyme may also be a hybrid enzyme, which comprises at least two fragments from different enzymes, for example a catalytic domain of one enzyme and a starch binding domain of a different enzyme or two fragments each fragment comprising part of the catalytic domain of the enzymes. In some embodiments, the fusion polypeptide of the invention comprises a Tat signal peptide, as recited herein, and an enzyme that is a protease, a carbohydrase, an isomerase, a glucoamylase, a kinase, an amidase, an esterase, or an oxidase.

In other embodiments, the fusion polypeptide caomprises a Tat signal peptide and a heterologous poplypeptide that is not naturally associated with a secretion signal peptide.

In other embodiment, the fusion polypeptide comprises a heterologous polypeptide that may be a therapeutic protein (i.e., a protein having a therapeutic biological activity). Examples of suitable therapeutic proteins include: erythropoietin, cytokines such as interferon-α, interferon-β, interferon-γ, interferon-ω, and granulocyte-CSF, GM-CSF, coagulation factors such as factor VIII, factor IX, and human protein C, antithrombin III, thrombin, soluble IgE receptor α-chain, IgG, IgG fragments, IgG fusions, IgM, IgA, interleukins, urokinase, chymase, and urea trypsin inhibitor, IGF-binding protein, epidermal growth factor, growth hormone-releasing factor, annexin V fusion protein, angiostatin, vascular endothelial growth factor-2, myeloid progenitor inhibitory factor-1, osteoprotegerin, α-1-antitrypsin, α-feto proteins, DNase II, kringle 3 of human plasminogen, glucocerebrosidase, TNF binding protein 1, follicle stimulating hormone, cytotoxic T lymphocyte associated antigen 4-Ig, transmembrane activator and calcium modulator and cyclophilin ligand, soluble TNF receptor Fc fusion, glucagon like protein 1 and IL-2 receptor agonist. Antibody proteins, e.g., monoclonal antibodies that may be humanized, are of particular interest.

The invention encompasses the polynucleotides that encode the fusion polypeptides. Thus, in some embodiments, the invention is directed to isolated polynucleotides that comprise a polynucleotide sequence encoding a Tat signal peptide, as

recited above, that is operably linked to a second nucleotide sequence encoding a heterologous polypeptide.

The Tat polynucleotides that encode the Tat polypeptides of the invention, as recited above, include the sequence information of the nucleic acid sequences generated by PCR using the primer pairs having SEQ ID NO: 15 and 16, 18 and 19, 21 and 22, 24 and 25, 27 and 28, 30 and 31, 33 and 34, 36 and 37, 39 and 40, 42 and 43, 45 and 46, 48 and 49, 51 and 52, 54 and 55, 57 and 58, 60 and 61, 63 and 64, 66 and 67, 69 and 70, 72 and 73, 75 and 76, 78 and 79, 81 and 82, 84 and 85, 87 and 88, 90 and 91, 93 and 94, 96 and 97, 99 and 100, 102 and 103, 105 and 106, 108 and 109, 111 and 112, 114 and 115, 117 and 118, 120 and 121, 123 and 124, 126 and 127, 129 and 130, 132 and 133, 135 and 136, 138 and 139, 141 and 142, 144 and 145, 147 and 148, 150 and 151, 153 and 154, 156 abd 157, 159 and 160, 162 and 163, 165 and 166, 168 and 169, 171 and 172, 174 and 175, 177 and 178, 180 and 181, 183 and 184, 186 and 187, 189 and 190, 192 and 193, 195 and 196, 198 and 199, and 201 and 202. The primer pairs are given in Table 2. The invention also encompasses polynucleotides that are generated by degenerate variants of the primer pairs recited herein.

The polynucleotides of the present invention also include, but are not limited to a polynucleotide comprising the protein coding sequence of the Tat signal peptides having SEQ ID NO: 254 -312. The polynucleotides of the present invention also include polynucleotides that hybridize under stringent conditions to the complement of any of the polynucleotides of the invention , a polynucleotide encoding any one of the Tat signal peptides of SEQ ID NO: 17, 20, 23 26 29 32 35, 38, 41, 44, 47, 50,53, 56, 59, 62, 65, 68, 71, 74, 77, 80, 83, 86, 89, 92, 95, 98, 101, 104, 107, 110, 113 116, 119, 122, 125, 128 131, 134, 137, 140, 143, 146, 149, 152, 155, 158, 161, 164, 167, 170, 173, 176, 179, 182, 185, 188, 191, 194, 197, 200, 203, and 204-253, a polynucleotide that is an allelic variant of any polynucleotide recited above, a polynucleotide which is a species homolog of any of the polypeptides recited above, or a polynucleotide that encodes a polypeptide comprising any one of the motifs described above. An allelic variant denotes any of two or more alternative forms of a gene occupying the same chromosomal locus. Allelic variation arises naturally through mutation, and may result in polymorphism within populations. Gene mutations can be silent (no change in the encoded polypeptide) or may encode polypeptides having altered amino acid sequences. An allelic variant of a polypeptide is a polypeptide encoded by an allelic variant of a gene.

The invention also encompasses polynucleotide fragments of the nucleic acid sequences of the invention. In addition to the use of these fragments for PCR, the

polynucleotide fragments can be used in various hybridization procedures or microarray procedures to identify or amplify identical or related parts of mRNA or DNA molecules from which the Tat signal peptides of the invention can be derived.

The polynucleotides of the invention additionally include the complement of any of the polynucleotides recited above.

The polynucleotides of the invention also provide polynucleotides including nucleotide sequences that are substantially equivalent to the Tat polynucleotides recited above. Polynucleotides according to the invention can have, e.g., at least about 65%, at least about 70%, at least about 75%, at least about 80%, more typically at least about 90%, and even more typically at least about 95%, sequence identity to a polynucleotide recited above. The invention also provides the complement of such polynucleotides. The polynucleotide can be DNA (genomic, cDNA, amplified, or synthetic) or RNA. Methods and algorithms for obtaining such polynucleotides are well known to those of skill in the art and can include, for example, methods for determining hybridization conditions which can routinely isolate polynucleotides of the desired sequence identities.

A Tat fusion polypeptide of the invention can be produced by standard recombinant DNA techniques. For example, DNA fragments coding for the different polypeptide sequences are ligated together in-frame in accordance with conventional techniques, e.g., by employing blunt-ended or stagger-ended termini for ligation, restriction enzyme digestion to provide for appropriate termini, filling-in of cohesive ends as appropriate, alkaline phosphatase treatment to avoid undesirable joining, and enzymatic ligation. In another embodiment, the fusion gene can be synthesized by conventional techniques including automated DNA synthesizers. Alternatively, PCR amplification of gene fragments can be carried out using anchor primers that give rise to complementary overhangs between two consecutive gene fragments that can subsequently be annealed and reamplified to generate a chimeric gene sequence (see, e.g., Ausubel, et al. (eds.) CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, John Wiley & Sons, 1992). A Tat signal peptide-encoding nucleic acid can be cloned into such an expression vector such that the fusion moiety e.g. polypeptide of interest, is linked in-frame to the Tat signal peptide. An expression vector comprising a Tat signal peptide –encoding polynucleotide can be any vector capable of expressing the polynucleotide encoding Tat signal peptide or Tat-fusion polypeptide in a selected host organism, and the choice of vector will depend on the host cell into which the expression vector is introduced. Thus, the invention encompasses expression vectors that comprise a first nucleotide sequence encoding a Tat

signal peptide, as recited herein, operably linked to a second nucleotide sequence encoding a heterologous polypeptide.

The expression vector typically includes the components of a cloning vector, such as, for example, an element that permits autonomous replication of the vector in the selected host organism and one or more phenotypically detectable markers for selection purposes. The expression vector normally comprises control nucleotide sequences encoding a promoter, operator, ribosome binding site, translation initiation signal and optionally, a repressor gene or one or more activator genes. For expression under the direction of control sequences, the nucleic acid sequence the modified enzyme is operably linked to the control sequences in proper manner with respect to expression.

Preferably, a polynucleotide in a vector is operably linked to a control sequence that is capable of providing for the expression of the coding sequence by the host cell, i.e. the vector is an expression vector. The control sequences may be modified, for example, by the addition of further transcriptional regulatory elements to make the level of transcription directed by the control sequences more responsive to transcriptional modulators. The control sequences may in particular comprise promoters.

In the vector, the nucleic acid sequence encoding for the Tat signal peptide or the Tat signal peptide fusion polypepitde is operably combined with a suitable promoter sequence. The promoter can be any DNA sequence having transcription activity in the host organism of choice and can be derived from genes that are homologous or heterologous to the host organism. Examples of suitable promoters for directing the transcription of the modified nucleotide sequence, such as modified enzyme nucleic acids, in a bacterial host include the promoter of the Streptomyces coelicolor agarase gene dagA promoters, the promoters of the Bacillus licheniformis .alpha.-amylase gene (amyL), the aprE promoter of Bacillus subtilis, the promoters of the Bacillus stearothermophilus maltogenic amylase gene (amyM, the promoters of the Bacillus amyloliquefaciens .alpha.-amylase gene (amyQ), the promoters of the Bacillus subtilis xylA and xylB genes and a promoter derived from a Lactococcus sp.-derived promoter including the P170 promoter.

The Tat fusion polypeptide may, in addition, can comprise a tag sequence that is fused to the C-terminus of the Tat fusion polypeptide to generate a tagged Tat fusion polypeptide. Such tag sequences can be used to identify transformants and/or to facilitate the purification of recombinant Tat fusion polypeptides. For example, the Tat fusion polypeptide it may be expressed to contain a tag such as those of maltose binding protein (MBP), glutathione-S-transferase (GST) or thioredoxin (TRX), or as a His tag. Kits for expression and purification of such fusion proteins are commercially available from New

England BioLab (Beverly, Mass.), Pharmacia (Piscataway, N.J.) and Invitrogen, respectively. The Tat fusion polypeptide can also be tagged with an epitope and subsequently purified by using a specific antibody directed to such epitope. One such epitope ("FLAG.RTM.) is commercially available from Kodak (New Haven, Conn.). Another tag that can be used in the invention is the c-myc tag, as is described in the examples.

This invention further provides expression vectors comprising at least a fragment of the polynucleotides set forth above and host cells or organisms transformed with these expression vectors. Useful vectors include plasmids, cosmids, lambda phage derivatives, phagemids, and the like, that are well known in the art. Accordingly, the invention also provides a vector including a polynucleotide of the invention and a host cell containing the polynucleotide. In general, the vector contains an origin of replication functional in at least one organism, convenient restriction endonuclease sites, and a selectable marker for the host cell.

The present invention further includes novel expression vectors comprising promoter elements operatively linked to polynucleotide sequences encoding a Tat fusion protein/polypeptide that comprises a Tat signal peptide of the invention and a protein of interest. The invention encompasses plasmids pTDW92, pTDW73, pTDW121, pTDW102, pTDW119, pTDW74, pTDW75, pTDW48, pTDW76, pTDW103, pTDW118, pTDW77, pTDW49, pTDW51, pTDW52, pTDW78, pTDW91, pTDW79, pTDW104, pTDW90, pTDW89, pTDW88, pTDW80, pTDW53, pTDW106, pTDW81, pTDW82, pTDW83, pTDW84, pTDW61, pTDW56, pTDW62, pTDW47, pTDW63, pTDW93, pTDW107, pTDW60, pTDW64, pTDW72, pTDW87, pTDW86, pTDW108, pTDW50, pTDW109, pTDW67, pTDW57, pTDW65, pTDW66, pTDW54, pTDW58, pTDW85, pTDW116, pTDW68, pTDW110, pTDW111, pTDW113, pTDW94, pTDW69, pTDW70, pTDW114, pTDW71, pTDW95, and pTDW115. The construction of the various expression vectors is disclosed in the examples.

Expression Systems

In some preferred embodiments, the expression host strain for heterologously expressed protein will be a *Streptomyces* strain. As used herein, the genus *Streptomyces* includes all members known to those skilled in the art, including but not limited to S. coelicolor, S. lividans, In some embodiments, the expression host strain is S. coelicolor. In other embodiments, the expression host strain is *S. lividans.* The genetic elements and tools required for heterologous expression of proteins is known in the art for *Streptomyces*, including expression vectors, promoters, as well as fermentation protocols (Glibert et al.,

(1995) Crit. Rev. Biotechnology 15:13 – 39). Other examples of suitable bacterial host organisms are gram positive bacterial species such as *Bacillaceae* including *Bacillus subtilis, Bacillus licheniformis, Bacillus lentus, Bacillus brevis, Bacillus stearothermophilus, Bacillus alkalophilus, Bacillus amyloliquefaciens, Bacillus coagulans, Bacillus lautus, Bacillus megaterium* and *Bacillus thuringiensis,* and other *Streptomyces* species such as *Streptomyces murinus, S. rubiginosus, S. griseus, S. avermitilis,* lactic acid bacterial species including *Lactococcus spp.* such as *Lactococcus lactis, Lactobacillus spp.* including *Lactobacillus reuteri, Leuconostoc spp., Pediococcus spp.* and *Streptococcus spp.*

In general a DNA sequence encoding a Tat signal sequence according to the invention is linked to the 5' end of the polynucleotide encoding the protein of interest, such that the signal sequence directs the secretion of the polypeptide sequence via the Tat pathway. Reference is made to Hopwood et al., (1985) Genetic Manipulation of *Streptomyces*: Laboratory Manual The John Innes Foundation, Norwich, UK, Fernandez-Abalos et al., (2003) *Microbiol.* 149:1623 – 1632; Connell, N.D. (2001) Curr. Opin. Biotechnol. 5:446 – 449; Binnie et al., (1997) Trends Biotechnol, 8:315 – 320; Van Mellaert (1993) FEMS Microbiol. Lett 114:121 – 128; and Baltz R., Chapter 6 "Gene Expression in Recombinant *Streptomyces*" in Gene Expression in Recombinant Microorganisms, Ed. Alan Smith, 1995, NY, M. Dekker).

Cell Cultures

The present invention further provides host cells genetically transformed with the vectors of the invention, which may be, for example, an expression vector that contains the Tat signal polynucleotides and/or the Tat fusion polynucleotides of the invention. Thus, the invention encompasses a bacterial host cell transformed with an expression vector, as recited above. The vector may be, for example, in the form of a plasmid, a viral particle, a phage etc. The host cells and transformed cells can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying Tat signal and/or Tat fusion polynucleotides. The specific culture conditions, such as temperature, pH and the like will be apparent to those skilled in the art. In addition, preferred culture conditions may be found in the scientific literature such as Hopwood (2000) Practical *Streptomyces* Genetics, John Innes Foundation, Norwich UK; Hardwood et al., (1990) Molecular Biological Methods for *Bacillus,* John Wiley and from the American Type Culture Collection (ATCC). The present invention still further provides host cells genetically engineered to express the polynucleotides of the invention, wherein such

polynucleotides are in operative association with a regulatory sequence heterologous to the host cell which drives expression of the polynucleotides in the cell.

The host cell can be a higher eukaryotic host cell, such as a mammalian cell, or a lower eukaryotic host cell, such as a yeast cell. Preferably, the host cell is a prokaryotic cell, such as a bacterial cell. The bacterial host cells may be from gram positive or gram negative bacteria. Preferably, the invention encompasses host cells from gram positive bacteria. A number of types of gram positive cells that may act as suitable host cells for expression of the Tat signal peptides and/or Tat fusion polypeptides include, for example, the *Streptomyces, Bacillus* and *Lactococcus* species recited herein. For industrial purposes the bacteria generally used are Bacillus subtilis, S. lividans. Preferably, the polypeptides of the invention are expressed in *S. lividans* cells. If the protein is made in bacteria, it may be necessary to modify the protein produced therein, for example by phosphorylation or glycosylation of the appropriate sites, in order to obtain a biologically active protein. Such covalent attachments may be accomplished using known chemical or enzymatic methods.

Identification of Transformants

Although the presence/absence of marker gene expression suggests that a gene of interest is also present, its presence and expression should be confirmed. For example, if the nucleic acid encoding heterologous protein, such as agarase is inserted within a marker gene sequence, recombinant cells containing the insert can be identified by the absence of marker gene function. Alternatively, a marker gene can be placed in tandem with nucleic acid encoding the secretion factor under the control of a single promoter. Expression of the marker gene in response to induction or selection usually indicates expression of the secretion factor as well.

Alternatively, host cells which contain the coding sequence for a secretion factor and express the protein may be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridization and protein bioassay or immunoassay techniques which include membrane-based, solution-based, or chip-based technologies for the detection and/or quantification of the nucleic acid or protein.

Secretion Assays

Means for determining the levels of secretion of a heterologous or homologous protein in a gram-positive host cell and detecting secreted proteins include, using either

polyclonal or monoclonal antibodies specific for the protein. Examples include enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA) and fluorescent activated cell sorting (FACS). These and other assays are described, among other places, in Hampton R et al (1990, Serological Methods, a Laboratory Manual, APS Press, St Paul MN) and Maddox DE et al (1983, J Exp Med 158:1211).

A wide variety of labels and conjugation techniques are known by those skilled in the art and can be used in various nucleic and amino acid assays. Means for producing labeled hybridization or PCR probes for detecting specific polynucleotide sequences include oligo labeling, nick translation, end-labeling or PCR amplification using a labeled nucleotide. Alternatively, the nucleotide sequence, or any portion of it, may be cloned into a vector for the production of an mRNA probe. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes in vitro by addition of an appropriate RNA polymerase such as T7, T3 or SP6 and labeled nucleotides.

A number of companies such as Pharmacia Biotech (Piscataway NJ), Promega (Madison WI), and US Biochemical Corp (Cleveland OH) supply commercial kits and protocols for these procedures. Suitable reporter molecules or labels include those radionuclides, enzymes, fluorescent, chemiluminescent, or chromogenic agents as well as substrates, cofactors, inhibitors, magnetic particles and the like. Patents teaching the use of such labels include US Patents 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149 and 4,366,241. Also, recombinant immunoglobulins may be produced as shown in US Patent No. 4,816,567 and incorporated herein by reference. An enzyme reporter assay for identifying a secreted polypeptide is described herein. While the present invention describes an agarase reporter assay for determining the presence of secreted polypeptides, it is understood that any assay that uses any one reporter described above can be used to identify the secreted polypetides. The enzymatic activity of a secreted TAT fusion polypeptide can be determined by contacting the secreted polypeptide with a substrate and detecting the reaction product produced by the action of the enzyme on the substrate. In addition, verifying that the correct enzyme or other reporter molecule is secreted can be accomplished by performing mass spectroscopy of the secreted protein.

Purification of Proteins

Host cells transformed with polynucleotide sequences encoding heterologous or homologous protein may be cultured under conditions suitable for the expression and recovery of the encoded protein from cell culture. "Recovering" a polypeptide from a culture medium refers to collecting the polypeptide in the culture medium into which it was

secreted by the host cell. The polypeptide can also be recovered from a lysate prepared from the host cells and further purified. A secreted polypeptide may be recovered from the cell wall fraction prepared according to methods known in the art. One skilled in the art can readily follow known methods for isolating polypeptides and proteins in order to obtain one of the isolated polypeptides or proteins of the present invention. These include, but are not limited to, immunochromatography, HPLC, size-exclusion chromatography, ion-exchange chromatography, and immuno-affinity chromatography. See, e.g., Scopes, Protein Purification: Principles and Practice, Springer-Verlag (1994); Sambrook, et al., in Molecular Cloning: A Laboratory Manual; Ausubel et al., Current Protocols in Molecular Biology. The protein produced by a recombinant host cell comprising a secretion factor of the present invention will be secreted into the culture media.

The invention also encompasses fusion polypeptides that comprise a Tat signal peptide and a heterologous peptide and a polypeptide domain that will facilitate purification of soluble proteins (Kroll DJ et al (1993) DNA Cell Biol 12:441-53). Such purification facilitating domains include, but are not limited to, metal chelating peptides such as histidine-tryptophan modules that allow purification on immobilized metals (Porath J (1992) Protein Expr Purif 3:263-281), protein A domains that allow purification on immobilized immunoglobulin, and the domain utilized in the FLAGS extension/affinity purification system (Immunex Corp, Seattle WA). The inclusion of a cleavable linker sequence such as Factor XA or enterokinase (Invitrogen, San Diego CA) between the purification domain and the heterologous protein can be used to facilitate purification.

In yet another aspect, the invention is directed to a method for producing a heterologous polypeptide comprising culturing host cells in culture medium under conditions suitable for producing the heterologous polypeptide, wherein the host cells contain an expression vector that comprises a first nucleotide sequence encoding a TAT signal peptide encompassed by the invention operatively linked to a second nucleotide sequence encoding a heterologous polypeptide, and producing the heterologous polypeptide. In some embodiments, the method uses a TAT signal peptide of the invention that comprises the sequence motif $(X^{-1})RR(X^{+2})(X^{+3})(X^{+4})$, wherein R is arginine, X^{-1} is amino acid M, H, A, P, K, R, N, T, G, S, D, Q or E; X^{+2} is amino acid A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is Q, I, L, V, M or F, and wherein the motif is not within the first 35N' terminal residues of the amino acid sequence of the polypeptide.

In other embodiments, the heterologous polypeptide that is produced by the method of the invention includes a TAT signal peptide that comprises the sequence motif

$(X^{-1})RR(X^{+2})(X^{+3})(X^{+4})$, wherein R is arginine, X^{-1} is amino acid H, A, P, K, R, N, T, G, S, D, Q E or L; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is T, G or A, and wherein the motif is within the first 35N' terminal residues of the amino acid sequence of the polypeptide. In some embodiments of this aspect, the step of producing the heterologous polypeptide comprises recovering the polypeptide from the culture medium.

In other embodiments of this aspect, the host cell is a prokaryotic cell. In other embodiments, the host cell is a *Streptomyces* bacterial cell. In yet other embodiments, the host cell is a *S. coelicolor* or an *S. lividans* bacterial cell. In further embodiments of this aspect, the method of the invention produces a heterologous polypeptide that can be an enzyme a growth factor or a hormone.

It is contemplated that any signal sequence that directs a nascent polypeptide into a secretory pathway may be used in the present invention. It is to be understood that the any one of the newly identified signal sequences are encompassed by the invention. The signal peptides that are specifically contemplated by the instant invention included the Tat-dependent signal peptides. Specific examples include but are not limited to the Tat signal peptides listed in Tables 2, 3, and 6.

EXPERIMENTAL

The following preparations and examples are given to enable those skilled in the art to more clearly understand and practice the present invention. They should not be considered as limiting the scope and/or spirit of the invention, but merely as being illustrative and representative thereof.

In the experimental disclosure which follows, the following abbreviations apply: eq (equivalents); M (Molar); µM (micromolar); N (Normal); mol (moles); mmol (millimoles); µmol (micromoles); nmol (nanomoles); g (grams); mg (milligrams); kg (kilograms); µg (micrograms); L (liters); ml (milliliters); µl (microliters); cm (centimeters); mm (millimeters); µm (micrometers); nm (nanometers); °C. (degrees Centigrade); h (hours); min (minutes); sec (seconds); msec (milliseconds); TLC (thin layer achromatography); TY, trypton/yeast extract; Ap, ampicillin; DTT, dithiotreitol; Em, erythromycin; HPDM, high phosphate defined medium; IPG, immobilized pH gradient; IPTG, isopropyl-β-D-thiogalactopyranoside; Km, kanamycin; LPDM, low phosphate defined medium; MM, minimal medium; OD, optical density; PAGE, polyacrylamide gel electrophoresis; PCR, polymerase chain reaction; Sp, spectinomycin; SSM, Schaeffer's sporulation medium; 2D, two-dimensional.

Example 1

Strains and Growth Conditions

Derivatives of *S. coelicolor* strain M145 (Bentley et al., (2002) Nature 417;141-147; Kieser T. et al., (2000) Practical Streptomyces Genetics, The John Innes Foundation, Norfolk, UK) and *S. lividans* 1326 were used throughout the examples. Strains were cultured on standard laboratory growth media: minimal medium (Kieser et al., (2000) Practical Streptomyces Genetics – The John Innes Found, Norfolk, UK), complete medium (CM) Hopwood DA (1967) Bacteriol Rev 31:373-403), Mannitol soya flour (MS) (Hobbs et al., (1989) Appl Microbiol Biotechnol 31:272-277) or R5 Bierman et al (1992) Gene 116:43-49) for growth on plates and Tryptone soya broth (TSB) and Yeast extract-malt extract (YEME) for liquid growth (Hopwood DA, *supra*). Antibiotics were used at a final concentration of 50 µg/ml unless otherwise stated.

Example 2

Construction of TAT mutants and gross phenotypic analysis of Δ*tatC* mutant *S. coelicor* strains

To determine the impact of Tat-mediated protein export in *S. coelicolor*, marked mutations in each of the *tatA, tatB* and *tatC* genes and umarked in-frame deletions of *tatB* and *tatC* were constructed.

S. coelicolor tat deletion strains were prepared as follows. Antibiotic-marked and unmarked in-frame deletions were constructed using the method of Datsenko and Wanner (Datsenko et al., (2000) Proc. Natl. Acad. Sci. USA 97:6640 - 6645) modified as described by Gust et al., (Gust et al., (2003) Proc. Natl. Acad. Sci. USA 100:1541 – 1546). Cosmid 141, which carries the tatA and tatC genes, and cosmid P8, which carries the tatB gene were mutated with apramycin replacement cassettes prepared for apramycin replacement cassettes prepared for *tatA*, *tatB*, and *tatC* genes by using primers listed in Table 1. The mutated cosmids were transferred by mating into *S. coelicolor* M145, and single-crossover recombinants were selected for on MS medium containing apramycin and kanamycin (Berks et al., (2003) Adv Microbiol Physiol 47:187-254). Double crossover recombinants subsequently were selected by several rounds of growth on nonselective media followed by selection for colonies that were apramycin-resistant and kanamycin sensitive. Loss of the *tat* genes in each strain was confirmed by PCR, and the strains were designated TP1 (Δ*tatC*::ApraR), TP2 (Δ*tatA*::ApraR), and TP3 (Δ*tatB*::ApraR). Markerless strains were

constructed by the method of Gust *et al.* (2), loss of the apramycin-resistance cassette was confirmed by PCR, and the strains were designated TP4 (Δ*tatC*) and TP5 (Δ*tatB*).

The apramycin resistance cassette from pIJ773 (Gust et al., (2003) Proc. Natl. Acad. Sci. USA 100:1541 – 1546) was amplified using the tatA, tatB or tatC primers, which resulted in a PCR fragments where the final 39bp at each end corresponded to the flanking regions of the target genes. The PCR products were transformed by electroporation into strain BW25113/pIJ790 harboring either cosmid 141, which carries the *tatA* and *tatC* genes, orP8, which carries the *tatB* gene, and transformants were selected for apramycin resistance. The mutated cosmids were extracted, transformed into *E. coli* strain ET12567 carrying helper plasmid pUZ8002 and mated into *S. coelicolor* M145. Single crossovers were selected by several rounds of growth on non-selective media followed by selection for colonies that were apramycin resistant and kanamycin sensitive (Gust et al. (2003) *supra*). The loss of the tat genes in each strain was confirmed by PCR and the strains were designated TP1 (Δ*tatC*::Apra^R), TP2 (Δ*tatA*::Apra^R) and TP3 (Δ*tatB*::Apra^R).

Marker-less strains were constructed by transforming either cosmid I41 harboring the Δ*tatC*::Apra^R allele or cosmid P8 harboring the Δ*tatB*::Apra^R allele into *E. coli* strain DH5α carrying plasmid pBT340, inducing expression of the FLP recombinase by growth at 42°C and subsequent testing for colonies that were apramycin sensitive. The cosmids carrying the unmarked alleles were then transformed into E. coli strain ET12567/pUZ8002 and mated into the appropriate S. coelicolor ApraR-marked tat deletion strain with selection for kanamycin resistance. Kanamycin resistant colonies were grown for several rounds without selection and colonies that were sensitive to both apramycin and kanamycin were identified. The loss of the apramycin resistance cassette was confirmed by PCR and the strains were designated TP4 (Δ*tatC*) and TP5 (Δ*tatB*).

The gross phenotype of all of the *coelicolor tat* mutants were similar to those observed for the *tatB* and *tatC* mutants of *S. lividans* (Schaerlaekens et al., (2001) J Bacteriol 183:6727-6732; (Schaerlaekens et al.,(2004) Microbiology 150:21-31; and Figure 1).

When cultured on solid MS, the *S. coelicolor tat* mutant strains formed very small colonies that appeared to hypersporulate (Figure 1 A). Conversely, on the hyperosmotic solid R5 medium (Figure 1B), *tat* mutants failed to sporulate and produced very little of the blue-pigmented antibiotic actinorhodin. In TSB liquid medium, the *tat* mutants grew in a very dispersed manner compared with the M145, which normally grows as pellets. Consistent with the pronounced developmental defect noted on R5 agar, *tat* mutant strains failed to grow in liquid YEME medium (which contains 34% sucrose) unless sucrose was

excluded. *S. coelicolor tat* mutant mycelium were markedly more fragile than M145 mycelium and were particularly prone to lysis on shaking.

Therefore, to study the extracellular proteins of these fragile strains, it was necessary to grow the cells on solid media only and focus analysis on the exported proteins associated with the cell wall.

TABLE 1

Primers pairs used for generating tat mutants and agarase constructs

Primers for deletion of *tatA*	
TATAF	acc cac cca gcc gcc tcg gtg aga agg taa aga ctt atg att ccg ggg atc cgt cga cc (SEQ ID NO: 1)
TATAR	atc tcg tgc ggc ggg ccg gag atc acc ggc cct gcg tca tgt agg ctg gag ctg ctt c (SEQ ID NO: 2)
Primers for deletion of *tatB*	
TATBF	ccg tgg tgg cgg ccc gga cat ccc aag gag ctt cag gtg att ccg ggg atc cgt cga cc (SEQ ID NO: 3)
TATBR	ttc ccg gcc cgc ggg ccg ggc gtg gga gga gcg ggg tca tgt agg ctg gag ctg ctt c (SEQ ID NO: 4)
Primers for deletion of *tatC*	
TATCF	cct gcc cgc aac aag gag aag gac ccc gag ggg cgg atg att ccg ggg atc cgt cga cc (SEQ ID NO: 5)
TATCR	ggg cgg cga cgc gtg ccg tca ccg ccc cgg cgg tcc tca tgt agg ctg gag ctg ctt c (SEQ ID NO: 6)
Primers for construction of agarase reporter	
Agarase-F	ggc ggc aag ctt aga tct acc gat tgt cac cct gcg (SEQ ID NO: 7)
AgaraseLead	ggc cgg gaa ttc cat atg cgt tct cct tct tcg att cc (SEQ ID NO:8)
Agaraseleadb	ggc ggc gga tcc gca gac ctc gaa tgg gaa c (SEQ ID NO: 9)
AgaraseRxba	ggc ggc tct aga tct gtt ccg tga ggt gc (SEQ ID NO:10)
AadAFnde	ggc ggc cat atg cta ttt gcc gac tac ctt gg (SEQ ID NO:11)
AadARbam	ggc ggc gga tcc ctg acg ccg ttg gat aca cc (SEQ ID NO:12)
Primers for construction of myc-tagged agarase	
kddagaf	gga att cca tat ggt caa ccg acg tga tct (SEQ ID NO: 13)
kddagamyc	gag gat ccc tac aga tcc tcc tcc gag atg agc ttc tgc tcc acg gcc tga tac gtc ctg a (SEQ ID NO:14)

Example 3

Construction of reporter plasmids

The primer sequences used to assemble a reporter construct based on the promoter and structural gene for agarase, *dagA* are listed in table 1. Initially a PCR

product (amplified with primers Agarase-F and AgaraseL eadnde covering the dagA promoter was digested with HindIII and EcoRI and cloned into similarly digested pBluescriptII (Stratagene). A second PCR product, corresponding to the AadA streptomycin resistance gene of pIJ778 plus its promoter (amplified using primers AadAFnde and AadARbam with pIJ778 as the template), was digested with NdeI and BamHI and cloned into the above construct that had been pre-digested with the same enzymes. A third PCR fragment (amplified with primers AgaraseLEADBAM and AgaraseRxba) covering the region encoding the DagA mature protein sequence was digested with BamHI and XbaI and cloned into this to give plasmid pTDW45. Plasmid pTDW45 carries a fragment that corresponds to the *dagA* promoter and region of *dagA* encoding the mature protein sequence separated by aadA gene and flanked by BglII sites. The BglII fragment from pTDW45 was cloned into pSET152 (Bierman et al., (1992) Gene 116:43 - 49) that has been previously digested with BamHI to give pTDW46. This shuttle vector can replicate in *E. coli* and be mated into *S. coelicolor* where it integrates into the chromosome site-specifically. The aadA gene can be removed from this plasmid by digestion with NdeI and BamHI and replaced with DNA fragments encoding different signal peptides. A list of oligonucleotide primers and the signal peptides they amplify and the subsequent derivatives of pTDW46 carrying these signal peptides are listed in Table 2.

Plasmid pIJ6902dagA carries a C-terminally myc eptitope-tagged derivative of agarase under control of the thiostrepton-inducible promoter PtipA. The Agarase gene was amplified using primers kddagaf and kddagamyc (Table 1) digested with NdeI and BamHI and cloned into similarly digested pIJ6902 (Huang et al., (2005) Mol Microbiol 58:1276-1287). The pIJ6902dagA (PtipA-dagAmyc allele) was transferred to the ΦC31 site on the chromosome of M145 and TP4 as described in Bierman et al., *supra.*

TABLE 2

Primer pairs for amplifying Tat signal peptides

Gene*	Oligonucleotides	Signal peptide**	Plasmid***
SCO0432	F=ggc gcg cat atg aga cgg atc cgc ctg ttg (SEQ ID NO: 15) R=ggc gcg gga tcc ggt gct cgc gag tgc ggg tac (SEQ ID NO: 16)	MRRIRLLAAISAGLALAAGA↓VAPVPALAST (SEQ ID NO: 17)	pTDW92

SCO0474	F=ggc gcg cat atg gaa ccg acc gtc cgc ac (SEQ ID NO: 18) R=ggc gcg gga tcc ctc ggt cgc cgc ggg ttc (SEQ ID NO:19)	MEPTVRTHPTRRLPTALVLT AALLATGCSEQS↓DDGREPA ATE (SEQ ID NO: 20)	pTDW73
SCO0494	F=ggc cgc cat atg ctc ctc aga aca acg (SEQ ID NO: 21) R=ggc gcg gga tcc cac ggt gac cgg gaa ggc (SEQ ID NO: 22)	MLLRTTRTKPWRRLAAALSA AALGVGLLAGCGSDS↓DDP ADEAGGGTPAAAGAFPVTV (SEQ ID NO: 23)	pTDW121
SCO0736	F=ggc cgc cat atg ggg gac ata cgc aga c (SEQ ID NO: 24) R=ggc cgc gga tcc gca gct cgc ggc ctg ggc (SEQ ID NO: 25)	MGDIRRRGAVALGVTALVAP LTLALTA↓APAQAASC (SEQ ID NO: 26)	pTDW102
SCO0930	F=ggc gcg cat atg aag acc tcc tgg cgg agc (SEQ ID NO: 27) R=ggc gcg gga tcc gcc ggg ggc ggc cgt ggc (SEQ ID NO: 28)	MKTSWRSASLVAGAAALLAL TTACG↓QDGGGPTGSQNVG ATAAPG (SEQ ID NO: 29)	pTDW119
SCO1172	F=ggc gcg cat atg gaa agg gag agg aca gc (SEQ ID NO:30) R=ggc gcg gga tcc gta gtc gac ggc acg gac (SEQ ID NO:31)	MERERTAPVPTRRRLLKGA ALATVPYTLLSGTRAAA↓QV RAVDY (SEQ ID NO:32)	pTDW74
SCO1196	F=ggc gcg cat atg gcc ctc acc acc cga cg (SEQ ID NO:33) R=ggc gcg gga tcc ctg gga cgc cag cgc ggt acc (SEQ ID NO:34)	MALTTRRRALTTLGAALTGA VALPA↓GTALASQ (SEQ ID NO: 35)	pTDW75
SCO1230	F=ggc cgg cat atg agg aag agc agc ata cgg cgg agg gcg acc gcc ttc ggc acg gcc gga gca ctg gtc acc g (SEQ ID NO: 36) R=ggc cgg gga tcc ggc gct cgc ggc ggg tgc cga gac ggc gcc ggc gat cag cgt ggc ggt gac cag tgc tcc ggc c (SEQ ID NO: 37)	MRKSSIRRRATAFGTAGALV TATLIAGA↓VSAPAASA (SEQ ID NO: 38)	pTDW48

SCO1356	F=ggc gcg cat atg acc agc gca ccg ttt c (SEQ ID NO: 39) R=ggc gcg gga tcc gcc gcc gtc gcc gga cgc (SEQ ID NO: 40)	MTSAPFHPAAGPARRTVVA AAGAAGLTAVLAACS↓DSDD GASGDGG (SEQ ID NO: 41)	pTDW76
SCO1396	F=ggc gcg cat atg aca cga ctc tcc gcc (SEQ ID NO: 42) R=ggc gcg gga tcc ctc gtg ccg tgc ctg cgc (SEQ ID NO: 43)	MTRLSAALRGLATTFAALLA VTAVPAAS↓APATAQARHE (SEQ ID NO: 44)	pTDW103
SCO1432	F=ggc gcg cat atg caa cga cgg acc cgc ccg acc cgg aca (SEQ ID NO: 45) R=ggc gcg aga tct tgc ggc ccg ggc ggt cgc cgg gat c (SEQ ID NO: 46)	MQRRTRPTRTPGGLPLMSR VEQPSRRALLAAAVAAAAVT GGA↓IPATARAA (SEQ ID NO: 47)	pTDW118
SCO1565	F=ggc gcg cat atg gga acg cag gag tcg gac (SEQ ID NO: 48) R=ggc gcg gga tcc ggt gcc ggc ggc tct cgc cgt g (SEQ ID NO: 49)	MGTQESDERAGGGTGRRA LLGAAVLGAGGAVLGLPGTA RA↓AGT (SEQ ID NO: 50)	pTDW77
SCO1590 I	F=ggc ggc cat atg ggc ggt gtc tcg cgc cgt gcc ttc acg gtg gcg gcg ttg tcg gcg (SEQ ID NO: 51) R=ggc ggc gga tcc cgc cga cgc ctc ggg cac gag cgt gaa cgc cga caa cgc cgc cac cgt (SEQ ID NO: 52)	MGGVSRRAFTVAALSAFTLV PEA↓SAA (SEQ ID NO: 53)	pTDW49
SCO1590 II	F= ggc ggc cat atg ggc ggt gtc tcg cgc cgt gcc ttc acg gtg gcg gcg ttg tcg gcg (SEQ ID NO: 54) R=ggc ggc gga tcc ggg ggt cgc cgc cga cgc ctc ggg cac gag cgt gaa cgc cga caa cgc cgc cac cgt (SEQ ID NO: 55)	MGGVSRRAFTVAALSAFTLV PEASAA↓TP (SEQ ID NO: 56)	pTDW51

SCO1639	F=ggc gcg cat atg cgc cga cgc tca ctc ctc atc gcc gtc ccc acg gga ctg gtc acg ctc gcc gcc tgc ggt gac agc gac gac tcc gg (SEQ ID NO: 57) R=ggc gcg gga tcc cgc ggc cga ggt ggc cga cgc gtc ggg cga cgg gct ctc gct cgc gtt gct cga cga gcc gga gtc gtc gct gtc acc (SEQ ID NO: 58)	MRRRSLLIAVPTGLVTLAAC GDSDDSGSSSNA↓SESPSP DASATSAA (SEQ ID NO: 59)	pTDW52
SCO1824	F=ggc gcg cat atg acc gct ccc ctc tcg (SEQ ID NO: 60) R=ggc gcg gga tcc ggg cgc ggc ctc ggc cgc (SEQ ID NO: 61)	MTAPLSRHRRALAIPAGLAV AASLAFLPGTPAA↓ATPAAE AAP (SEQ ID NO: 62)	pTDW78
SCO1906	F=ggc gcg cat atg tcg ctc act cgc agg (SEQ ID NO: 63) R=ggc gcg gga tcc cgt gga cgc gag ggc gtt g (SEQ ID NO: 64)	MSLTRRDFAGRSALTGAGV VLAGSVGALA↓TAPNALAST (SEQ ID NO: 65)	pTDW91
SCO1948	F=ggc gcg cat atg aga cga aga gcg aga tc (SEQ ID NO: 66) R=ggc gcg gga tcc ggg ttg tgc ctg ggc gat g (SEQ ID NO: 67)	MRRRARSILAVGALLIGGAS FA↓PIAQAQP (SEQ ID NO: 68)	pTDW79
SCO1968	F=ggc gcg cat atg cac gtg cgc gca gta g (SEQ ID NO: 69) R=ggc gcg gga tcc cgc ctc gtc cgc ccg cgc gtg (SEQ ID NO: 70)	MHVRAVAVTTTALLGVALT↓ APLSHARADEA (SEQ ID NO: 71)	pTDW104
SCO2068	F=ggc gcg cat atg acc agt cga cac aga gc (SEQ ID NO: 72) R=ggc gcg gga tcc ggg ggc cgc gtc ggc cgc (SEQ ID NO: 73)	MTSRHRASENSRTPSRRTV VKAAAAGAVLAAPLAAA↓LP AGAADAAP (SEQ ID NO: 74)	pTDW90

SCO2226	F=ggc gcg cat atg cct gcc acg cgc cgt ac (SEQ ID NO: 75) R=ggc gcg gga tcc ggt gtc ggc ccg tgc ggc gag (SEQ ID NO: 76)	MPATRRTARVRRVAAVTVT ALAAA↓LLPPLAARADT (SEQ ID NO: 77)	pTDW89
SCO2286	F=ggc gcg cat atg aca ccc gcg aac cac (SEQ ID NO: 78) R=ggc gcg gga tcc ctc ggc ggc gct cgc ggt g (SEQ ID NO: 79)	MTPANHQAPTSAPSPAPSQ SSHAPELRAAARSLGRRRFL TVTGAAAALAFAVNLPA↓AG TASAAE (SEQ ID NO: 80)	pTDW88
SCO2383	F=ggc gcg cat atg cgt ctc aga cgc ccc cgc ag (SEQ ID NO: 81) R=ggc gcg gga tcc gct gtc ggc gac ggc act g (SEQ ID NO: 82)	MRLRRPRRPSDGPPGRSG ARRLRVASAPLIPLTVIGAVV AGLIGPEPPSDG↓PLPGTSA VADS (SEQ ID NO: 83)	pTDW80
SCO2446	F=ggc gcg cat atg agc aca gga cca gtc aga cgc ggg gcg gcc ctc ctg tcg gcg ggg ctc gtg (SEQ ID NO: 84) R=ggc gcg gga tcc ctg cgc cgc cga ggt ccc ggc cgg cag cag cgc gat cac gag ccc cgc cga cag gag (SEQ ID NO: 85)	MSTGPVRRGAALLSAGLVIA LLPA↓GTSAAQ (SEQ ID NO: 86)	pTDW53
SCO2591	F=ggc gcg cat atg ttc aca tca cgc gcg cgc ctc aga (SEQ ID NO: 87) R=ggc gcg gga tcc gtc ggc ggc gac ggc ggg ag (SEQ ID NO: 88)	MFTSRARLRTRRSRLLTCTA LAVAAGMLVTT↓PAVAAD (SEQ ID NO: 89)	pTDW106
SCO2637	F=ggc gcg cat atg acc aac agc ccc cag (SEQ ID NO: 90) R=ggc gcg gga tcc gtc ggc ggc gag ggc cac (SEQ ID NO: 91)	MTNSPQREPIPGARRAARL ALATGLAAALA↓AVGPVPVA LAAD (SEQ ID NO: 92)	pTDW81
SCO2758	F=ggc gcg cat atg cac cac agc agc acg g (SEQ ID NO: 93) R=ggc gcg gga tcc ggc gtc ggc gta ggc cgt ac (SEQ ID NO: 94)	MHHSSTAGTGSTAQPSRRS VLTATAVVTAALA↓AGGGTA YADA (SEQ ID NO: 95)	pTDW82

SCO2780	F=ggc gcg cat atg tcc cac gcc agc gct ac (SEQ ID NO: 96 R=ggc gcg gga tcc gtc gcc cgc ccc atc gct c (SEQ ID NO: 97)	MSHASATHPTRRGILAAGG ALGLGAVLAACG↓DGDGKS DGA GD (SEQ ID NO: 98)	pTDW83
SCO2786	F=ggc gcg cat atg aga cct cat cga cgg cac (SEQ ID NO: 99) R=ggc gcg gga tcc cgg ttc cgc ggc cgc ctt c (SEQ ID NO: 100)	MRPHRRHHRTTPRITRLLGS LLLVAAVGAMTTGA↓APVRK AAAEP (SEQ ID NO: 101)	pTDW84
SCO2821	F=ggc gcg cat atg aga cga cca gtc gcg ctg (SEQ ID NO: 102) R=ggc gcg gga tcc gcc ggc cgc tgc cgc cgc ctc (SEQ ID NO: 103)	MRRPVALRLQAALGTLALAA A↓AGVVLTMPEAAAAAG (SEQ ID NO: 104)	pTDW61
SCO3053	F= ggc gcg cat atg tcc aga ttc cgt tca cgc gcc gcg gtc acc gcg ggt ctg acg gcc atc gcg gtg gcc (SEQ ID NO: 105) R=ggc gcg gga tcc ggt ggc ggc ctg gac ggg tcc gac gac gag gca gcc ggc ggc cac cgc gat ggc cgt cag (SEQ ID NO: 106)	MSRFRSRAAVTAGLTAIAVA AGCLVVG↓PVQAAT (SEQ ID NO: 107)	pTDW56
SCO3456	F=ggc gcg cat atg cgg cac agc gga cgg (SEQ ID NO: 108) R=ggc gcg gga tcc ctc gac ccc cgt ggt atc (SEQ ID NO: 109)	MRHSGRRVKRATVLAAGAA VAGLLTAA↓CSSPDTTGVE (SEQ ID NO: 110)	pTDW62
SCO3471 (dagA)	F=ggc cgg cat atg gtc aac cga cgt gat ctc atc aag tgg agt gcc gtc gca ctc gga gcg ggt g (SEQ ID NO: 111) R=ggc cgg gga tcc gag gtc tgc ggc atg agc ggc ggg tgc ggg acc cgc gag ccc cgc acc cgc tcc gag tgc gac g (SEQ ID NO: 112)	MVNRRDLIKWSAVALGAGA GLAGPAPAAHA↓ADL (SEQ ID NO: 113)	pTDW47

SCO3484	F=ggc gcg cat atg acc aag cct gtt gtg (SEQ ID NO: 114) R=ggc gcg gga tcc acc gga gcc ttg gga cga g (SEQ ID NO: 115)	MTKPVVPSGVSRRGFLGGS LGVAGAVLLAA↓CSGGGNS SQGSG (SEQ ID NO: 116)	pTDW63
SCO3790 long	F=ggc gcg cat atg cgc aag ctc ctg ccg ttg (SEQ ID NO: 117) R=ggc gcg gga tcc ggg cgc ggc ctg cgc cga c (SEQ ID NO: 118)	MRKLLPLIGTPSGSHPGGRS AMTCRFRCGDACFHEVPNT SSNEYVGDVIAGALSRRSM MRAAAVVTVAAAGAGAVGV AG↓APSAQAAP (SEQ ID NO: 119)	pTDW93
SCO3790 short	F=ggc gcg cat atg ccg aac acc agc tcc aac g (SEQ ID NO: 120) R= ggc gcg gga tcc ggg cgc ggc ctg cgc cga c (SEQ ID NO: 121)	MPNTSSNEYVGDVIAGALSR RSMMRAAAVVTVAAAGAGA VGVAG↓APSAQAAP (SEQ ID NO: 122)	pTDW107
SCO4142	F=ggc gcg cat atg aac cgg cgg gcc ctc gct c (SEQ ID NO:123) R=ggc gcg gga tcc gct gtt ggc ggc ggc gga g (SEQ ID NO: 124)	MNRRALALGALAVSGALALT ACG↓SDDTGGNSGSDSSSA AANS (SEQ ID NO: 125)	pTDW60
SCO4152	F=ggc gcg cat atg cca gtc aca gca cag (SEQ ID NO: 126) R=ggc gcg gga tcc ccc gcc cgc gct cgc cga g (SEQ ID NO: 127)	MPVTAQPHQPRRRRRTSRL LVVAATVVTAGVLAAAALPA ↓SASAGG (SEQ ID NO: 128)	pTDW64
SCO4672	F=ggc gcg cat atg acc cac acc tca cgg (SEQ ID NO: 129) R=ggc gcg gga tcc cgg agc ggc gga cgc agc (SEQ ID NO: 130)	MTHTSRRGVLAAFAASAAT VPFGGAAAASPARATATAA SA↓VDPAAASAAP (SEQ ID NO: 131)	pTDW72
SCO4884	F=ggc gcg cat atg cgt cgg aca tca aga c (SEQ ID NO: 132) R=ggc gcg gga tcc gtc gtc gcc gct ctc gcc gct g	MRRTSRLIRVAVGVASLALA ATA↓CGGTSGESGDD (SEQ ID NO: 134)	pTDW87

	(SEQ ID NO: 133)		
SCO4885	F=ggc gcg cat atg cgc cgg att tcc cgg atc (SEQ ID NO: 135) R=ggc gcg gga tcc gga ctc cga acc ggc gtc (SEQ ID NO: 136)	MRRISRITVAGAATASLALALAACG↓GTSTDSGSES (SEQ ID NO: 137)	pTDW86
SCO5074	F=ggc cgc cat atg acc tcg tcg ctg cac (SEQ ID NO: 138) R=ggc cgc gga tcc gac gct cgc ggc gtg cgc (SEQ ID NO: 139)	MTSSLHHAIRLTTASAIALGGLVTLGTS↓AHAASV (SEQ ID NO: 140)	pTDW108
SCO5113	F=ggc ggc cat atg agc att ctc cgt aac cgc acc gcc acg gcc gcc ata gtc gcg gtc gca gcc ggc gcc ctg (SEQ ID NO: 141) R=ggc ggc gga tccggc gga gct gtc gct gtc gcc acc acc gca ggc ggt gag ggt cag ggc gcc ggc tgc gac cgc (SEQ ID NO:142)	MSILRNRTATAAIVAVAAGALTLTACGG↓GDSDSSA (SEQ ID NO: 143)	pTDW50
SCO5447	F=ggc cgc cat atg ttg agc agc agc act c (SEQ ID NO: 144) R=ggc cgc gga tcc gtc ggg ggc cgc ggt ggc (SEQ ID NO: 145)	MLSSSTPHRRTSHTTSHTSRRTAHRRAAAVALAGVAALIATAVQS↓GTATAAPD (SEQ ID NO: 146)	pTDW109
SCO5461	F=ggc gcg cat atg atc acc act agt ctg c (SEQ ID NO: 147) R=ggc gcg gga tcc cgg ggc ggc ctt ggc ggg (SEQ ID NO: 148)	MITTSLRRRTAAAVLSLSAVLATTAATAPGAAPAPSA↓APAKAAP (SEQ ID NO: 149)	pTDW67
SCO5660	F=ggc gcg cat atg cgt gcc cgt agc cat gca gtt ctc gcc gcc tcg gtc gtc ctc gcc ctg gcc g (SEQ ID NO: 150) R=ggc gcg gga tcc cgg cgc ggc gaa ggc ggg ggc cgc cag cgg gcc ggc ggc cag ggc gag gac gac c (SEQ ID NO: 151)	MRARSHAVLAASVVLALAAGPLA↓APAFAAP (SEQ ID NO: 152)	pTDW57

SCO6009	F=ggc gcg cat atg aac acg cgt atg cgt c (SEQ ID NO: 153) R=ggc gcg gga tcc ctc ctt ggc gct gcc aca g (SEQ ID NO: 154)	MNTRMRRAAVAVATTAMAV SLAA↓CGSAKE (SEQ ID NO: 155)	pTDW65
SCO6052	F=ggc gcg cat atg tcc acg tcc cgc aac g (SEQ ID NO: 156) R=ggc gcg gga tcc ggg ctg ggc ggc ggc ggt g (SEQ ID NO: 157)	MSTSRNAATRRQVLARTGA LGAGIAFTGALSELFA↓GTAA AQP (SEQ ID NO: 158)	pTDW66
SCO6198	F=ggc gcg cat atg gac acc act tac tgg agc aga cgt cgc gtc ctc acc gtg ctc ggc gcc gcc acc gcg gcg ac (SEQ ID NO: 159) R=ggc gcg gga tcc ggc ggc gag agc gcg gga cgg ggc ggc gag agg gat cga ggt cgc cgc ggt ggc ggc gcc (SEQ ID NO: 160)	MDTTYWSRRRVLTVLGAAT AATSIPLA↓APSRALAA (SEQ ID NO: 161)	pTDW54
SCO6199	F=ggc gcg cat atg cac cgc aca aga ccc ggg ggc cac ccc cga tcc ata cgc ttc gcc aca ctc gcg atc tcc ctg a (SEQ ID NO: 162) R=ggc gcg gga tcc ggt gac ggc gac ggc cgg tga cgc cgt cat gag cgc gga gcc cgc ggt cag gga gat cgc gag tgt g (SEQ ID NO: 163)	MHRTRPGGHPRSIRFATLAI SLTAGSALMT↓ASPAVAVT (SEQ ID NO: 164)	pTDW58
SCO6272	F=ggc gcg cat atg acc gag gtg tcc cgc (SEQ ID NO: 165) R=ggc gcg gga tcc ccc tgc cgc ggc cgc tgc (SEQ ID NO: 166)	MTEVSRRKLMKGAAVSGGA LALPALGAPPATA↓APAAGP EDLPGPAAAAAG (SEQ ID NO: 167)	pTDW85
SCO6281	F=ggc gcg cat atg tcc gcc att tct cgc (SEQ ID NO: 168) R=ggc gcg gga tcc ggc gat cgc cgc ggc ggg tg (SEQ ID NO: 169)	MSAISRRRFIRHGAVAGGAA VALPALGGWAGEAFA↓AQP SPAAAIA (SEQ ID NO: 170)	pTDW116
SCO6418	F=ggc gcg cat atg gac aga cga gcc ctc (SEQ ID NO: 171) R=ggc gcg gga tcc ggt cgc ggc gtc cgt ccc (SEQ ID NO: 172)	MDRRALMLATGGLLGAAGA AQLTAAPSAVA↓AGTDAAT (SEQ ID NO: 173)	pTDW68

SCO6457	F=ggc gcg cat atg ccg cac tcg ccc gtg tc (SEQ ID NO: 174) R=ggc gcg gga tcc ctc gtc ggc cgc cgc ccg ccg tac (SEQ ID NO: 175)	MPHSPVSPAESPAPQPGRP RPVVSRRRLLEGGAAVLGA LALSASPLTAQA↓AVRRAAA DE (SEQ ID NO: 176)	pTDW110
SCO6580 long	F=ggc cgc cat atg acc ccg ttc acc gac (SEQ ID NO: 177) R=ggc cgc gga tcc ccg gtc gtg ggc gag cgc (SEQ ID NO: 178)	MTPFTDSSRTDAGTDPSAD GPGESLRRALGVNRRRFLS TCTAVAAGAVAAPVFG↓ASP ALAHDR (SEQ ID NO: 179)	pTDW111
SCO6580 short	F=ggc gcg cat atg aac cgg cgc cgc ttc (SEQ ID NO: 180) R= as above; (SEQ ID NO: 181)	MNRRRFLSTCTAVAAGAVA APVFG↓ASPALAHDR (SEQ ID NO: 182)	pTDW113
SCO6594	F=ggc gcg cat atg ccg tct tcc ccc acg (SEQ ID NO: 183) R=ggc gcg gga tcc gga gtc ggc ggc ggc ggt tc (SEQ ID NO: 184)	MPSSPTPSPTPSGAPEPSG VRRRSLLAAAAAAPVLASLA GA↓GTAAADS (SEQ ID NO: 185)	pTDW94
SCO6644	F=ggc gcg cat atg cgc tcc cac cgc cgc (SEQ ID NO: 186) R=ggc gcg gga tcc acc gtc acc gga cgc gga a (SEQ ID NO: 187)	MRSHRRPRLLAPFLLVPLMA GCFASGGG↓DDSASGDG (SEQ ID NO: 188)	pTDW69
SCO6691	F=ggc gcg cat atg gcc gaa gtg aac cgg (SEQ ID NO: 189) R=ggc gcg gga tcc cgg gag tgc cgc ggc ccg gtc (SEQ ID NO: 190)	MAEVNRRRFLQLAGATTAF SALS↓ASIDRAAALP (SEQ ID NO: 191)	pTDW70
SCO6738	F=ggc cgc cat atg aac cgc ctg cgc acc (SEQ ID NO: 192) R=ggc cgc gga tcc gga gcc gtc ggc gtg ggc (SEQ ID NO: 193)	MNRLRTLTATTAALAASLLV PSLLTA↓PAHADGS (SEQ ID NO: 194)	pTDW114

SCO7399	F=ggc gcg cat atg aga cgc ctc ctg ctc (SEQ ID NO: 195) R=ggc gcg gga tcc ctt gtc ggc ggc ggg ttc (SEQ ID NO: 196)	MRRLLLTAAATTAAALTLAA CG↓TTEPAADK (SEQ ID NO: 197)	pTDW71
SCO7631	F=ggc gcg cat atg agc aca gag gaa cag (SEQ ID NO: 198) R=ggc gcg gga tcc gcc ggc cgc gtc ggc cgc ctg (SEQ ID NO: 199)	MSTEEQSSGRRHPDRRVLL RAAVAVPAAGAAVAGTAAV PA↓QAADAAG (SEQ ID NO: 200)	pTDW95
SCO7677	F=ggc cgc cat atg tcc aga cag atc gat c (SEQ ID NO: 201) R=ggc cgc gga tcc tcc ggg gtt tcc ggc gga g (SEQ ID NO: 202)	MSRQIDRRSFLRRGAAGAA ALAVGPGLLAA↓CSTDEPGS AGNPG (SEQ ID NO: 203)	pTDW115

*The SCO number for each protein whose signal peptide was cloned in frame with mature agarase is given.

** The signal peptide coding sequence that was amplified for each secreted protein is shown, the arrow indicates the most likely position of signal peptide cleavage as determined by SignalP

*** The plasmid designation of the final construct carrying the particular signal peptide-agarase fusion is given.

Example 4

Identification of Tat-dependent proteins by 2D-gel electrophoresis (2-DGE)

To identify proteins secreted via the TAT pathway, extracellular proteins derived from the *S. coelicolor* parent strain (M145) and the *tatC* mutant strain were analyzed by two-dimensional gel electrophoresis (2-DGE).

Protein were prepared for 2D gel analysis or MUDPIT by following the method of (Hesketh et al., (2002) Mol. Microbiol. 46:917 – 932; Yu et al., (2003) Anal. Chem. 75:6023 –6028; and Perkins et al., (1999) Electrophoresis 20:3551 - 3567) S. coelicolor strains M145 or TP1 *tatC* were grown by inoculating 10^6 spores onto sterile cellophane disks placed on the surface of complete medium (CM), R5 or mannitol soya (MS) media. After incubation for 48 hours at 30°C the biomass was scraped from the disks, dispersed in 5M LiCl solution and left on ice for 30 minutes. The suspensions were vortexed for 2 – 3 minutes and the biomass removed by centrifugation (1800 g for 5 min) followed by

passage through a 0.45 um filter. Protein were precipitated from the LiCl solution by addition of trichloroacetic acid to a final concentration of 20%, incubated on ice for 30 min and centrifuged at 1800g for 15 min. After centrifugation, the mixture had settled into two phases. The upper phase was removed and discarded and water was added to the lower phase to adjust it back to the original volume. The solution turned cloudy so it was again centrifuged at full speed for 15 min, after which the precipitated protein formed a pellet. The pellet was washed three times with -20°C acetone and then air-dried.

For 2D gel analysis, the protein samples were resuspended in IEF sample-buffer (8M urea, 0.5% CHAPS, 0.2% DTT, 0.5% IPG buffer pH 4 – 7 (Amersham Biosciences), 0.002% bromophenol blue) and protein concentration was determined using the Biorad DC protein assay. The proteins in the samples were separated in the first dimension by their differing isoelectric points. 500 – 1000 µg samples of protein were loaded onto Amersham Biosciences 18cm Immobline Drystrip pH 4 – 7 or pH 6 – 11 iso-electric focusing gels and resolved by electrophoresis for 33 kVh using an Amersham Biosciences Ettan IPGphor iso-electric focusing unit. Foci=ussed strips were treated as described by Hesketh et al., (2002) *supra* and then separated in the second dimension by SDS-PAGE using Amersham Biosciences DALT 12.5% gels. The gels were stained using a colloidal Coomassie Blue stain and scanned with a Proxpress Proteomic Imaging System scanner for later comparison. Protein spots of interest were subsequently excised from the gel, digested with trypsin and identified by MALDI-TOF peptide mass fingerprint analysis as described previously (Hesketh et al., (2002) *supra*).

Typical 2D gel electrophoretograms for the cell wall-associated proteome of these strains cultured on R5 medium are shown in Figure 2 and for MS and CM media in Figure 3. Significant differences in the staining patterns were observed between the two strains, and proteins that were present in M145 but absent from the *tatC* strain clearly represented candidates for Tat substrates. Putative Tat-targeted proteins were identified by MALDI-TOF mass spectrometry (after in-gel digestion with trypsin); several of them are marked in Figs. 2 and 3.

A total of 98 proteins that migrated with unique positions were identified in the M145 washes. In all, 37 of these proteins had identifiable N-terminal signal peptides, and 34 of those signal peptides contained RR dipeptides (Table 3). Cross-referencing with the predicted Tat substrates encoded by the *S. coelicolor* genome and identified by the programs TatP and TATFIND (Rose et al., (2002) Mol Microbiol 45:943-950; Bendtsen et al., (2005) BMC Bioinformatics 6:167-175), suggested that 21 of this group of 37 had plausible Tat-targeting signal peptides.

In addition to the 37 putative exported proteins, the remaining 61 proteins unique to the *tat*$^+$ strain included ribosomal subunit proteins and proteins of the thioredoxin pathway (Table 5). The identification of such obviously cytoplasmic proteins undoubtedly reflected contamination of the cell wall washes with cytoplasmic proteins from lysed bacteria. Cytoplasmic protein contamination in this type of analysis is almost inevitable given the *S. coelicolor* life cycle, which involves altruistic lysis of the mycelium to release nutrients for the continued growth of the aerial parts of the colony (Manteca et al., (2006) Res Microbiol 157:143-152). Moreover, such cytoplasmic protein contamination has been reported during analyses of extracellular proteins from *B. subtilis* (Antelmann et al., (2001) Genome Res 11;1483-1502) and *Mycobacterium tuberculosis* (Rosenkrands et al., (2000) Electrophoresis 21:935-948). Indeed, cell lysis, as well as substrate modification (see below) and up-regulation of Sec substrates, also may contribute to the presence of additional extracellular proteins in the Δ*tatC* strain that are absent in M145 (Table 4). Several of the exported proteins were detected as multiple spots in the M145 sample, possibly as a result of posttranslational modification or proteolysis, and it was not uncommon for one or more of the additional spots for a particular protein to be absent from cell wall fractions of the Δ*tatC* strain. Therefore, because several putative extracellular proteases are predicted TAT substrates (Dilks et al., (2003) J Bacteriol 185:1478-1483), the lack of a particular protein spot in the 2-DGE analysis of the Δ*tatC* strain might be indicative of the lack of postexport protein modification rather than a lack of TAT translocation *per se*.

TABLE 3

Exported proteins observed in the cell wall-associated fraction of the wild type (M145) strain identified by either 2-DGE and/or MudPIT

Sco number	Method of observation	Product	Putative signal peptide
Twin-arginine signal peptides			
SCO0432 TatP	2D	Probable secreted peptidase	MRRIRLLAAISAGLALAAGAVAPV PALAS (SEQ ID NO: 204)
SCO0474	MudPIT	Putative lipoprotein	MEPTVRTHPTRRLPTALVLTAAL LATGCSEQSDDGREPAA (SEQ ID NO: 205)
SCO1172 TatFind	2D	Putative amidase (putative secreted	MERERTAPVPTRRRLLKGAALA TVPYTLLSGTRAAAQ (SEQ ID NO: 206)

		protein)	
SCO1196 TatP and TatFind	2D	Putative secreted protein	MALTTRRRALTTLGAALTGAVAL PAGTALAS (SEQ ID NO: 207)
SCO1356 TatFind	MudPIT	Putative iron sulfur protein (putative secreted protein)	MTSAPFHPAAGPARRTVVAAAG AAGLTAVLAACSDSDDGASGD (SEQ ID NO: 208)
SCO1432 TatP and Tat Find	2D and MudPIT	Putative membrane protein.	MQRRTRPTRTPGGLPLMSRVEQ PSRRALLAAAVAAAAVTGGAIPA TARAAS (SEQ ID NO: 209)
SCO1565 TatP and TatFind	2D and MudPIT	Putative glycerophospho ryl diester phosphodiester ase	MGTQESDERAGGGTGRRALLG AAVLGAGGAVLGLPGTARAA (SEQ ID NO: 210)
SCO1590 TatP	2D	Putative secreted protein (transglycosidas e)	MGGVSRRAFTVAALSAFTLVPE ASAA (SEQ ID NO: 211)
SCO1639 TatFind	2D and MudPIT	Putative secreted peptidyl-prolyl cis-trans isomeraseprotei n	MRRRSLLIAVPTGLVTLAACGDS DDSGSSSNASESPSPDASATSA A (SEQ ID NO: 212)
SCO1824	MudPIT	Secreted subtilisin-like protease	MTAPLSRHRRALAIPAGLAVAAS LAFLPGTPAAATPAAEAA (SEQ ID NO: 213)
SCO1906	2D	Putative secreted protein (phoX phosphatase)	MSLTRRDFAGRSALTGAGVVLA GSVGALATAPNALAS (SEQ ID NO: 214)
SCO1948	2D	Putative zinc-binding carboxypeptidas e	MRRRARSILAVGALLIGGASFAPI AQAQ (SEQ ID NO: 215)
SCO2068 TatP and TatFind	2D and MudPIT	Putative secreted alkaline phosphatase	MTSRHRASENSRTPSRRTVVKA AAAGAVLAAPLAAALPAGAADAA P (SEQ ID NO: 216)
SCO2226	2D	Putative bi-functional protein (secreted alpha-amylase/dextrin ase)	MPATRRTARVRRVAAVTVTALA AALLPPLAARAD (SEQ ID NO: 217)

SCO2286*	2D	Putative alkaline phosphatase	MTPANHQAPTSAPSPAPSQSSH APELRAAARSLG**RR**RFLTVTGAA AALAFAVNLPAAGTASAA (SEQ ID NO: 218)
SCO2383	MudPIT	Putative secreted protein (PF00652 Ricin B lectin)	MRLRRPRRPSDGPPGRSGA**RR** LRVASAPLIPLTVIGAVVAGLIGP EPPSDGPLPGTSAVAD (SEQ ID NO: 219)
SCO2637	MudPIT	Putative serine protease (pfam00082, Peptidase_S8, Subtilase family)	MTNSPQREPIPGA**RR**AARLALAT GLAAALAAVGPVPVALAA (SEQ ID NO: 220)
SCO2758 TatP and TatFind	2D	Beta-N-acetylglucosaminidase (putative secreted protein)	MHHSSTAGTGSTAQPS**RR**SVLT ATAVVTAALAAGGGTAYAD (SEQ ID NO: 221)
SCO2780† TatFind	2D and MudPIT	Putative secreted protein (cd01146, FhuD, Fe^{3+}-siderophore binding domain FhuD)	MSHASATHPT**RR**GILAAGGALGL GAVLAACGDGDGKSDGA (SEQ ID NO: 222)
SCO2786	2D	Beta-N-acetylhexosaminidase (Transglycosidase)	MRPH**RR**HHRTTPRITRLLGSLLL VAAVGAMTTGAAPVRKAAAE (SEQ ID NO: 223)
SCO2821	2D	Putative secreted pectate lyase	M**RR**PVALRLQAALGTLALAAAA GVVLTMPEAAAAA (SEQ ID NO: 224)
SCO3456	MudPIT	Putative secreted protein (ABC transporter substrate binding protein)	MRHSG**RR**VKRATVLAAGAAVAG LLTAACSSPDTTGV (SEQ ID NO: 225)
SCO3484 TatP and TatFind	2D	Putative secreted sugar-binding protein	MTKPVVPSGVS**RR**GFLGGSLGV AGAVLLAACSGGGNSSQGS (SEQ ID NO: 226)
SCO3790 #TatP and TatFind	2D	Conserved hypothetical protein	MRKLLPLIGTPSGSHPGGRSAM TCRFRCGDACFHEVPNTSSNEY VGDVIAGALS**RR**SMMRAAAVVT

		of PhoX type	VAAAGAGAVGVAGAPSAQAA (SEQ ID NO: 227)
SCO4142†	2D	Phosphate-binding protein precursor	MNRRALALGALAVSGALALTAC GSDDTGGNSGSDSSSAAAN (SEQ ID NO: 228)
SCO4152	2D and MudPIT	Putative secreted 5'-nucleotidase	MPVTAQPHQPRRRRRTSRLLVV AATVVTAGVLAAAALPASASAG (SEQ ID NO: 229)
SCO4672 TatP and TatFind	2D and MudPIT	Putative secreted protein	MTHTSRRGVLAAFAASAATVPF GGAAAASPARATATAASAVDPA AASAA (SEQ ID NO: 230)
SCO4884	2D and MudPIT	Putative lipoprotein	MRRTSRLIRVAVGVASLALAATA CGGTSGESGD (SEQ ID NO: 231)
SCO4885	2D and MudPIT	Putative lipoprotein	MRRISRITVAGAATASLALALAAC GGTSTDSGSE (SEQ ID NO: 232)
SCO5461	2D	Putative secreted protein (ADP ribosylation)	MITTSLRRRTAAAVLSLSAVLATT AATAPGAAPAPSAAPAKAA (SEQ ID NO: 233)
SCO6009	2D	Solute-binding protein	MNTRMRRAAVAVATTAMAVSLA ACGSAK (SEQ ID NO: 234)
SCO6052 TatP and TatFind	2D	Putative membrane protein	MSTSRNAATRRQVLARTGALGA GIAFTGALSELFAGTAAAQ (SEQ ID NO: 235)
SCO6198† TatP and TatFind	MudPIT	Putative secreted protein	MDTTYWSRRRVLTVLGAATAAT SIPLAAPSRALAA (SEQ ID NO: 236)
SCO6272* TatP and TatFind	2D	Putative secreted FAD-binding protein	MTEVSRRKLMKGAAVSGGALAL PALGAPPATAAPAAGPEDLPGP AAAAAG (SEQ ID NO: 237)
SCO6281 TatP and TatFind	2D and MudPIT	Putative FAD-binding protein	MSAISRRRFIRHGAVAGGAAVAL PALGGWAGEAFAA (SEQ ID NO: 238)
SCO6418 TatP and TatFind	2D and MudPIT	Hypothetical protein (pectin lyase like)	MDRRALMLATGGLLGAAGAAQL TAAPSAVAA (SEQ ID NO: 239)
SCO6457 TatP and TatFind	2D	Putative beta-galactosidase (Galactose mutarotase-like)	MPHSPVSPAESPAPQPGRPRPV VSRRRLLEGGAAVLGALALSASP LTAQAAVRRAAAD (SEQ ID NO: 240)

SCO6580 #TatP and TatFind	2D	Hypothetical protein (Xylose isomerase-like)	MTPFTDSSRTDAGTDPSADGPG ESLRRALGVN**RR**RFLSTCTAVAA GAVAAPVFGASPALAH (SEQ ID NO: 241)
SCO6594 TatP and TatFind	MudPIT	Putative secreted protein	MPSSPTPSPTPSGAPEPSGV**RR** RSLLAAAAAAPVLASLAGAGTAA AD (SEQ ID NO: 242)
SCO6644	MudPIT	Putative solute-binding lipoprotein	MRSH**RR**PRLLAPFLLVPLMAGC FASGGGDDSASGD (SEQ ID NO: 243)
SCO6691 TatP and TatFind	2D	Putative phospholipase C	MAEVN**RR**RFLQLAGATTAFSAL SASIDRAAAL (SEQ ID NO: 244)
SCO7399	2D	Possible binding-protein-dependent transport lipoprotein (FhuD).	M**RR**LLLTAAATTAAALTLAACGT TEPAAD (SEQ ID NO: 245)
SCO7631 TatP	2D	Putative secreted protein	MSTEEQSSGRRHPD**RR**VLLRAA VAVPAAGAAVAGTAAVPAQAAD AA (SEQ ID NO: 246)

Non twin-arginine signal peptides			
SCO0930	MudPIT	Putative lipoprotein	MKTSWRSASLVAGAAALLALTTA CGQDGGGPTGSQNVGATAA (SEQ ID NO: 247)
SCO1396	MudPIT	Putative D-alanyl-D-alanine dipeptidase	MTRLSAALRGLATTFAALLAVTA VPAASAPATAQAR (SEQ ID NO: 248)
SCO1968†	2D and MudPIT	Putative secreted hydrolase	MHVRAVAVTTTALLGVALTAPLS HARAD (SEQ ID NO: 249)
SCO3967	MudPIT	Conserved Hypothetical membrane protein	MKASRIAAVGAVSGIAVLALSAP AFAHVSVQPEGAAAK (SEQ ID NO: 250)
SCO4010	MudPIT	Putative secreted protein (subtilisin inhibitor)	MFQVTRTPAARLLGAAALSAAAL ASVCAAPSAAVAGAG (SEQ ID NO: 251)
SCO5074	2D and MudPIT	Putative dehydratase	MTSSLHHAIRLTTASAIALGGLVT LGTSAHAA (SEQ ID NO: 252)
SCO6738	2D	Putative Carboxypeptidase (putative secreted protein)	MNRLRTLTATTAALAASLLVPSLL TAPAHAD (SEQ ID NO: 253)

Proteins are listed by SCO number and the signal peptide sequences are given in the right column. Twin-arginine dipeptides, where present, are shown in bold. Where multiple arginine dipeptides are present, only the most plausible one is marked. Method of observation indicates which technique was used to identify a particular protein (2D is two-dimensional gel electrophoresis; MudPIT is multidimensional protein identification technology).

There is some ambiguity about the assigned start codons for SCO3790 and SCO6580. For each of these proteins there is an alternative GTG start codon within the predicted signal peptide coding sequences. These are shown as valine residues that are highlighted in bold underline. Signal peptides initiating at these alternative start codons were tested in the agarase export assay (Fig 5) and in each case shown to mediate higher levels of agarase export than the full length signal peptides.

*These signal peptides have been shown to mediate Tat-dependent export when fused to the mature sequence of *S. lividans* xylanase C (Li H, Jacques PE, Ghinet MG, Brzezinski R, Morosoli R (2005) *Microbiology* 151: 2189-2198).

† These proteins have also been identified in the extracellular fraction of *S. coelicolor* grown in liquid medium (Kim DW, Chater K, Lee KJ, Hesketh, A (2005) *J Bacteriol* 187: 2957-2966).

#These proteins are only detected by TatP and TatFind when truncated *in silico* at the N-terminus to the underlined residue.

The annotation TatFind and/or TatP provided for SCO numbers indicates whether a particular protein has been identified as a putative Tat substrate by either of the two prediction programs TatFind 1.4 or TatP1.0 (information regarding the programs can be found on the world wide web at signalfind.org and cbs.dtu.dk/services/TatP-1.0, respectively).

TABLE 4

Exported proteins observed in the cell wall-associated fraction of the mutant *tatC* strain identified by either 2-DGE and/or MudPIT

Sco number	Method of observation	Product	Putative signal peptide
Twin-arginine signal peptides			
SCO0131 †	2D	Putative secreted protein (DNase I-like)	MFRRTLPVLAAAVGLAA AAAGPATAD (SEQ ID NO: 254)
SCO0462	MudPIT (mutual)	Putative oxidoreductase	MRIAVIGAGGVGGYFGA RLAAAGNEVTFVARGG HLAAIRRSGLVVHSPLGE LRTSPDSVVAS (SEQ ID NO: 255)
SCO0494	2D and MudPIT (mutual)	Putative iron-siderophore binding lipoprotein (FhuD type)	MLLRTTRTKPWRRLAAA LSAAALGVGLLAGCGSD SD (SEQ ID NO: 256)
SCO0736 TatP and TatFind	MudPIT (mutual)	Putative secreted protein	MGDIRRRGAVALGVTAL VAPLTLALTAAPAQAA (SEQ ID NO: 257)
SCO1230	2D and MudPIT (mutual)	Putative secreted tripeptidylaminopeptidase	LRKSSIRRRATAFGTAGA LVTATLIAGAVSAPAASA A (SEQ ID NO: 258)

TatP			
SCO1725	MudPIT (mutual)	Putative secreted hydrolase (similar to scabies esterase)	MRRFRLVGFLSSLVLAA GAALTGAATAQAA (SEQ ID NO: 259)
SCO1860 TatP	2D and MudPIT (mutual)	Putative secreted protein	MSLMRAPHACLTGGPTT LNGNTFRMPARRLATVA AATALAAGPATLAGAGS AHAT (SEQ ID NO: 260)
SCO2116†	2D and MudPIT *(~tatC)*	Putative secreted protein similar in parts to bacterial AmpD (amidase)	LSAEHRRNRNKKRLLVY GVASAVVVAATTGTLALA SPGLLGLDADQAAAA (SEQ ID NO: 261)
SCO2150	MudPIT *(~tatC)*	Cytochrome C heme-binding subunit	MKKLSARRRHPLAALVVL LLALACTGGLYAAFAPAS KAQAD (SEQ ID NO: 262)
SCO2217 TatFind	2D and MudPIT (mutual)	Putative secreted protein	MGRNTRKRRTPLATKIVA GAAALAIGGGGLVWN (SEQ ID NO: 263)
SCO2270	2D and MudPIT *(~tatC)*	Putative membrane protein	MPVRRRRPTALAAAVAT AAALGTTALAALGGATAA SAA (SEQ ID NO: 264)
SCO2446	2D and MudPIT (mutual)	Putative secreted peptidase (pfam00082, Peptidase_S8, Subtilase family)	MSTGPVRRGAALLSAGL VIALLPAGTSAAQ (SEQ ID NO: 265)
SCO2505	MudPIT (mutual)	Putative ABC-transporter metal-binding lipoprotein (cd01019, ZnuA, Zinc binding protein)	MNVRRRRISGIAVTAATA LGLGTLSACSSDSSAA (SEQ ID NO: 266)
SCO2591	2D* and MudPIT (~tatC)	Putative secreted protein	LFTSRARLRTRRSRLLTC TALAVAAGMLVTTPAVAA (SEQ ID NO: 267)
SCO2920	MudPIT (mutual)	Putative secreted protease pfam05547, Peptidase_M6, Immune inhibitor A peptidase M6	MTRRSWTFRTAATTVAF AAAAATFSAAGVAQAD (SEQ ID NO: 268)
SCO3540	2D	Proteinase (putative secreted protein)	MDTRRTHRRTRTGGTRF RATLLTAALLATACSAGG ASTSAGSPAAKAA (SEQ ID NO: 269)

SCO3667 TatP	MudPIT (mutual)	Putative secreted solute-binding protein	MSSRARRIAISVAASLMT FSLASCGMLGASESSD (SEQ ID NO: 270)
SCO4584	2D and MudPIT (mutual)	Putative membrane protein (lolA lipo-protein sorting protein)	MAPIDSDDNTTAGAGEE LRAGRRKAARYVVPVAV MGVAAATIGLVPALA (SEQ ID NO: 271)
SCO5022 TatFind	MudPIT *(~tatC)*	Putative lipoprotein	MNRRRPTLLAALTLTAAA ALTLSACGS (SEQ ID NO: 272)
SCO5260	MudPIT *(~tatC)*	Secreted protein (periplasmic binding protein AtrA)	MTARSTRRTTAAHSRLA AVGAIAVAGALLLTGCGD QTENGD (SEQ ID NO: 273)
SCO5447	2D	Putative neutral zinc metalloprotease	MLSSSTPHRRTSHTTSH TSRRTAHRRAAAVALAG VAALIATAVQSGTATAA (SEQ ID NO: 274)
SCO5477†	2D and MudPIT (mutual)	Putative oligopeptide-binding lipoprotein	MTTQRTSGRRKQALAAA AVVAALLTTAACGGGGD DDSGGG (SEQ ID NO: 275)
SCO5650	MudPIT *(~tatC)*	Putative membrane protein	MAVLSIISRTASRRGVRA LGVVAASAALALSASGNA LAC (SEQ ID NO: 276)
SCO6096	MudPIT *(~tatC)*	Putative lipoprotein	MPATSALRRTLAVIAALP LLTLAACGYGSQAK (SEQ ID NO: 277)
SCO6108	2D and MudPIT (mutual)	Esterase fusH	MATLIPKKGSTLLNKGIRT RRARGALAGGTVLTAAA ALLTAVPAAQAI (SEQ ID NO: 278)
SCO6197 TatP	2D and MudPIT (mutual)	Putative secreted protein	MSTTSVNGSRSKRRVLL RRALPVCVAAGVASIVVF GSGSSPDAAAR (SEQ ID NO: 279)
SCO6407	2D and MudPIT (mutual)	Putative gamma-glutamyltranspeptidase (putative secreted protein)	MRRPVARNLTVLAVSAA VVVSVGAAAP (SEQ ID NO: 280)
SCO6572	2D	Putative glycosyl hydrolase (putative secreted protein)	MRSVGAGRWRRVASQV GRSVTVWAVSTLTMLVL AGILPDFRLQSPDGDSAT DIAMTAAVGAGA (SEQ ID NO: 281)
SCO6590	2D and MudPIT (mutual)	Putative secreted esterase (trypsin like serine protease)	MSSIRRSRSAVLGAVLTA GSLALSTAAAVAVTGGTP VADSD (SEQ ID NO: 282)

SCO6608	2D and MudPIT (mutual)	Putative secreted protein (exopolysaccharide biosynthesis)	MKLRKRRLCAATGVVLG ALTALAALPPAAVAS (SEQ ID NO: 283)
SCO6736	2D and MudPIT (mutual)	Putative metallopeptidase	MPLRLGAFVHRRFIVPGA LGAASLMLAIPASAAA (SEQ ID NO: 284)
SCO7657 TatP	2D and MudPIT (mutual)	Putative secreted protein	MTRTRTTRLRRPLLTGST AVAAMAVTAGLVAATAE PAKAA (SEQ ID NO: 285)
SCO7677† TatP and TatFind	2D and MudPIT (mutual)	Putative secreted solute-binding protein	MSRQIDRRSFLRRGAAG AAALAVGPGLLAACSTDE PGSA (SEQ ID NO: 286)
Non twin-arginine signal peptides			
SCO0297†	2D and MudPIT (mutual)	Putative secreted protein	MRSTGPVSRIGRSLAAVT ATAAAGAVAALGLAASPA AAA (SEQ ID NO: 287)
SCO0324	2D	Putative secreted protein (superfamily Quinoprotein amine dehydrogenase)	MSPKSKRTRAALAALVV AAGSLVTAGAARAE (SEQ ID NO: 288)
SCO0472	MudPIT (mutual)	Putative secreted protein (Quinoprotein amine dehydrogenase)	LNTSIRNRAVTGTALALA VSAVLTACGGDTSSSD (SEQ ID NO: 289)
SCO0762†	2D and MudPIT (mutual)	Putative secreted protein	MRNTARWAATLGLTATA VCGPLAGASLAS (SEQ ID NO: 2890)
SCO2008	MudPIT (mutual)	Putative branched chain amino acid binding protein	MPGKGLIVRQRSLIAITAA LAAGALTLTACGS (SEQ ID NO: 291)
SCO2096	MudPIT *(~tatC)*	Putative membrane protein (Transglutaminase/protease-like	MMSGRARLAMCAWAAT LMAACALLPLVEPATWIL QAA (SEQ ID NO: 292)
SCO2271	MudPIT *(~tatC)*	Hypothetical secreted protein	MPARLRALVTVLCAAVLA ALLPAAAAHAE (SEQ ID NO: 293)
SCO2483	2D	Putative secreted protein	MHSTTSRGAVLATLLALI GVFATYRALPAGAD (SEQ ID NO: 294)
SCO2892	2D and MudPIT (mutual)	Putative secreted protein	MTRGRDGGAGAPPTKH RALLAAIVTLIVAISAAIYA GASAD

			(SEQ ID NO: 295)
SCO3053	2D and MudPIT (mutual)	Putative secreted esterase (FusH-like)	MSRFRSRAAVTAGLTAIA VAAGCLVVGPVQAA (SEQ ID NO: 296)
SCO3176	2D	Putative membrane protein	MRPSSLRSPWTRAAVAG AAAAGLAAAGLAGVAHA A (SEQ ID NO: 297)
SCO3487	2D	Putative hydrolase	MTVHKRACTTPPPRASR SFRVRWPVLIAAACAGLV LATTSPPAVAA (SEQ ID NO: 298)
SCO3966	MudPIT *(~tatC)*	Putative secreted protein (thioredoxin-like)	MRKKTFATAALLAAACLT LSACGSGADSD (SEQ ID NO: 299)
SCO4589	2D	Putative aminopeptidase (putative secreted protein)(Zn dependent)	MQLLPSGRALTAGAVAV VTLMAGGSAAGAASAPV HPATPTGTAAAAAA (SEQ ID NO: 300)
SCO4891	MudPIT *(~tatC)*	Putative secreted protein	MRTRTKLTTVLASGTLAA GLLVGGASSASAG (SEQ ID NO: 301)
SCO5113[†]	2D and MudPIT (mutual)	BldKB, putative ABC transport system lipoprotein	MSILRNRTATAAIVAVAA GALTLTACGGGDSDSSA K (SEQ ID NO: 302)
SCO5232	MudPIT *(~tatC)*	Putative sugar transporter sugar binding protein	MKRKLIAAIGIAGMMVSIA ACGG (SEQ ID NO: 303)
SCO5420	2D and MudPIT *(~tatC)*	Cholesterol esterase	MESQVRGGTRWKRFAV VMVPSVAATAAIGVALAQ GALAA (SEQ ID NO: 304)
SCO5660	MudPIT (mutual)	Putative secreted peptidase	MRARSHAVLAASVVLAL AAGPLAAPAFAA (SEQ ID NO: 305)
SCO5776	MudPIT (mutual)	Glutamate binding protein	MKLRKVSAASAAVLALAL TATACGG (SEQ ID NO: 306)
SCO6109	2D and MudPIT (mutual)	Putative secreted hydrolase (fusH-like)	MFAMPRSARLTALSAALL AGALASSSTPALATAGA (SEQ ID NO: 307)

SCO6199	2D and MudPIT (mutual)	Secreted esterase	MHRTRPGGHPRSIRFAT LAISLTAGSALMTASPAV AVTGPAAAD (SEQ ID NO: 308)
SCO6591	2D and MudPIT (mutual)	Putative secreted protein	MKKRSSFVRVLGAAATA GALAWAVLAQGNAGAA (SEQ ID NO: 309)
SCO6710	2D	Putative glycosyl hydrolase	MWAALLAGLLMVLGLQA TSASGQPAKAPAAAAA (SEQ ID NO: 310)
SCO6723	2D and MudPIT (mutual)	Putative oxidoreductase	MALAGAVCAAALAALTAV PAQAH (SEQ ID NO: 311)
SCO7250	2D and MudPIT (mutual)	Conserved hypothetical protein (Amidase_2, N-acetylmuramoyl-L-alanine amidase)	MPALATAALLLPLLGAAP SAAAPTAAP (SEQ ID NO: 312)

The table lists proteins by SCO number and the signal peptide sequences are given in the right column. Method of observation indicates which technique was used to identify a particular protein (2D is two-dimensional gel electrophoresis; MudPIT is multidimensional protein identification technology. Bracketed after MudPIT is indicated *"tatC'* if the protein was found in the cell-wall fraction of the *'tatC* mutant strain only, and 'mututal' if it was found in the cell-wall fractions of both the M145 and the *'tatC* mutant strains. The annotation TatFind and/or TatP provided for SCO numbers indicates whether a particular protein has been identified as a putative Tat substrate by either of the two prediction programs TatFind or TatP.

*This protein migrated in a unique position on two-dimensional gel analysis of the extracellular fraction from the M145 strain. However it was detected in the *tatC* mutant strain by MudPIT analysis.

†These proteins have also been identified in the extracellular fraction of *S. coelicolor* grown in liquid medium (Kim et al., (2005) *J Bacteriol* 187: 2957-2966).

Example 5

Identification of Tat-dependent proteins by Multidimensional Protein Identification Technology (MudPIT)

In parallel to the traditional 2-D gel electrophoresis approach, the cell wall proteome of the *S. coelicolor* strains was analyzed by multidimensional protein identification technology
(MudPIT), a sensitive modern technique used to separate and identify even low-abundance proteins from complex mixtures.

Samples of precipitated protein (1 mg) prepared as described above, were dissolved by addition of 100 µl 100mM Tris-HCl, pH 8.0 200 µl 0.4% RapiGest TM (Waters Ltd, Elstree, herts, UK) in 20mM Tric-HCl, pH 8.0 10µl of 40 mM EDTA in water and 40 µl of 45 mMDTT in 100mM Tris-Hcl pH 8.0. Once the pellets dissolved the samples were heated in a water bath at 90°C for 30 min then cooled to room temperature. 40µl of 100mM iodoacetamide in 100mM Tris-HCl pH8.0 was added and the sample incubated in the dark for 15 min. 10 µg trypsin (modified sequence grade porcine, Promega, UK) in 10 µl 20, < Tris-HCl pH8.0 was added to each sample which were then incubated at 37°C. After 16 h further 10 µg of trypsin was added and the samples incubated for a further 8h. The RapiGest was denatured prior to MudPIT mass spectrometry by the addition of 40 µl of 500 mMHCl to give a concentration between 30 – 50 mM (pH < 2) followed by incubation at 37°C for 45 min. The cloudy samples containing hydrolyzed detergent were centrifuged at 13,000 rpm for 10 min in a BioFuge TM (Heraeus) and the supernatants carefully removed for chromatographic separation.

Samples were loaded onto a biphasic column comprising a strong cation exchange phase (SCX) and a reverse phase. Peptides were eluted stepwise from the SCX phase by using increasing concentrations of salt onto the reverse phase. A reverse-phase gradient then was generated and peptides were eluted into a Q-ToF2 mass spectrometer (Micromass, Manchester, U.K.) The data from each reverse-phase gradient was combined and searched by using MASCOT (Matrix Science, London, U.K. (Perkins et al., Electrophoresis 20:3551-3567 (1999)). The details for the analysis of the samples are as follows.

For chromatographic separation, a 75 µm PicoFrot capillary (NewObjective, Inc.) was packed first with 90 mm of Symmetry C18 300Å reverse-phase material (Waters, Ltd, UK) followed by 30 mm of PartiSphere strong cation exchange material (Whatman) using a pressure packing device. The resulting biphasic micro capillary column was equilibrated to

5% acetonitrile/0.1% formic acid. After loading the sample the column was mounted on the Z-spray ion source of a Q-ToF2 mass spectrometer (Micromass, Manchester, UK) and in-line with a capillary HPLC system (CapLC, Water Ltd., UK). A fully automated 9-step chromatography run was carried out, with the mass spectrometer operating in data-dependent modes during each reverse phase elution. The three buffer solutions used for chromatography were 0.1% formic acid (buffer A), 100% acetonitrile/0.1% formic acid (buffer b) and 500mM ammonium acetate/0.1% formic acid (buffer c). Elution was performed using increasing concentrations of buffer C followed by a reverse-phase gradient. Buffer C concentrations were 0, 10, 25, 50, 75, 100, 200, 300 and 500mM. The reverse phase gradients consisted of the following profile: t = 0 min, 5%B; t=80, 40%B: t=90, 80%B.

The eluted peptides were detected in MS mode in order to select precursor ions for collision-induced dissociation (MS/MS). Every 1.2s the three most intense signals were subjected to MS/MS analysis. The resulting MS/MS data was deconvoluted using MaxEnt3 (Micromass) and converted into a text file listing the mass, intensity and charge state of the parent ions and mass and intensity of the associated fragment ions using ProteinLynx (Micromass). These data were search using the MASCOT search tool (Matrix Science Ltd, London) against SCODB (the S. coelicolor protein database) using appropriate parameters. The sole fixed modification was carboxyamidomethyl (Cys) and the only variable modification was oxidation (Met). The enzyme was selected as trypsin and the maximum number of missed cleavages was set as 3. Peptide M_W tolerance and MS/MS fragment ion tolerance were set as + 0.25Da.

The criteria for protein identification were based on the manufacturer's definitions (Matrix Science, Ltd.) Candidate peptides with probability based Mowse scores exceeding threshold ($p< 0.05$) indicated a significant homology and were referred to as "hits". Protein scores were derived from peptide ion scores as a non-probabilistic basis for ranking proteins.

MudPIT analysis of the M145 and Δ*tatC* strains grown on CM medium identified 308 and 279 individual proteins from the cell wall washes, respectively, of which 188 were common to both samples (Tables 3-5). Of the 120 remaining proteins exclusively present in M145 cell wall sample, 20 corresponded to proteins bearing plausible N-terminal signal peptides that contained two consecutive arginines in their sequences (including 11 that had been identified by 2-DGE analysis), strongly suggesting that this group represents the Tat substrates (Table 3).

TABLE 5

Cytoplasmic proteins observed in the extracellular fraction of *S. coelicolor* identified by either 2-DGE and/or MudPIT

Protein SCO number	Method of observation	Annotation
SCO0256	MudPIT (Mutual)	Putative short chain oxidoreductase
SCO0168	2D	Possible regulator protein
SCO0179	2D	Putative zinc-containing dehydrogenase
SCO0180	2D	Conserved hypothetical protein SCJ1.29c
SCO0199	2D	Putative alcohol dehydrogenase
SCO0204	2D	Putative luxR family two-component response regulator
SCO0260	2D and MudPIT *(~tatC)*	Conserved hypothetical protein
SCO0379†	2D and MudPIT (Mutual)	Catalase (EC 1.11.1.6)
SCO0400	2D	Putative epimerase
SCO0492	MudPIT *(~tatC)*	Putative peptide synthetase
SCO0499†	MudPIT *(~tatC)*	Putative formyltransferase
SCO0525	MudPIT *(~tatC)*	Conserved hypothetical protein SCF11.05
SCO0527	2D and MudPIT (mutual)	Cold shock protein
SCO0570	MudPIT *(~tatC)*	50S ribosomal protein L33
SCO0592	2D	Hypothetical protein SCF55.16c
SCO0596	2D	Putative DNA-binding protein
SCO0597	2D	Conserved hypothetical protein SCF55.21 c
SCO0640	MudPIT *(~tatC)*	Conserved hypothetical protein SCF56.24c
SCO0641	MudPIT (mutual)	Tellurium resistance protein
SCO0682	MudPIT (M145)	Hypothetical protein SCF15.03c
SCO0714	2D	Putative oxidoreductase
SCO0740	2D	Putative hydrolase
SCO0741	2D	Putative oxidoreductase
SCO0788	MudPIT (mutual)	Hypothetical protein 3SCF60.20c

SCO0797	2D	Putative integral membrane protein.
SCO0852	2D and MudPIT *(~tatC)*	Putative aldolase
SCO0870	MudPIT (M145)	Putative two-component system response regulator
SCO0884	MudPIT *(~tatC)*	Probable oxidoreductase
SCO0885	MudPIT *(~tatC)*	Thioredoxin
SCO0906	2D	Hypothetical protein SCM1.39c
SCO0945	2D	Putative formamidopyrimidine-DNA glycosylase
SCO0975	MudPIT (M145)	6-phosphogluconate 1-dehydrogenase
SCO0977	MudPIT (mutual)	Conserved hypothetical protein
SCO0999	MudPIT (mutual)	Superoxide dismutase
SCO1081	2D and MudPIT *(~tatC)*	Putative electron transfer flavoprotein, alpha subunit
SCO1082	MudPIT *(~tatC)*	Putative electron transfer flavoprotein, beta subunit
SCO1085	MudPIT *(~tatC)*	Putative acyltransferase
SCO1116	MudPIT (mutual)	Hypothetical protein
SCO1123	2D and MudPIT *(~tatC)*	Hypothetical protein 2SCG38.16
SCO1198	MudPIT *(~tatC)*	Putative acyl-CoA dehydrogenase
SCO1199	MudPIT *(~tatC)*	Putative oxidoreductase
SCO1337	MudPIT (M145)	Putative oxidoreductase
SCO1384	2D and MudPIT (mutual)	Conserved hypothetical protein
SCO1390	MudPIT (mutual)	Putative PTS system sugar phosphotransferase component IIA
SCO1391	MudPIT (M145)	Phosphoenolpyruvate-protein phosphotransferase
SCO1438	2D	ATP phosphoribosyltransferase
SCO1453	MudPIT *(~tatC)*	Hypothetical protein
SCO1462	MudPIT (M145)	Hypothetical protein SCL6.19c
SCO1476	2D	S-adenosylmethionine synthetase
SCO1480	MudPIT (mutual)	Conserved hypothetical protein
SCO1489	2D	Putative DNA-binding protein

SCO1496	2D	Chorismate synthase
SCO1505	MudPIT (mutual)	30S ribosomal protein S4
SCO1508	2D	Histidyl tRNA synthetase
SCO1569	MudPIT *(~tatC)*	Putative oxidoreductase
SCO1594	MudPIT (mutual)	Putative phenylalanyl-tRNA synthetase beta chain
SCO1598	MudPIT (M145)	50S ribosomal protein L20
SCO1600	MudPIT (M145)	Putative translation initiation factor IF-3
SCO1620	MudPIT *(~tatC)*	Glycine betaine transport system permease protein
SCO1623	MudPIT (M145)	Conserved hypothetical protein SCI41.06
SCO1626	MudPIT (mutual)	Putative cytochrome P450
SCO1628	MudPIT (M145)	Conserved hypothetical protein
SCO1643	MudPIT (mutual)	20S proteasome alpha-subunit
SCO1644	MudPIT (mutual)	20S proteasome beta-subunit precursor
SCO1646	MudPIT (mutual)	Hypothetical protein
SCO1651	MudPIT (M145)	Conserved hypothetical protein SCI41.34c
SCO1654	2D	Putative two-component response regulator
SCO1766	MudPIT (M145)	Putative glycohydrolase
SCO1767	2D	Putative DNA hydrolase
SCO1773	2D	Putative L-alanine dehydrogenase
SCO1775	2D	Conserved hypothetical protein
SCO1788	MudPIT (mutual)	Conserved hypothetical protein
SCO1794	2D and MudPIT (M145)	Conserved hypothetical protein
SCO1809	MudPIT (mutual)	Conserved hypothetical protein
SCO1839	MudPIT (mutual)	Putative transcriptional regulator
SCO1870	MudPIT (M145)	Hypothetical protein SCI39.17c
SCO1921	MudPIT (mutual)	Putative aminotransferase
SCO1922	MudPIT (mutual)	Putative ABC transporter ATP-binding subunit

SCO1923	MudPIT (mutual)	Putative dioxygenase ferredoxin subunit
SCO1935	2D	Transketolase A
SCO1936	MudPIT (mutual)	Putative transaldolase
SCO1937	2D	Putative glucose-6-phosphate 1-dehydrogenase
SCO1945†	2D and MudPIT (mutual)	Triosephosphate isomerase
SCO1946	2D and MudPIT (mutual)	Phosphoglycerate kinase
SCO1947†	2D and MudPIT (mutual)	Glyceraldehyde-3-phosphate dehydrogenase
SCO1998	2D and MudPIT (mutual)	30S ribosomal protein S1
SCO2011	MudPIT (M145)	Putative branched chain amino acid transport ATP-binding protein
SCO2012	MudPIT (mutual)	Putative branched chain amino acid transport ATP-binding protein
SCO2039	MudPIT (M145)	Indoleglycerol phosphate synthase
SCO2045	2D	Conserved hypothetical protein
SCO2050	MudPIT (M145)	Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase
SCO2077	MudPIT *(~tatC)*	Hypothetical protein
SCO2105	MudPIT *(~tatC)*	Putative transcriptional regulatory protein
SCO2126	MudPIT *(~tatC)*	Glucokinase
SCO2149	MudPIT (M145)	Rieske iron-sulfur protein
SCO2167	MudPIT *(~tatC)*	Hypothetical protein
SCO2168	MudPIT *(~tatC)*	Hypothetical protein
SCO2180†	2D and MudPIT (mutual)	Putative dihydrolipoamide dehydrogenase
SCO2181	2D and MudPIT (M145)	Putative dihydrolipoamide succinyltransferase

SCO2183	MudPIT *(~tatC)*	Putative pyruvate dehydrogenase E1 component
SCO2198†	2D and MudPIT (mutual)	Glutamine synthetase I
SCO2206	MudPIT (M145)	ArsC, arsenate reductase
SCO2210	2D and MudPIT (M145)	Glutamine synthetase
SCO2262	2D and MudPIT (M145)	Putative oxidoreductase
SCO2265	MudPIT *(~tatC)*	Hypothetical protein SCC75A.1 1c.
SCO2339	MudPIT (mutual)	Hypothetical protein
SCO2367	2D and MudPIT (mutual)	Conserved hypothetical protein
SCO2368	2D and MudPIT (mutual)	Conserved hypothetical protein
SCO2369	MudPIT *(~tatC)*	Putative thiol-specific antioxidant protein
SCO2370	MudPIT (mutual)	Conserved hypothetical protein
SCO2380	MudPIT *(~tatC)*	Conserved hypothetical protein SC4A7.08
SCO2382	MudPIT *(~tatC)*	Hypothetical protein SC4A7.10
SCO2389	MudPIT (mutual)	Acyl carrier protein
SCO2394	2D and MudPIT (M145)	Putative secreted protein
SCO2407	2D	Putative aldose 1-epimerase
SCO2469	MudPIT (mutual)	Putative reductase
SCO2504	2D	Glycyl-tRNA synthetase
SCO2506	MudPIT *(~tatC)*	Putative metal transport ABC transporter
SCO2529	2D	Putative metalloprotease.
SCO2548	MudPIT (mutual)	Putative Hit-family protein.
SCO2554	MudPIT (M145)	DnaJ protein.
SCO2576	MudPIT *(~tatC)*	Phosphoglycerate mutase
SCO2582	MudPIT (mutual)	Conserved hypothetical protein SCC 123.20.
SCO2596	MudPIT (mutual)	50S ribosomal protein L27

SCO2597	MudPIT (M145)	Ribosomal protein L21
SCO2612	2D and MudPIT (M145)	Nucleoside diphosphate kinase
SCO2627	MudPIT *(~tatC)*	Putative sugar-phosphate isomerase
SCO2633	2D and MudPIT (mutual)	Superoxide dismutase [Fe-Zn] (EC 1.15.1.1)
SCO2634	MudPIT (mutual)	Conserved hypothetical protein SC8E4A.04c
SCO2635	MudPIT *(~tatC)*	Putative aminopeptidase
SCO2643†	MudPIT (mutual)	Aminopeptidase N
SCO2696	MudPIT (M145)	Putative 2-hydroxyacid-family dehydrogenase
SCO2736†	MudPIT (mutual)	Citrate synthase.
SCO2822	2D and MudPIT (M145)	Putative decarboxylase
SCO2849	MudPIT (M145)	Conserved hypothetical protein SCE20.23.
SCO2901	2D and MudPIT (mutual)	Hypothetical protein
SCO2902	2D	Conserved hypothetical protein
SCO2904	2D and MudPIT (mutual)	Putative ribonuclease PH
SCO2905	MudPIT (mutual)	Hypothetical protein
SCO2911	MudPIT (mutual)	Conserved hypothetical protein
SCO2950	MudPIT (mutual)	DNA-binding protein Hu (hs1)
SCO2951	2D and MudPIT (mutual)	Putative malate oxidoreductase
SCO2999	MudPIT *(~tatC)*	Conserved hypothetical protein
SCO3023	MudPIT (M145)	Adenosylhomocysteinase
SCO3027	MudPIT (mutual)	Hypothetical protein SCD84.08c
SCO3096†	2D and MudPIT (mutual)	Enolase

SCO3122	MudPIT (M145)	Putative nucleotidyltransferase
SCO3123†	MudPIT *(~tatC)*	Ribose-phosphate pyrophosphokinase
SCO3124	MudPIT (mutual)	Ribosomal L25p family protein
SCO3127	MudPIT (mutual)	Phosphoenolpyruvate carboxylase
SCO3182	MudPIT (M145)	UTP-glucose-1-phosphate uridylyltransferase
SCO3187	MudPIT *(~tatC)*	Conserved hypothetical protein SCE22.04
SCO3303	MudPIT *(~tatC)*	Putative lysyl-tRNA synthetase
SCO3304	MudPIT *(~tatC)*	Putative arginyl-tRNA synthetase
SCO3328	MudPIT (M145)	Hypothetical protein
SCO3337	2D	Pyrroline-5-carboxylate reductase
SCO3345	2D	Dihydroxy acid dehydratase
SCO3373	MudPIT (M145)	Putative Clp-family ATP-binding protease
SCO3407	2D	Conserved hypothetical protein
SCO3409	MudPIT (mutual)	Putative inorganic pyrophosphatase
SCO3425	MudPIT (M145)	Putative 30S ribosomal protein S18
SCO3427	MudPIT (M145)	Putative 50S ribosomal protein L31
SCO3429	MudPIT (mutual)	Putative 50S ribosomal protein L28
SCO3438	MudPIT (M145)	Hypothetical protein
SCO3443	2D	Putative pyridine nucleotide-disulphide oxidoreductase
SCO3463	MudPIT (M145)	Putative phosphorylase
SCO3490	2D	Transposase
SCO3549	MudPIT (M145)	Putative anti-sigma factor antagonist
SCO3549	2D and MudPIT *(~tatC)*	putative anti-sigma factor antagonist
SCO3569	MudPIT (mutual)	Putative endonuclease
SCO3571	2D	Putative transcriptional regulator
SCO3575	MudPIT (M145)	Conserved hypothetical protein
SCO3617	MudPIT (M145)	Hypothetical protein
SCO3622	2D	Putative aminotransferase

SCO3629	2D and MudPIT (M145)	Adenylosuccinate synthetase
SCO3647	2D and MudPIT (M145)	Conserved hypothetical protein
SCO3649†	2D and MudPIT (mutual)	Putative fructose 1,6-bisphosphate aldolase
SCO3659	MudPIT *(~tatC)*	Hypothetical protein
SCO3661	MudPIT (M145)	ATP-dependent protease ATP-binding subunit
SCO3671	MudPIT (mutual)	Heat shock protein 70
SCO3731	MudPIT (mutual)	Cold-shock protein
SCO3734	MudPIT (M145)	Hypothetical protein
SCO3748	MudPIT (mutual)	Cold shock protein
SCO3767	2D and MudPIT (mutual)	Conserved hypothetical protein
SCO3790†	2D	Conserved hypothetical protein
SCO3792†	MudPIT (mutual)	Putative methionyl tRNA synthetase
SCO3793	MudPIT (M145)	Conserved hypothetical protein
SCO3823	2D	Putative quinone oxidoreductase
SCO3834	MudPIT (mutual)	Putative 3-Hydroxyacyl-CoA dehydrogenase
SCO3877	MudPIT *(~tatC)*	Putative 6-phosphogluconate dehydrogenase
SCO3878	MudPIT (mutual)	DNA polymerase III, beta chain
SCO3889	MudPIT (mutual)	Thioredoxin
SCO3890	MudPIT (M145)	Thioredoxin reductase (NADPH)
SCO3899	2D and MudPIT (M145)	Hypothetical protein
SCO3906	MudPIT (mutual)	Putative 30S ribosomal protein S6
SCO3907	2D and MudPIT (M145)	Putative single-strand DNA-binding protein
SCO3909	MudPIT (mutual)	50S ribosomal protein L9
SCO3913	2D	Conserved hypothetical protein
SCO3917	MudPIT (M145)	Conserved hypothetical protein

SCO3920	MudPIT *(~tatC)*	Putative cystathionine/methionine gamma-synthase/lyase
SCO3956	MudPIT *(~tatC)*	Putative ABC transporter ATP-binding protein
SCO3974	MudPIT *(~tatC)*	Hypothetical protein
SCO4036	MudPIT (mutual)	Hypothetical protein
SCO4039	MudPIT (mutual)	Hypothetical protein 2SCD60.05c
SCO4041	MudPIT (M145)	Putative uracil phosphoribosyltransferase
SCO4043	MudPIT *(~tatC)*	Conserved hypothetical protein
SCO4048	MudPIT (M145)	Conserved hypothetical protein
SCO4068	2D	Phosphoribosylamine-glycine ligase (EC 6.3.4.13)
SCO4086	MudPIT (M145)	Amidophosphoribosyltransferase
SCO4087	2D and MudPIT (mutual)	Phosphoribosylformylglycinamidine cyclo-ligase
SCO4089†	MudPIT *(~tatC)*	Valine dehydrogenase (EC 1.4.1.-)
SCO4091	MudPIT (mutual)	Putative DNA-binding protein
SCO4109	2D and MudPIT (mutual)	Putative oxidoreductase
SCO4164	MudPIT (mutual)	Putative thiosulfate sulfurtransferase
SCO4165	MudPIT (mutual)	Conserved hypothetical protein
SCO4173	MudPIT *(~tatC)*	Hypothetical protein SCD66.10c
SCO4175	2D and MudPIT *(~tatC)*	Hypothetical protein SCD66.12c
SCO4199	MudPIT (M145)	Hypothetical protein
SCO4228	2D	Putative phosphate transport system regulator
SCO4240	2D	ABC transporter ATP-binding protein

SCO4246	MudPIT (M145)	Hypothetical protein SCD8A.19c
SCO4277	2D and MudPIT (mutual)	Putative tellurium resistance protein
SCO4294	MudPIT *(~tatC)*	Conserved hypothetical protein
SCO4295	MudPIT (M145)	Cold shock protein
SCO4296	2D and MudPIT (mutual)	Chaperonin 2
SCO4366†	2D and MudPIT (M145)	Putative phosphoserine Aminotransferase
SCO4411	MudPIT (mutual)	Putative calcium binding protein
SCO4441	2D	Putative DNA-binding protein
SCO4505	2D and MudPIT (mutual)	Cold shock protein
SCO4509	MudPIT (mutual)	Hypothetical protein SCD35.16
SCO4580	MudPIT (mutual)	Putative fumarylacetoacetase
SCO4614	2D and MudPIT (mutual)	Conserved hypothetical protein
SCO4645	MudPIT (mutual)	Aspartate aminotransferase
SCO4649	MudPIT (mutual)	50S ribosomal protein L1
SCO4652	2D and MudPIT (mutual)	50S ribosomal protein L10
SCO4653	MudPIT (mutual)	50S ribosomal protein L7/L12
SCO4655	MudPIT *(~tatC)*	DNA-directed RNA polymerase beta' chain (fragment)
SCO4659	MudPIT (mutual)	30S ribosomal protein S12
SCO4660	MudPIT (mutual)	30S ribosomal protein S7
SCO4662	2D and MudPIT (mutual)	Elongation factor TU-1
SCO4675	MudPIT (mutual)	Conserved hypothetical protein SCD40A.21 c
SCO4687	MudPIT (mutual)	Hypothetical protein SCD31.12c
SCO4701	MudPIT (M145)	30S ribosomal protein S10
SCO4702	MudPIT (mutual)	50S ribosomal protein L3
SCO4703	2D and MudPIT (mutual)	50S ribosomal protein L4
SCO4704	MudPIT (mutual)	50S ribosomal protein L23
SCO4705	MudPIT (mutual)	50S ribosomal protein L2
SCO4706	MudPIT (mutual)	30S ribosomal protein S19

SCO4707	MudPIT (M145)	50S ribosomal protein L22
SCO4709	MudPIT (M145)	50S ribosomal protein L16
SCO4710	MudPIT (M145)	50S ribosomal protein L29
SCO4711	MudPIT (M145)	30S ribosomal protein S17
SCO4713	MudPIT (mutual)	50S ribosomal protein L24
SCO4714	MudPIT (M145)	50S ribosomal protein L5
SCO4716	MudPIT (mutual)	30S ribosomal protein S8
SCO4717	MudPIT (mutual)	50S ribosomal protein L6
SCO4718	MudPIT (mutual)	50S ribosomal protein L18
SCO4719	MudPIT (mutual)	30S ribosomal protein S5
SCO4720	MudPIT (mutual)	50S ribosomal protein L30
SCO4721	MudPIT (mutual)	50S ribosomal protein L15
SCO4723	MudPIT (M145)	Adenylate kinase
SCO4725	MudPIT (mutual)	Translational initiation factor IF1
SCO4726	MudPIT (M145)	50S ribosomal protein L36
SCO4727	MudPIT (mutual)	30S ribosomal protein S13
SCO4729	2D and MudPIT (mutual)	DNA-directed RNA polymerase alpha chain
SCO4730	MudPIT (mutual)	50S ribosomal protein L17
SCO4734	MudPIT (mutual)	50S ribosomal protein L13
SCO4735	MudPIT (mutual)	30S ribosomal protein S9
SCO4761	2D and MudPIT (mutual)	10 kD chaperonin cpn10
SCO4762	MudPIT (mutual)	60 kD chaperonin cpn60
SCO4770	2D and MudPIT (M145)	Inosine 5' monophosphate dehydrogenase
SCO4771†	2D	Putative inosine-5'-monophosphate dehydrogenase
SCO4800	MudPIT *(~tatC)*	Isobutiryl CoA mutase, small subunit
SCO4808	MudPIT (mutual)	Succinyl-CoA synthetase beta chain
SCO4809†	MudPIT (mutual)	Succinyl CoA synthetase alpha chain
SCO4814	MudPIT (M145)	Bifunctional purine biosynthesis protein

SCO4824	2D and MudPIT (mutual)	Bifunctional protein (methylenetetrahydrofolate dehydrogenase and methenyltetrahydrofolate cyclohydrolase.
SCO4827†	MudPIT (mutual)	Malate dehydrogenase
SCO4855	MudPIT (M145)	Putative succinate dehydrogenase iron-sulfur subunit
SCO4856	MudPIT *(~tatC)*	Putative succinate dehydrogenase flavoprotein subunit
SCO4860	MudPIT (mutual)	Putative secreted hydrolase
SCO4907	2D	Transcriptional regulatory protein
SCO4921	MudPIT (M145)	Putative acyl-CoA carboxylase complex A subunit
SCO4932	MudPIT *(~tatC)*	Histidine ammonia-lyase
SCO4956	MudPIT (mutual)	Putative peptide methionine sulfoxide reductase
SCO4958†	2D and MudPIT (M145)	Cystathionine gamma-synthase
SCO4990	MudPIT (M145)	Conserved hypothetical protein
SCO5031	MudPIT (mutual)	Alkyl hydroperoxide reductase system hypothetical protein
SCO5042	MudPIT *(~tatC)*	Fumarate hydratase C
SCO5059	2D	Polyphosphate glucokinase
SCO5075	2D	Putative oxidoreductase
SCO5077	MudPIT (mutual)	Hypothetical protein
SCO5079	2D	Conserved hypothetical protein
SCO5101	MudPIT (M145)	Conserved hypothetical protein

SCO5102	MudPIT (mutual)	Putative mutT-like protein
SCO5135	2D and MudPIT (mutual)	Ferredoxin
SCO5145	MudPIT (mutual)	Conserved hypothetical protein
SCO5167	2D	Conserved hypothetical protein
SCO5169	MudPIT (mutual)	Putative ATP-binding protein
SCO5187	2D and MudPIT (mutual)	Putative glutaredoxin-like protein
SCO5199	MudPIT *(~tatC)*	Conserved hypothetical protein
SCO5212†	MudPIT (mutual)	3-phosphoshikimate 1-carboxyvinyltransferase
SCO5217	MudPIT *(~tatC)*	Anti-sigma factor
SCO5221	2D	Putative polypeptide deformylase
SCO5236	MudPIT (M145)	Putative glucosamine phosphate isomerase
SCO5249	2D	Putative nucleotide-binding protein
SCO5251	2D	Acetyltransferase
SCO5254	2D	Superoxide dismutase
SCO5261	MudPIT (M145)	Putative malate oxidoreductase
SCO5262†	MudPIT *(~tatC)*	Putative dehydrogenase
SCO5281	MudPIT *(~tatC)*	Putative 2-oxoglutarate dehydrogenase
SCO5359	MudPIT (mutual)	50S ribosomal protein L31
SCO5362	MudPIT *(~tatC)*	Conserved hypothetical protein
SCO5370	2D	ATP synthase delta chain
SCO5371	2D	ATP synthase alpha chain
SCO5373	MudPIT *(~tatC)*	ATP synthase beta chain
SCO5374	MudPIT *(~tatC)*	ATP synthase epsilon chain
SCO5396	2D	Putative cellulose-binding protein
SCO5398	MudPIT (M145)	Conserved hypothetical protein
SCO5405	2D	Putative transcriptional regulator

SCO5444	MudPIT *(~tatC)*	Putative glycogen phosphorylase
SCO5464	MudPIT (mutual)	Possible calcium binding protein
SCO5465	MudPIT (mutual)	Conserved hypothetical protein
SCO5470	MudPIT (mutual)	Serine hydroxymethyltransferase
SCO5490	MudPIT *(~tatC)*	Conserved hypothetical protein
SCO5514	MudPIT (mutual)	Acetolactate synthase small subunit
SCO5515	MudPIT (mutual)	Probable D-3-phosphoglycerate dehydrogenase
SCO5522	2D and MudPIT (M145)	3-isopropylmalate dehydrogenase
SCO5545	MudPIT (mutual)	Hypothetical protein SC1C2.26
SCO5547	2D	Glutamyl-tRNA synthetase
SCO5553	MudPIT *(~tatC)*	3-isopropylmalate dehydratase large subunit
SCO5561	MudPIT (mutual)	Hypothetical protein SC7A1.05c
SCO5564	MudPIT (M145)	Putative 50S ribosomal protein L28
SCO5570	2D and MudPIT (M145)	Hypothetical protein
SCO5571	MudPIT *(~tatC)*	50S ribosomal protein L32
SCO5584	2D and MudPIT (M145)	Nitrogen regulatory protein P-II
SCO5591	MudPIT (mutual)	30S ribosomal protein S16
SCO5607	MudPIT *(~tatC)*	Putative transcriptional regulator

SCO5624	MudPIT (M145)	30S ribosomal protein S2
SCO5651	MudPIT (M145)	Hypothetical protein SC6A9.16
SCO5655	2D	Putative aminotransferase
SCO5699	2D	Prolyl tRNA synthetase
SCO5703	2D	Hypothetical protein SC5H4.27
SCO5723	MudPIT *(~tatC)*	Putative regulator, BldB
SCO5736	MudPIT (mutual)	30S ribosomal protein S15
SCO5737†	2D and MudPIT (mutual)	Guanosine pentaphosphate synthetase/polyribonucleotide nucleotidyltransferase
SCO5744	2D	Putative dihydrodipicolinate synthase
SCO5745	2D	Conserved hypothetical protein SC9A10.09
SCO5759	MudPIT *(~tatC)*	Hypothetical protein SC7C7.14
SCO5777	MudPIT (M145)	Glutamate uptake system ATP-binding protein
SCO5806	MudPIT *(~tatC)*	Conserved hypothetical protein SC4H2.27c
SCO5838	MudPIT *(~tatC)*	Putative protease
SCO5841	MudPIT (mutual)	Phosphocarrier protein hpr
SCO5843	2D	Hypothetical protein SC9B10.10
SCO5853	2D	Conserved hypothetical protein SC9B 10.20c
SCO5862	2D	Two-component regulator CutR

SCO5971	MudPIT *(~tatC)*	Conserved hypothetical protein
SCO5976	MudPIT (M145)	Ornithine carbamoyltransferase
SCO5984	2D	Putative acyl-CoA dehydrogenase
SCO5999†	2D and MudPIT (mutual)	Aconitase
SCO6031	MudPIT (mutual)	Uroporphyrinogen decarboxylase
SCO6042	MudPIT *(~tatC)*	Conserved hypothetical protein SC1B5.02
SCO6046	MudPIT (M145)	Hypothetical protein
SCO6061	MudPIT (M145)	Putative oxidoreductase
SCO6084	2D	Putative DNA polymerase
SCO6145	2D	Hypothetical protein SC1A9.09
SCO6176	2D and MudPIT (mutual)	Conserved hypothetical protein
SCO6206	2D	Putative oxidoreductase
SCO6283	MudPIT (M145)	Conserved hypothetical protein
SCO6324	2D	Putative hydrolase
SCO6409	MudPIT (mutual)	Methionine aminopeptidase
SCO6412	2D	Putative aminotransferase
SCO6415	2D	Putative D-hydantoinase
SCO6482	2D and MudPIT (mutual)	Conserved hypothetical protein
SCO6488	MudPIT (M145)	Putative acyl-peptide hydrolase
SCO6531	2D and MudPIT (mutual)	Putative ATP/GTP binding protein
SCO6549	MudPIT (mutual)	Hypothetical protein SC5C7.34
SCO6551	MudPIT (mutual)	Putative oxidoreductase
SCO6606	MudPIT (M145)	Conserved hypothetical protein SC1F2.03c
SCO6650	2D	Hypothetical protein SC4G2.24
SCO6658	MudPIT (M145)	6-phosphogluconate dehydrogenase
SCO6663	MudPIT (M145)	Transketolase B
SCO6753	2D	Putative nucleotide sugar-1-phosphate transferase

SCO6776	2D	Conserved hypothetical protein
SCO6928	2D and MudPIT *(~tatC)*	Putative O-methyltransferase.
SCO7000†	MudPIT (mutual)	Isocitrate dehydrogenase
SCO7072	MudPIT (M145)	Conserved hypothetical protein
SCO7271	MudPIT *(~tatC)*	Putative ion channel subunit
SCO7286	MudPIT *(~tatC)*	Putative oxidoreducatse
SCO7293	2D	Conserved hypothetical protein
SCO7481	2D and MudPIT (mutual)	Conserved hypothetical protein
SCO7510†	2D and MudPIT (mutual)	Peptidyl-prolyl cis-trans isomerase
SCO7511	2D	Glyceraldehyde 3-phosphate dehydrogenase
SCO7647	MudPIT *(~tatC)*	Putative calcium binding protein
SCO7654	MudPIT (M145)	Putative oxidoreductase
SCO7676	MudPIT (mutual)	Putative ferredoxin
SCO7683	MudPIT (M145)	Putative non-ribosomal peptide synthase
SCO7699	2D	Putative nucleotide-binding protein
SCO7700	2D	Putative cyclase
SCO7728	2D and MudPIT *(~tatC)*	Conserved hypothetical protein
SCO7730	MudPIT (mutual)	Conserved hypothetical protein

The table lists proteins by SCO number. Method of observation indicates which technique was used to identify a particular protein. 2D is two-dimensional gel electrophoresis; MudPIT is multidimensional protein identification technology. Bracketed after MudPIT is indicated *"tatC'* if the protein was found in the cell-wall fraction of the *tatC* mutant strain only, 'M145' if the protein was found in the cell-wall fraction of the M145 strain only and 'mutual' if it was found in the cell-wall fractions of both the M145 and the *tatC* mutant strains.

†These proteins have also been identified in the extracellular fraction of *S. coelicolor* grown in liquid medium (Kim DW, Chater K, Lee KJ, Hesketh, A (2005) *J Bacteriol* 187: 2957-2966).

Example 6

A Tat Transport Assay Based on Agarase

To identify bona-fide Tat-targeting signal peptides associated with the group of 43 proteins identified by the proteomic assays 2-DGE and MudPIT, a reporter-based assay for Tat transport was designed.

Sec-dependent signal peptides are not normally recognized by the Tat machinery, and Tat-targeted proteins usually are folded and, therefore, are not normally compatible for Sec-dependent export. Thus, a reporter-based assay for Tat transport was designed to address directly whether the group of 43 putative Tat substrates were indeed synthesized with bona fide Tat signal peptides.

Embedding of *S. coelicolor* colonies into agar is a phenomenon associated with secretion of the enzyme agarase, which degrades agar to smaller oligosaccharides. As the agar is broken down around the colony, this causes the colony to sink down into the medium. Agarase is encoded by the *dagA* gene, and the protein product bears an N-terminal signal peptide containing an apparent twin-arginine motif MVNRRDLIKWSAVALGAGAGLAGPAPAAHA↓AD (SEQ ID NO: 313) (Figure 4A).
To determine whether agarase activity could be used in a reporter assay to test the ability of a signal peptide to secrete a protein via the TAT pathway, Applicants tested whether agarase is a TAT-dependent substrate.

a) First, extracellular agarase activity of wild type M145 *S. coelicor* was compared to that in a *tatC* mutant (TP1 (Δ*tatC*::Apra^R).

S. coelicor strains M145 and TP1 were grown on MM-C minimal medium for 5 days and were stained with lugol solution.

First, it was seen that colonies of the S. coelicor tat mutants failed to sink into solid media suggesting that DagA is a TAT substrate. Subsequent staining of the agar-containing plates with lugol, an iodine based reagent, showed a clear halo corresponding to the degradation of agar around the wild type strain (M145), while no zone of clearing was observed around colonies of the *tat* mutant strains (Fig. 4B).

These data provide strong and visual evidence that that the *tat* mutants have no significant extracellular agarase activity.

b) Second, to ensure that agarase activity was expressed in the *tat* mutants, the *dagA* gene was placed under control of the *tipA* thistrepton-inducible promoter and incorporated onto the chromosome in a single copy at the ϕ C31 attachment site. After integration of the construct into S. coelicolor strain M145 (wild type) and TP4 (Δ*tatC*), harboring pIJ6902-*dagA* in a single copy, thiostrepton was added to induce expression of

agarase. Transcription of *dagA* in *S. coelicolor* is highly regulated and reported to be repressed by glucose in the growth medium (serving-Gonzalez et al., (199) Microbiology 140:2555-2565). The bacteria were grown on mimimal medium containing 0.5% glucose, apramycin and thiostrepton for 3 days before being stained with lugol solution.

As shown in Fig. 4C, agarase activity was produced in the wild type strain M145, while no active agarase was secreted in the TP4 (*ΔtatC*) background.

These data demonstrate that *S. coelicolor* DagA is a TAT-dependent extracellular protein.

c) Third, to exploit agarase as a screen for Tat-dependent protein export the ability of the DagA mature protein to be targeted via alternative export routes was tested.

The TAT targeting signal peptide of DagA was swapped for the Sec-targeting signal peptides of three *S. coelicolor* proteins: SCO3053, SCO5660, and SCO6199, and the agarase activity determined as described above.

No extracellular agarase activity could be detected when the mature portion of the enzyme was targeted to the Sec machinery (Fig. 5; see Fig 6), most likely as a result of folding of agarase before export.

Taken together, these data show that DagA can be used as a reporter exclusively for TAT-mediated protein secretion.

Example 7

Idenfication of bona fide TAT-exported proteins and signal peptides using the TAT Transport Assay Based on Agarase.

The two proteomic techniques above identified putative (a total of 43) proteins that could have been transported by the Tat pathway (Table 3). To identify whether these proteins contained bona fide Tat-targeting signal peptides, the reporter-transport assay based on agarase activity was used. Fusions of the signal peptides of each of the putative TAT-targeted proteins to mature agarase were constructed and expressed in a nonagarase-producing host strain (Streptomyces lividans 1326) and assessed for the production of agarase activity using the lugol-based plate test.

Tat dependency of *Streptomyces coelicolor* agarase was demonstrated by growing strain M145 and TP4 harboring either pIL6902 or pIJ6902dagA on minimal medium containing glucose, which strongly represses expression of the native agarase. Approximately 10^4 spores of each strain were streaked or spotted onto minimal medium containing glucose, additionally supplemented with apramycin and thiostrepton (to induce expression of the myc-tagged agarase) and grown at 37°C for 72 hours. And then stained

with Lugol solution (Sigma) for 45 min. Plasmids encoding signal peptide–agarase fusion proteins (listed in Table 2) were mated into *S. lividans tat*$^+$ 1326 and 10-164 (as 1326, *msiK*$^-$), and the 10-164 isogenic *tatC* mutant (2Faury et al., (2004) Biochem Biophys Acta 1699:155-162), all of which do not possess native agarase. Plates containing mature spores of the resultant strains, along with *S. coelicolor* M145 and TP4 were used to inoculate, with a dissecting needle, MM-C minimal Medium (MM-C) (10g Agar, 1g $(NH_4)_2SO_4$, 0.5g K_2HPO, 0.2g $FeSO_4.7H_2O$ in 1L) lacking a carbon source other than agar. The inoculated plates were grown for 5 days at 30C and then stained with Lugol solution (Sigma) for 45 min before photography on a light box. Relative activities were estimated by determining the diameter of the zones of clearing using the measure tool of the image manipulation program GIMP (open source software distributed under a GNU general public license at <www.gimp.org>). One sample was included in every batch of results as a standard to which the diameters of the zones of clearing were adjusted. Each zone of clearing was measured twice with each measurement being at right angles to the other. The mean of the two measurements was taken and used to calculate R^2 (R = radius) as an estimate of the concentration of agarase. All measurements were then expressed as a percentage of the activity of pTDW47, which is the reporter pTDW46 with the native agarase signal peptide reintroduced. Nine measurements were made for each data point and the standard error of the mean was calculated for each point.

(a) Bona fide signal peptides:

In all, 25 of the 43 signal peptides clearly directed Tat-dependent secretion of DagA, consistent with their being bona fide Tat signal peptides (Fig. 5 and Table 6; a representative sample of these also was tested in a *S. lividans tatC* strain (Faury et al., *supra*), further demonstrating their Tat dependence; Fig. 6). Of these 25, TATFIND predicted 20 and TatP predicted 18 to be Tat signal peptides. Both prediction programs failed to identify two of the Tat-active signal peptides (SCO2286 and SCO3790) because of their unusual length; however, *in silico* N-terminal truncation of these signal peptides restored the ability of both programs to recognize them as Tat-targeting signal peptides. Moreover, it is clear when comparing the results from the agarase assays (Figs. 5 and 6) that secretion efficiency varies significantly with signal peptide primary structure. Thus, comparisons of heterologous protein translocation by using only a single Tat or Sec signal peptide, as recently performed for hTNFalpha and hIL10 in *S. lividans* (Schaerlaekens et al., (2004) J Biotechnol 112:279-288), may not reflect the true potential of each pathway for protein transport. The remaining 18 signal peptides of the 43 that contained consecutive arginine residues could not mediate secretion of active agarase, and of these, TATFIND

predicted 0 and TatP predicted only 1 to be a genuine Tat-targeting signal peptide (Table 3). These 18 signal peptides therefore were considered not to be Tat-targeting signals and the corresponding passenger proteins are not detected in the cell wall fraction of the Δ*tatC* strain presumably because of pleiotropic effects.

Of the exported proteins identified in cell wall analyses of the Δ*tatC* strain (Table 4), two (SCO0736 and SCO7677) were predicted by both TATFIND and TatP programs to contain Tat-targeting signal peptides. These signal peptides were subjected to the agarase test, and both conferred extracellular agarase activity to the *S. lividans* host strain, bringing the total of verified Tat-targeting signal peptides to 27 (Fig. 5). The apparent presence of these final two proteins in the Δ *tatC* strain cell wall might have been due to minor contamination of washes because of cell lysis. Finally, rare examples of Tat-dependent signal peptides lacking consecutive arginine residues have been reported (e.g., Ignatova et al., (2002) 291:146-149). A total of seven exported proteins were identified in the cell wall fraction of M145 that did not contain obvious twin-arginine motifs in their signal peptides (Table 3). The agarase test was applied to six of these seven signal peptides, but none were found to direct export of active agarase.

(b) TAT-exported proteins:

Taken altogether, Applicants have unambiguously identified 27 Tat-dependent proteins in *S. coelicolor*. This number represents 30% of all of the exported proteins that we detected in the cell wall fraction of the *tat*$^+$ strain and clearly demonstrates that the Tat pathway is a major protein translocation route. The agarase reporter assay developed here and used to validate possible Tat-targeting signal peptides is particularly powerful because it is not only facile and rapid but also semiquantitative and, therefore, provides a ready measure of transport efficiency for a particular Tat-targeting signal peptide. It is also anticipated that the agarase reporter system will facilitate the exploitation of the Tat pathway for heterologous protein production. The identified Tat-exported proteins listed in Table 6 represent a broad spectrum of functional classes: several are predicted to be involved in phosphate and carbohydrate metabolism, nutrient transport, and lipid metabolism. However, unlike the well characterized *E. coli* system, only 3 of the 27 Tat substrates detected here are likely to be cofactor-containing and, remarkably, 2 of these (SCO6272 and SCO6281) are associated with a type I modular polyketide synthase gene luster, indicating a role in secondary metabolism and, therefore, are not expected *a priori* to be exported. Substrates of the twin-arginine translocation pathway have hitherto not been found associated with a secondary metabolite gene cluster. Other notable Tat substrates identified here include two proteins involved in peptidoglycan metabolism, SCO0736 and

SCO1172, the latter being a probable cell wall amidase. Considering the connectivity between nonexport of amidases and the integrity of the cell envelope in *E. coli tat* mutants (Ize et al., (2003) Mol Microbiol 48:1183-1193), the identification of SCO1172 as a Tat substrate may well account for the fragility of the *Streptomyces tat* mutants. Moreover, as with the *E. coli* model, it may perhaps be possible to "rescue" the fragile phenotype of the *Streptomyces tat* mutants by increased expression of a Sec-dependent amidase. This likely would enable a further proteomic analysis of the "Tat secretome" in this organism and the assignment of Tat-targeted proteins not associated with the cell wall. A major surprise was that 4 of the 27 Tat substrates are annotated as lipoproteins (Bendtsen et al., (2005) BMC bioinformatics 6:167-175) (Table 6). Again, this *in vivo* data demonstrates the Tat dependence of putative lipoproteins and implies that class I and class II cellular signal peptidases can recognize and cleave Tat-targeting signal peptides. Interestingly, an outer membrane-localized dimethyl sulfoxide reductase, a Tat substrate that is also strongly predicted to be a lipoprotein, has been described in *Shewanella oneidensis* (Gralnick et al., (2006) Proc Natl Acad Sci USA 103:4669-4674).

Overall, these data demonstrate that the Tat pathway can export a diverse range of proteins and further reinforces the contention that the Tat pathway is a truly general protein export system in this organism. The ability of the *Streptomyces* Tat pathway to export proteins requiring various anti- and possibly posttranslocation modifications, as well as two of the largest single Tat substrates ever reported (SCO6198 and SCO6457, at 116 and 146 kDa, respectively), underscores the potential of the Tat system in Streptomycetes for heterologous protein production.

TABLE 6

Tat-dependent proteins identified in *S. coelicolor*

Protein	Description
SCO0736	Putative secreted protein
SCO1172	Putative amidase (putative secreted protein)
SCO1196	Putative secreted protein
SCO1356*†	Putative iron sulfur protein (putative secreted protein)
SCO1432	Putative membrane protein
SCO1565	Putative glycerophosphoryl diester

	phosphodiesterase
SCO1590	Putative secreted protein (transglycosidase)
SCO1639†	Putative secreted peptidyl-prolyl *cis-trans* isomerase protein
SCO1906	Putative secreted protein (PhoX phosphatase)
SCO2068	Putative secreted alkaline phosphatase
SCO2286	Putative alkaline phosphatase
SCO2758	β-*N*-acetylglucosaminidase (putative secreted protein)
SCO2780†	Putative secreted protein (cd01146, FhuD, Fe3_-siderophore binding domain FhuD)
SCO2786	β -*N*-acetylhexosaminidase (Transglycosidase)
SCO3484†	Putative secreted sugar-binding protein
SCO3790	Conserved hypothetical protein of PhoX type
SCO4672	Putative secreted protein
SCO6052	Putative membrane protein
SCO6198	Putative secreted protein
SCO6272*	Putative secreted FAD-binding protein
SCO6281*	Putative FAD-binding protein
SCO6457	Putative _-galactosidase (Galactose mutarotase-like)
SCO6580	Hypothetical protein (Xylose isomerase-like)
SCO6594	Putative secreted protein
SCO6691	Putative phospholipase C
SCO7631	Putative secreted protein
SCO7677	Putative secreted solute-binding protein

*Cofactor containing.
†Putative lipoproteins.

What is claimed is:

1. An isolated TAT signal peptide comprising the sequence motif (X^{-1})RR(X^{+2})(X^{+3})(X^{+4}), wherein R is arginine, X^{-1} is amino acid M, H, A, P, K, R, N, T, G, S, D, Q or E; X^{+2} is amino acid A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is Q, I, L, V, M or F, and wherein said motif is not within the first 35N' terminal residues of the amino acid sequence of the polypeptide.

2. An isolated TAT signal peptide comprising the sequence motif (X^{-1})RR(X^{+2})(X^{+3})(X^{+4}), wherein R is arginine, X^{-1} is amino acid H, A, P, K, R, N, T, G, S, D, Q E or L; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is T, G or A, and wherein the motif is within the first 35N' terminal residues of the amino acid sequence of the polypeptide.

3. The TAT signal peptide of Claim 2, wherein when X^{-1} is H then X^{+4} is A.

4. The TAT signal peptide of Claim 2, wherein when X^{-1} is L then X^{+4} is G.

5. An isolated TAT signal peptide comprising the sequence motif (X^{-1})RR(X^{+2})(X^{+3})(X^{+4}), wherein RR represents two adjacent arginine residue and X designates positions restrict to other selected amino acids: X^{-1} is M, H, A, P, K, R, N, T, G, S, D, Q, or E; X^{+2} is a polar amino acid residue; and X^{+3} and X^{+4} are non-polar amino acid residues, and wherein the motif is not within the first 35N' terminal residues of the amino acid sequence of the polypeptide.

6. An isolated variant of the TAT signal peptide of Claim 5.

7. An isolated TAT signal peptide comprising the amino acid sequence of the signal peptide of proteins SCO2286 (SEQ ID NO: 218), SCO3790long (SEQ ID NO: 227), SCO6580long (SEQ ID NO:241), SCO1590 (SEQ ID NO: 211), SCO1824 (SEQ ID NO: 213), SCO6580short (SEQ ID NO: 182), or SCO3790short (SEQ ID NO: 122).

8. An isolated polynucleotide comprising a polynucleotide sequence encoding a signal peptide of Claims 1, 2, 5, or 6.

9. The isolated polynucleotide of Claim 8, wherein said nucleotide sequence encoding a TAT signal polypeptide is operably linked to a second nucleotide sequence encoding a heterologous polypeptide.

10. An expression vector comprising a first nucleotide sequence encoding a TAT signal polypeptide of Claims 1, 2, or 5 operably linked to a second nucleotide sequence encoding a heterologous polypeptide.

11. A fusion polypeptide comprising a TAT signal peptide of Claim 1, 2, or 5 and a heterologous polypeptide.

12. The fusion polypeptide of Claim 11, wherein said TAT signal peptide is the secretory leader sequence of polypeptides that are naturally expressed by *Streptomyces.*

13. The fusion polypeptide of Claim 11, wherein said heterologous polypeptide is an enzyme, a growth factor or a hormone.

14. The fusion polypeptide of Claim 13, wherein said enzyme is a protease, a carbohydrase, an isomerase, a glucoamylase, a kinase, an amidase, an esterase, or an oxidase.

15. The fusion polypeptide of Claim 11, wherein said heterologous polypeptide is not naturally associated with a secretion signal peptide.

16. A bacterial host cell genetically transformed with an expression vector of Claim 10.

17. A method for producing a heterologous polypeptide comprising:

(a) culturing host cells comprising an expression vector comprising a first nucleotide sequence encoding a TAT signal peptide operatively linked to a second nucleotide sequence encoding a heterologous polypeptide,

wherein said TAT signal peptide comprises the sequence motif $(X^{-1})RR(X^{+2})(X^{+3})(X^{+4})$, wherein R is arginine, X^{-1} is amino acid M, H, A, P, K, R, N, T, G, S, D, Q or E; X^{+2} is amino acid A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is Q, I, L, V, M or F, and wherein said motif is not within the first 35N' terminal residues of the amino acid sequence of the polypeptide; and

(b) producing said heterologous polypeptide.

18. A method for producing a heterologous polypeptide comprising:

(a) culturing host cells comprising an expression vector comprising a first nucleotide sequence encoding a TAT signal peptide operatively linked to a second nucleotide sequence encoding a heterologous polypeptide,

wherein said TAT signal peptide comprises the sequence motif $(X^{-1})RR(X^{+2})(X^{+3})(X^{+4})$, wherein R is arginine, X^{-1} is amino acid H, A, P, K, R, N, T, G, S, D, Q E or L; X^{+2} is A, P, K, R, N, T, G, S, D, Q or E; X^{+3} is I, W, F, L, V, Y, M, C, H, A, P, N or T; and X^{+4} is T, G or A, and wherein the motif is within the first 35N' terminal residues of the amino acid sequence of the polypeptide; and

(b) producing said heterologous polypeptide.

19. The method of Claims 17 or 18, wherein said step of producing comprises recovering said heterologous polypeptide from the culture medium.

20. The method of Claim 17 or 18, wherein said host cells are prokaryotic cells.

21. The method of Claim 17 or 18, wherein said host cells are Streptomyces bacterial cells.

22. The method of Claim 17 or 18, wherein said host cells are *S. coelicolor* or *S. lividans* bacterial cells.

23. The method of Claim 17 or 18, wherein said heterologous polypeptide is an enzyme, a growth factor or a hormone.

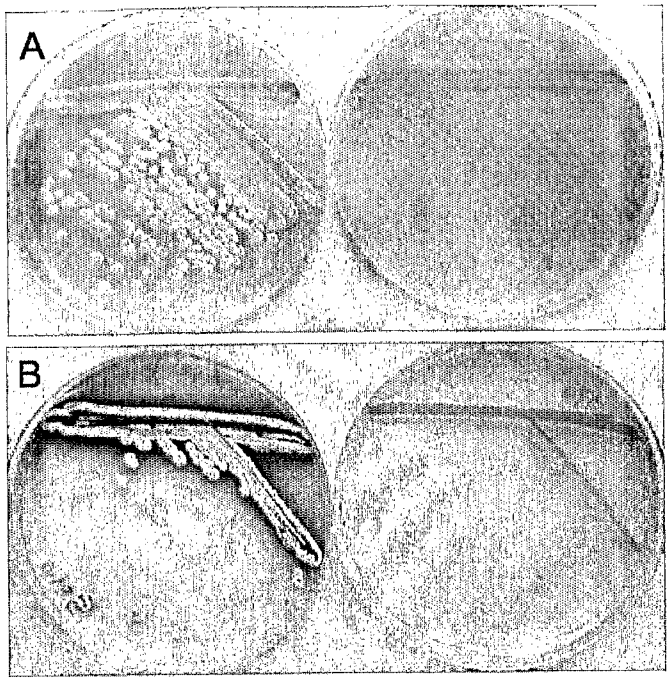


Figure 1

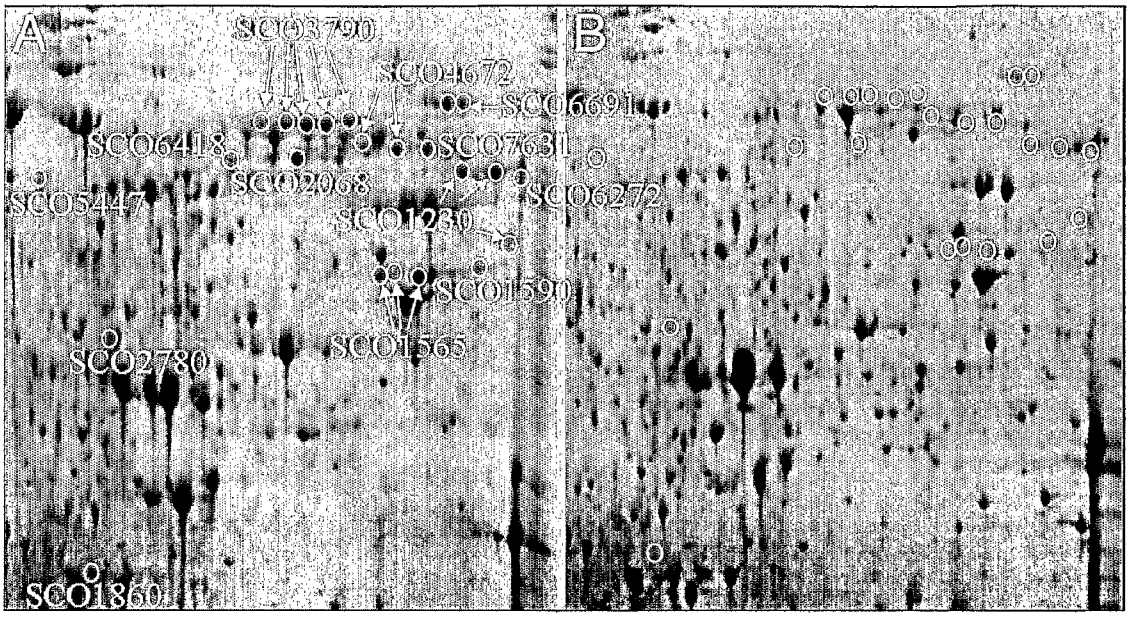

A
SCO3790
SCO4672
SCO6418
SCO7631
SCO2068
SCO5447
SCO1230
SCO1590
SCO2780
SCO1565
SCO1860
B
SCO6691
SCO6272

Figure 2

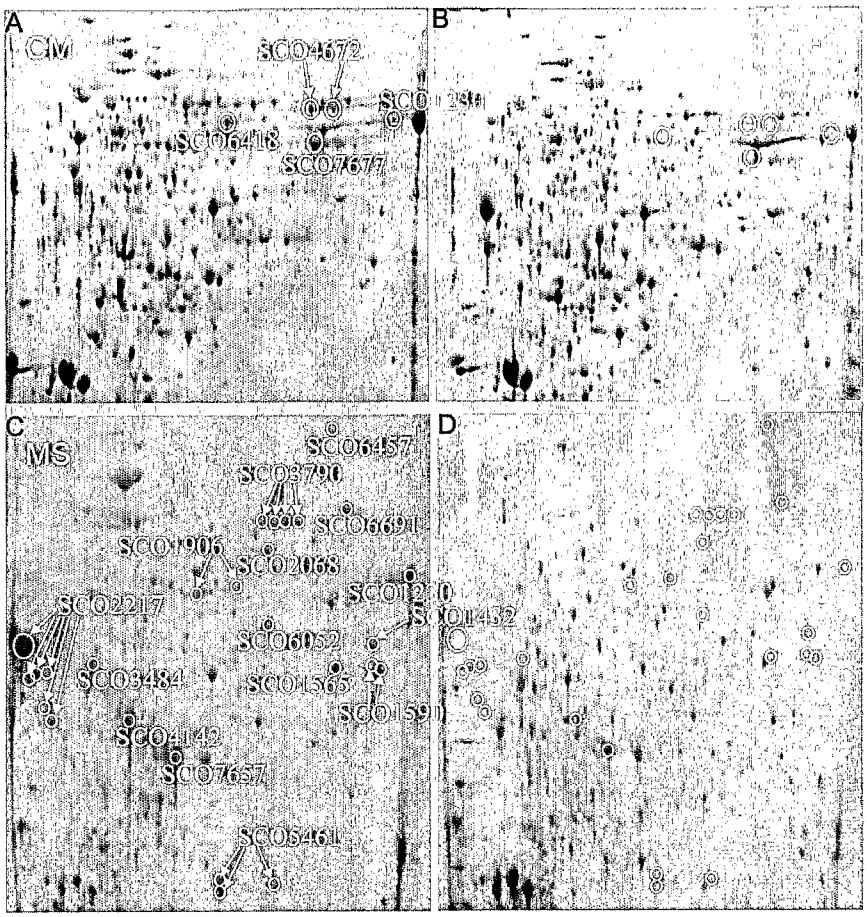

A
CM
SCO4672
SCO
SCO6418
SCO7677
B
C
MS
SCO6457
SCO3790
SCO6691
SCO1906
SCO2068
SCO2217
SCO6052
SCO3484
SCO1565
SCO4142
SCO7657
SCO5461
D
SCO1432

Figure 3

A) MVNRRDLIKWSAVALGAGAGLAGPAPAAHAIAD

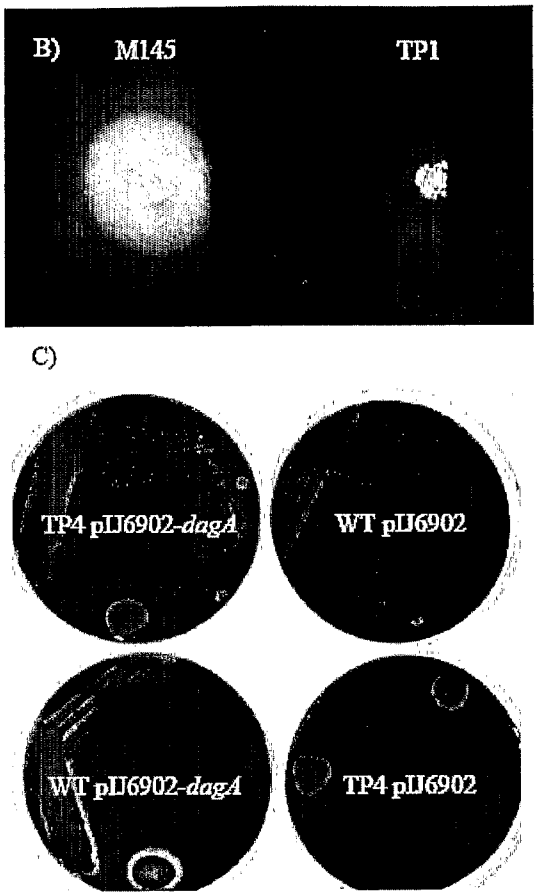


Figure 4

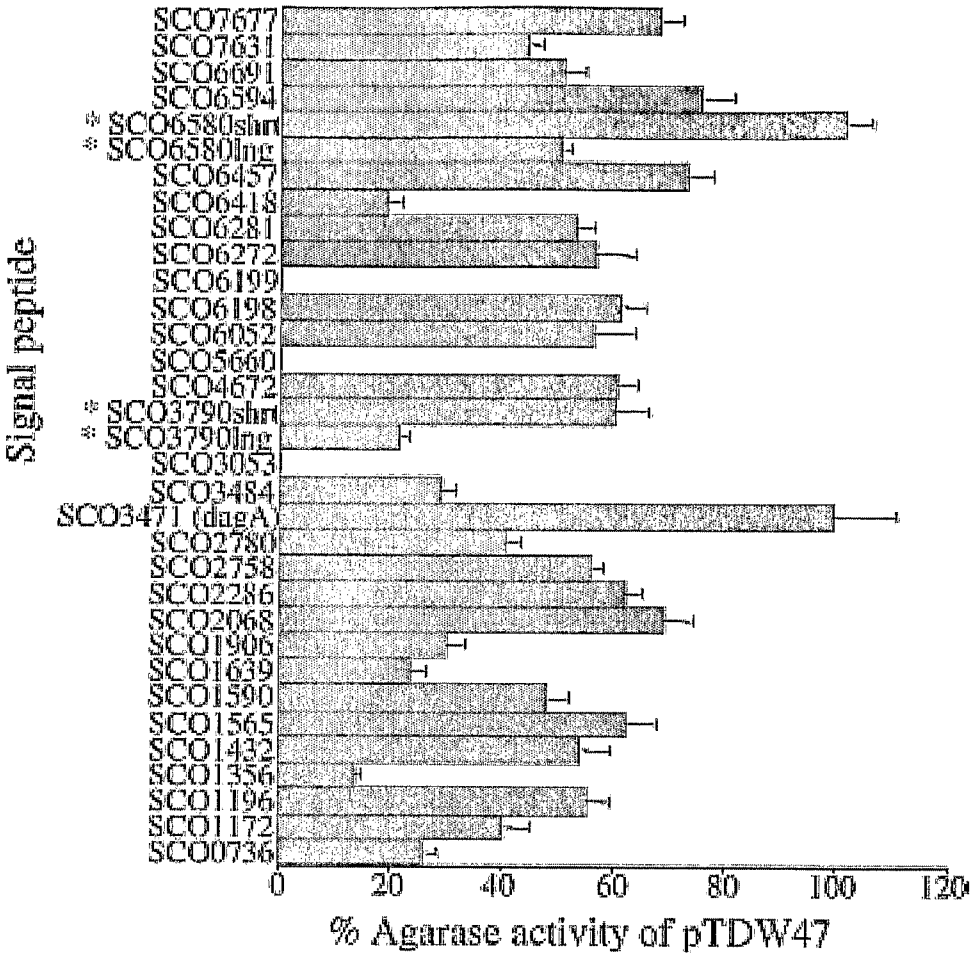

SCO7677
SCO7631
SCO6691
SCO6594
* SCO6580shrt
* SCO6580lng
SCO6457
SCO6418
SCO6281
SCO6272
SCO6199
SCO6198
SCO6052
SCO5660
SCO4672
* SCO3790shrt
* SCO3790lng
SCO3053
SCO3484
SCO3471 (dagA)
SCO2780
SCO2758
SCO2286
SCO2068
SCO1906
SCO1639
SCO1590
SCO1565
SCO1432
SCO1356
SCO1196
SCO1172
SCO0736
Signal peptide
0
20
40
60
80
100
120
% Agarase activity of pTDW47

Figure 5

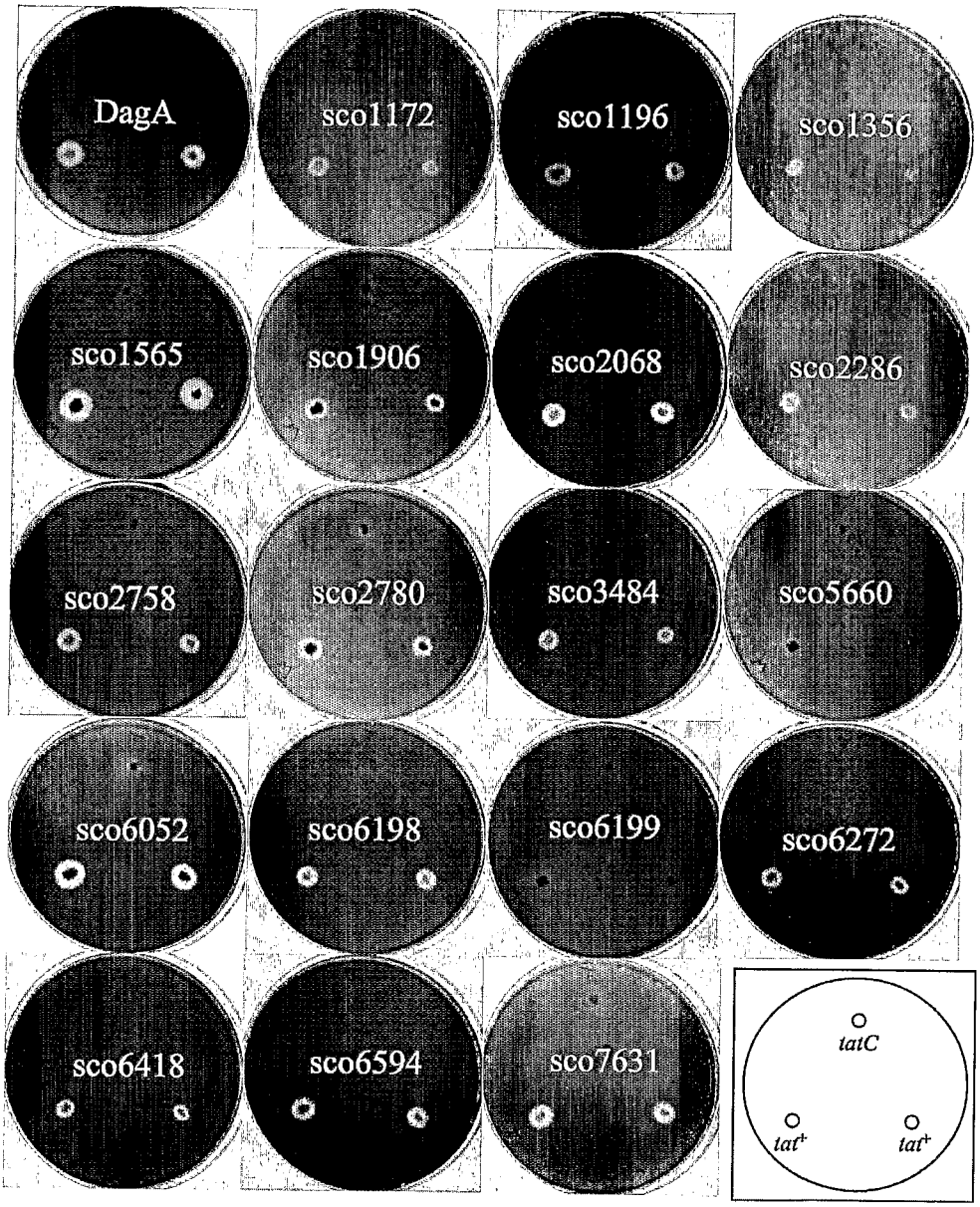

DagA
sco1172
sco1196
sco1356
sco1565
sco1906
sco2068
sco2286
sco2758
sco2780
sco3484
sco5660
sco6052
sco6198
sco6199
sco6272
sco6418
sco6594
sco7631
tatC
tat+
tat+

Figure 6