MBE Advance Access published June 14, 2011

Functional Diversification of the Twin-Arginine Translocation Pathway Mediates the Emergence of Novel Ecological Adaptations

Xiaowei Jiang[1] and **Mario A. Fares**[1,2]*

[1]Evolutionary Genetics and Bioinformatics Laboratory, Department of Genetics, Smurfit Institute of Genetics, University of Dublin, Trinity College Dublin, Dublin 2, Ireland

[2]Group of Integrative and Systems Biology, Instituto de Biología Molecular y Celular de Plantas (CSIC-Universidad Politécnica de Valencia), Valencia, Spain

*Corresponding author:

Dr. Mario A. Fares

Evolutionary Genetics and Bioinformatics Laboratory, Department of Genetics, Smurfit Institute of Genetics, University of Dublin, Trinity College Dublin, Dublin 2, Ireland

E-mail addresses: faresm@tcd.ie

Tel: +353-1-8963521

Abstract:

Microorganisms occupy a myriad of ecological niches that show an astonishing diversity. The molecular mechanisms underlying microbes' ecological diversity remain a fundamental conundrum in evolutionary biology. Evidence points to that the secretion of a particular set of proteins mediates microbes' interaction with the environment. Several systems are involved in this secretion, including the Sec secretion system and the Tat pathway. Shifts in the functions of proteins from the secretion systems may condition the set of secreted proteins and can, therefore, mediate adaptations to new ecological niches. In this manuscript we have investigated processes of functional divergence—a term used here to refer to the emergence of novel functions by the modification of ancestral ones—of Tat pathway proteins using a large set of microbes with different lifestyles. The application of a novel approach to identify functional divergence allowed us to distinguish molecular changes in the three Tat proteins among different groups of archaea and bacteria. We found these changes as well as the composition of secreted proteins to be correlated with differences in microbe's lifestyles. We identified major signatures of functional divergence in halophilic and thermophilic archaea as well as in pathogenic bacteria. The location of amino acids affected by functional divergence in functionally important domains of Tat proteins made it possible to find the link between the molecular changes in Tat, the set of secreted proteins and the environmental features of the microbes. We present evidence that links specific molecular changes in secretion mediating proteins of microbes to their ecological adaptations.

Introduction

Bacterial, archaeal and eukaryotic cells use two major secretion systems to transport proteins that are synthesised in the cytosol, namely the general secretion (Sec) pathway and the twin-arginine translocation (Tat) pathway. The Sec pathway transports proteins in an unfolded state, whereas the Sec-independent Tat pathway transports proteins that are partially or fully folded. The N-terminal signal peptide, as part of the substrate protein, plays a significant role in targeting proteins to the cytoplasmic membrane, where they can be integrated or alternatively transported across. This N-terminal peptide presents a tripartite configuration that includes an N-terminal region at the beginning, a hydrophobic region in the middle and a C-terminal region in the end. A typical Tat signal peptide has an arginine-arginine (twin-arginine) consensus motif in its N-terminal region (Yuan et al. 2010).

Two major differences exist in the mechanism of protein secretion of Sec-dependent and Tat systems. The Sec-dependent translocation pathway uses two modes of protein secretion: the co-translational translocation and the post-translational translocation. In the co-translational translocation pathway, which is conserved in both archaea and bacteria, proteins are translocated across or into the plasma membrane during protein synthesis. In contrast to this, in the post-translational translocation, proteins are first synthesized in the cytosol and stabilised in an unfolded state through the chaperonin activity of SecB and/or SecA for their subsequent translocation across the membrane (McFarland, Francetic, Kumamoto 1993; Eser, Ehrmann 2003). The translocation of proteins occurs through the pore formed by SecY protein and favoured by the ATPase activity of SecA. Importantly, SecB and/or SecA mediated translocation

pathway is only found in bacteria but not in archaea. The mechanism that archaea use to solve the problem of post-translational translocation—that is stabilising proteins in an unfolded or partially folded states for their subsequent translocation across the membrane, remains elusive (Irihimovitch, Eichler 2003).

The Sec-independent Tat pathway is conserved in bacteria, archaea, chloroplasts and plant mitochondria (Lee, Tullman-Ercek, Georgiou 2006), pointing to its essentiality in the viability of these organisms. It consists of three major components, TatA, TatB and TatC, all of which are present in most gram-negative bacteria. In contrast, most gram-positive bacteria and archaea have only two Tat components, TatA and TatC (Eijlander et al. 2009). Most of our knowledge regarding the mechanisms of substrate recognition and translocation by the Tat system is based on studies of the two model organisms, *Escherichia coli* and *Bacillus subtilis* (Panahandeh, Holzapfel, Mueller 2009). These studies show that Tat translocase uses Tat(B)C complex as a receptor site for binding Tat substrate proteins. After substrate binding, Tat(B)C recruits multiple TatA proteins to form a pore-containing protein complex that allows passing substrates through the pore driven by a proton motive force (Lee, Tullman-Ercek, Georgiou 2006; Tarry et al. 2009; Yuan et al. 2010). TatC was shown to determine the substrate specificity and thus plays a critical role in Tat-dependent protein secretion (Strauch, Georgiou 2007; Eijlander, Jongbloed, Kuipers 2009). What is the mechanism whereby Tat pathway monitors the folding fidelity of proteins for their correct translocation? In one study, authors showed that indeed TatA presents a proofreading function of its substrate FeS proteins NrfC and NapC, so that when such substrates are incorrectly folded they undergo rapid degradation (Matos, Robinson, Di Cola 2008).

Although the mechanism is not entirely clear, Tat pathway seems to mediate the emergence of new ecological adaptations in many bacteria and archaea. For example, recently an association between ecological adaptation and the Tat secretion pathway has been established in halophilic archaea *Halobacterium sp. NRC-1* (Rose et al. 2002). Studies showed that this group of archaea uses the Tat translocation pathway extensively to transport most of their secreted proteins (Dilks, Gimenez, Pohlschroder 2005). Haloarchaea live in an environment with "near-saturation" salt concentrations, a harsh condition for the proper function and secretion of proteins whose functions are important in later stages outside the cytosol. Interestingly, this correlates with the fact that most sequenced haloarchaea have two TatC homologs (TatCo and TatCt). In contrast, therefore, to many bacteria in which Tat transports a small fraction of proteins and are not essential for cell viability (Lee, Tullman-Ercek, Georgiou 2006), TatC may be essential to haloarchaea in buffering the environmental conditions.

Not only is Tat pathway a major route for protein secretion in haloarchaea, but also it is essential for cell viability (Dilks, Gimenez, Pohlschroder 2005; Thomas, Bolhuis 2006). Some other prokaryotes are known to use Tat pathway extensively, such as *Streptomyces* species (e.g. *S. coelicolor*) (Widdick et al. 2006). This difference in Tat essentiality between microbes seeks an explanation and can be crucial to understand ecological adaptation. What makes Tat essential and what molecular bases contribute to the different essentiality of Tat in different groups of organisms? Could we find molecular changes that are associated with environmental adaptations? Are there any other prokaryotes that present similar patterns in their Tat translocase? These questions remain largely unexplored.

Secretion systems play a significant role in prokaryotes environmental adaptation (Wooldridge 2009) and, as such, understanding the evolution of these systems is a crucial aim in evolutionary biology. Unfortunately, however, contrary to cytosolic proteins, a large amount of secreted proteins are structurally uncharacterized, posing difficulties to perform detailed evolutionary analyses (Elofsson, von Heijne 2007; Popot 2010). Alternative approaches are therefore required to circumvent the limitations to studying the evolution of these systems.

Here we hypothesise that the Tat pathway has diverged its function from the ancestral ones in microbes with particular ecological traits. As a case in point, we focus in two main types of microbes: halophilic archaea, that live in environments saturated with salt, and pathogenic bacteria, that may require the secretion of proteins to interfere with the immune response of the host. We use prokaryotic genome sequences, together with abundant secondary structure data, a new computational tool to identify divergence in protein functions and a wide range of experimental studies reporting crucial information to test this hypothesis.

Material and Methods

Sequences and genomes

Homologous protein sequences for TatA, TatB and TatC were retrieved using API (Application Program Interface) (http://www.genome.jp/kegg/soap/) service from KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database (Kanehisa et al. 2008). Genomes and protein table files were retrieved from NCBI (National Center for Biotechnology Information) Genome database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.ptt.tar.gz). Protein table files were later used for functionally annotating Tat-dependent substrates proteins with COG (Clusters of Orthologous Groups) (Tatusov et al. 2003). We also annotated all pathogenic species with their host (human, animal and plant) and the diseases they cause (ICD-10, International Statistical Classification of Diseases and Related Health Problems 10th Revision, http://apps.who.int/classifications/apps/icd/icd10online) wherever possible as this allowed us to classify bacteria as being pathogenic when appropriate. Amino acid sequences of all homologs were aligned using MUSCLE 3.7 (Edgar 2004) and alignments were manually checked using Jalview (Waterhouse et al. 2009). Gaps are removed in the functional divergence analysis.

Phylogeny reconstruction

Amino acid substitution model and model parameters were chosen using Prottest (Abascal, Zardoya, Posada 2005) for each of the Tat proteins: TatA, TatB and TatC. These parameters were then used by RaxML program (Stamatakis 2006) to estimate the best Maximum likelihood phylogenetic trees for the three proteins. The phylogenetic trees of TatA, TatB and TatC with

their corresponding protein sequence alignments were used as starting data for the analyses of functional divergence.

Analysis of Functional Divergence in Tat proteins

Functional divergence (FD) is a term that refers to a change in function as a result of an amino acid substitution event within a protein. Here we used a relaxed definition of FD to describe shifts in the ancestral function of a protein as a result of important amino acid substitutions in that protein—these shifts might not involve the emergence of a completely novel function but the modification of an existing one. FD can therefore refer to complete change in function after gene duplication or slight shifts in the function of a protein after speciation.

Previous studies have used the term FD to exclusively refer to shifts in the function of a protein after gene duplication—that is to say, FD between paralogs. Under this view, they classified FD into two main categories, type I and type II FD (Gu 1999). In type I FD amino acids are highly conserved at particular sites in one paralog while being highly stochastic in the other, indicating the acquisition of a function at these sites in one paralog (the conserved one) or the loss of the function in the other. In type II FD, amino acids are highly conserved in both paralogs at the same positions of the protein, indicating the divergence of two highly important functions between the paralogs.

Recently, a new method was implemented to indentify type I FD at the genome level between orthologs (Toft, Williams, Fares 2009) and was proved efficient in identifying conserved amino acid sites with FD in a large phylogenetic tree including orthologous as well as paralogous sequences (Williams et al. 2010). Briefly, the method uses a protein sequence alignment and a

phylogenetic tree including paralogs and orthologs. The method then compares a pair of clades in the tree sharing a common ancestral origin to their closest phylogenetic outgroup. The comparison is performed for each amino acid site in the protein and the strength or likelihood of the amino acid state is evaluated using the appropriate BLOSUM matrix (BLOcks of Amino Acid SUbstitution Matrix) (Henikoff, Henikoff 1992). BLOSUM matrices comprise the scores for the transition between the 20 amino acids, with positive scores meaning the transition is more frequent than expected, negative means the transition is less frequent than expected and 0 means these transitions are as frequent as expected. Using BLOSUM as a score matrix for the transitions between amino acids allows scoring both variable sites and conserved sites: one clade may seem variable yet the amino acid transitions between orthologous sequences within the clade may have taken place between biochemically close amino acids (positive scores). In the comparison between two clades to an outgroup (clade 1 and clade 2), functional divergence in clade 1 would be detected if the BLOSUM scores were positive within clade 1, negative between clade 1 and the outgroup and positive between clade 2 and the same outgroup. To test the significance of these score variances, we calculated the mean and standard error of the scores for clade 1 ($\bar{C}_1$ and SE_1) and clade 2 ($\bar{C}_2$ and SE_2) (Toft, Williams, Fares 2009). The significance of the difference in the means was obtained by calculating the Z-score for the comparison of the two clades as:

$$Z = \frac{\bar{C}_1 - \bar{C}_2}{SE_{C_1}} \quad (1)$$

Where $\overline{C}_{1,2}$ are the mean substitution scores for the transition from clades on either side of the bifurcation in the phylogenetic tree relative to the outgroup and SE_{C_1} is the standard error of clade 1.

Protein secondary structure prediction

Two approaches were used to predict protein secondary structures for TatC, TatA and TatB, respectively. We first used the SMART database to predict protein domains in TatC (Letunic, Doerks, Bork 2009). We then used the PSIPRED to predict protein structures of TatA and TatB (Bryson et al. 2005). The latter method is based on artificial neural network approaches and can reach accurate protein secondary structure predictions (http://bioinf.cs.ucl.ac.uk/index.php?id=779). Protein domains of *E. coli* TatC were predicted first and then mapped back to the protein sequence alignment to define protein domains for all the proteins in the alignment. These domains included three cytoplasmic domains (C1-C3), six transmembrane domains (T1-T6) and three periplasmic domains (P1-P3). Similarly, protein sequences of TatA and TatB from *E. coli* were used as reference sequences. N-terminal region, transmembrane domain, hinge region, the amphipathic helix and C-terminal domain were predicted for *E. coli* TatA and TatB, respectively. We then used the *E. coli* TatA and TatB as reference and mapped these structural regions in their multiple sequence alignments to define the protein domains for TatA and TatB. After FD analysis, we mapped amino acid sites under FD to these protein domains. To understand the relevance of the results of FD analyses, we built a two dimensional matrix, with the rows representing a clade under FD and the columns representing the number of sites under FD in each protein domain. Three such data matrices were made, one

per protein: TatC, TatA and TatB, respectively. We then used the heatmap.2 function in R (http://www.r-project.org) to cluster the rows and columns to find common patterns of FD among protein domains and clades in the matrix. The heatmap.2 function scales each element in a row in the data matrix based on a normalised Z-score, which is calculated as follows:

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_i} \tag{2}$$

Here, z_{ij} is the Z-score calculated for element x_{ij} in the matrix, $\bar{x}_i$ is the mean of the row i, σ_i is the standard deviation of all the elements in the row i.

Prediction of potential Sec and Tat substrate proteins

We used Tatp to predict potential Tat-dependent secreted proteins (secretome) (Bendtsen et al. 2005) and SignalP to predict Sec-dependent secretome (Emanuelsson et al. 2007). In predicting Sec-dependent substrates, we first used SignalP and then removed those predicted also by TatP to minimise false positives. Predicted substrates from Sec and Tat pathways were used to assess how different the two secretion systems, Tat and Sec, were and how much they explained about ecological adaptations in our organisms. We preferred using TatP instead of Tatfind (Rose et al. 2002) for the prediction of Tat-dependent secretome as the former approach is based on artificial neural networks that largely outperforms other approaches in pattern recognition tasks pertaining variant Tat signal peptides. In contrast, Tatfind is a regular expression based approach, which does not allow recognition of variant Tat signal peptides (Bendtsen et al. 2005).

Predicted Sec and Tat substrates were grouped according to their COG functional categories. To determine if the numbers of substrates for Tat-dependent secretion in each of the categories were significantly higher (enriched category for secreted substrates) or lower (impoverished category) than expected by chance, a χ^2 test was performed on each COG functional category with one degree of freedom. The statistical test was performed according to the following equation:

$$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

$$E_i = n_i \frac{N_{predicted}}{N_{COG}} \tag{4}$$

O_i is the observed number of predicted substrate proteins in COG category *i*. while E_i is the

expected number of such proteins in that category i . The total number of proteins in category i is indicated by n_i. $N_{predicted}$ is the total number of predicted substrate proteins, while N_{COG} is the total number of proteins in the proteome. Only proteins assigned within single COG functional categories were analysed.

Results and Discussion

Evidence for Functional Divergence in the Tat Translocation System

In this study we analysed Tat homologs (orthologs and paralogs) retrieved from 944 bacterial genomes, 68 archaeal genomes and 11 chloroplasts. We estimated the amount of functional divergence in each of the Tat proteins in the different lineages of the phylogenetic tree (see Material and Methods for details). Functional divergence (FD) refers here to the divergence of protein's function from its ancestral function by amino acid replacements in functionally important amino acid sites. We applied a method to identify functional divergence previously developed (Toft et al. 2009; Williams et al. 2010). Analysis of our tree allowed us to identify 204, 145 and 119 clades with at least one conserved amino acid site with evidence of FD for TatC, TatA and TatB, respectively (Supporting information: Table S1 to S3 and Figure S1 to S3). The number of clades under FD in each Tat phylogenetic tree (TatC: 204; TatA: 145; TatB: 119) seems to be correlated with TatC and TatA being the most important components in determining a minimum Tat translocase (Blaudeck et al. 2005; Panahandeh, Holzapfel, Mueller 2009). In particular, a mutant TatA can complement a strain with a deleted TatB without compromising substrates translocation (Blaudeck et al. 2005).

In total, we identified 167 bacterial pathogens (100% of the pathogenic bacteria in our data) and 8 haloarchaea (100% of the analysed haloarchaea) to be involved in a cluster with at least one site under FD. To understand the relationship between FD and ecological adaptation, we focused on the 5 most functionally divergent clades of the tree, that have at least 4 sequences in clade 1—that is to say, organism clades that presented the largest number of amino acid sites with

evidence of having diverged their functions from ancestral clades. We also show the relative phylogenetic positions of the clades under FD (Figure 1). Phylogenetic clades of halophilic archaea containing TatC homologs, TatCo and TatCt which resulted from a TatC gene duplication, have been identified to be the most functionally divergent of all clades (Figure 1A), namely the halophilic archaea have the greatest number of amino acid sites (32 sites for TatCo, TatCt has 24 sites) under FD compared with other lineages. This is interesting as haloarchaea had to adapt to an extreme environment and FD of TatC after duplication may have enabled such adaptations through the acquisition of novel functions or the modification of its original function. In contrast to TatC, thermophilic archaea *Sulfolobus* presented most of its FD signal in TatA (Figure 1B). Finally, *Corynebacteria* accumulated most of the functionally divergent changes in TatB (Figure 1C). Importantly, 8 out of the 11 clades with the strongest signal of FD included pathogenic bacterial strains (Figure 1), which suggests that Tat translocation pathway may play a role in bacterial pathogenesis. In summary, clades showing strong FD in Tat translocation system are unique as they include microorganisms living in extreme environments always challenged by ionic concentrations, temperature or host immune response (We discuss these results with detail in the following sections).

Strong functional divergence in TatC from Halophilic archaea

Haloarchaea presented substantial FD following the TatC gene duplication event that gave rise to TatCo (Table 1, amino acid sites under FD are mapped to Tat alignment, as shown in Figure S4) and TatCt (Table 2, amino acid sites under FD are mapped to Tat alignment, as shown in Figure S5) (Figure 1A). A close look to the distribution of amino acid sites with evidence of FD shows that the different proteins domain present significant differences in their content of FD (Figure

1A). In general, we could identify three main clusters of domains according to their content in functionally divergent amino acid sites (sites under functional divergence are shown in the supplementary Table S4 to S6). The first cluster consists of 2 protein domains (from left to right: P3 and C4, Figure 1A) that were impoverished for FD. The second cluster included 5 protein domains (from left to right: C3, T6, T5, T1 and T3, Figure 1A) that showed modest number of amino acid sites under FD. Finally, the third cluster comprised 6 protein domains (from left to right: C1, C2, P1, T2, T4 and P2, Figure 1A) and was highly enriched for FD.

Interestingly, halophilic archaea TatCo and TatCt proteins presented two different patterns of FD: TatCt showed strong FD in the second cluster of FD domains, affecting mostly TatC transmembrane domains while TatCo presented most of its FD in the N-terminal region of the protein (Figure 1A). Non-overlapping FD between the protein homologs resulting from the duplication of an ancestral gene is a strong indicator of sub-functionalization after gene duplication, in which the sum of functions of the two paralogs complements the ancestral function. We also found three pathogenic bacterial clades (*Leptospira*, *Bartonella* and *Rickettsia*) to be amongst the most functionally divergent ones (Supporting information: Table S4), with the third cluster (comprising protein domains C1, C2, P1, T2, T4 and P2, Figure 1A) being enriched for FD in these bacteria (Figure 1A). The N-terminal half of TatC has been shown to play an important role in precursor binding (Holzapfel et al. 2007). C1 and C2 were shown to be important for Tat-dependent protein transport (Buchanan et al. 2002). One study demonstrated that the mutation R16A in TatC (also called TatCd, a TatC homolog) in *B. subtilis* could abolish the secretion of a Tat-dependent substrate PhoD, homologous site of which we detected to be under FD (Table 2, see **Col.** 226) in haloarchaea (Eijlander et al. 2009). Moreover, mutations in

the C2 and P2 domains were shown to generate a Tat version capable of translocating proteins with variations in the RR motif of Tat signal peptide (Kreutzenbeck et al. 2007; Strauch, Georgiou 2007), a motif that needs to be generally conserved for the translocation to take place. These studies suggest that extensive FD in these domains of haloarchaea TatCo has contributed to the expansion of Tat substrates repertoire in this clade. The same rationale applies to the pathogenic bacteria *Leptospira*, *Bartonella* and Rickettsia (De Buck, Lammertyn, Anne 2008; Joshi et al. 2010; Reynolds et al. 2011).

TatC transmembrane domains have been shown to interact with other TatC or TatB to form complexes for signal peptide and substrate binding (Alami et al. 2003; Behrendt, Lindenstrauss, Bruser 2007; Punginelli et al. 2007). However, the exact function of these domains in substrate recognition and binding is less well understood. The observed FD in TatCt transmembrane domains may have led to the formation of a functionally different TatCt complex, which could interact with a specific group of Tat substrates in haloarchaea.

TatC was suggested to play a role in determining the specificity of Tat pathway dependent secretion (Jongbloed et al. 2000). In *B. subtilis* and *E.coli*, both TatCs were shown to serve as a primary RR motif recognition site (Mendel et al. 2008). Haloarchaea encodes two TatC paralogs, and both are among the most functionally divergent clades. Rose and colleagues first showed that *Halobacterium* sp. NRC-1 extensively uses the twin-arginine translocation pathway, instead of Sec, to transport most of its secreted proteins (Rose et al. 2002). Such shift of protein transport from Sec to Tat may have been crucial to solve the protein folding problem faced by microbes in high salt concentration environments: folding proteins first in the cytosol and then exporting

them from the cell. This provides an explanation to the essentiality of TatC proteins for cell viability in halophilic archaea (Dilks, Gimenez, Pohlschroder 2005; Thomas, Bolhuis 2006). Moreover, we analysed the secretomes of two extreme halophilic bacteria *Salinibacter ruber* and *Halorhodospira halophila* whose genomes are available. Strikingly, COG category P—that comprises proteins involved in inorganic ion transport and metabolism--is ranked as the largest group with known function in Sec-dependent secretome of *S. ruber* (Figure S6A). Conversely, in Tat-dependent secretome, functional category P is only ranked at 6th position (Figure S6B). Similarly, P is ranked at the second and fourth positions in Sec- and Tat-dependent secretomes in *H. halophila*, respectively (Figure S6C-D). Not surprisingly, none of the two halophilic bacteria, which present both secretion systems Sec and Tat, showed any sign of strong FD in their Tat components. In the post-translational translocation pathway in bacteria, SecB and SecA, that present chaperonin activity, can bind to the newly synthesized proteins and keep them in an unfolded state, which is required for their Sec-dependent translocation. In halophilic bacteria, SecB and/or SecA can act as holdases preventing the premature folding of the synthesized proteins, ensuring therefore correct protein translocation despite the high salt concentrations. Contrary to this, halophilic archaea do not have SecA and SecB proteins that could mitigate the effects of high salt concentrations on protein translocation (Dilks, Gimenez, Pohlschroder 2005). In these archaea proteins need therefore to be translocated in a folded state, which is possible through the Tat pathway that does not require unfolded proteins. This, however, would be possible provided that the substrate-specificity of TatC has relaxed (for example, recognition of RR motif in a signal peptide of Tat-substrate proteins would be no longer a requirement), possibly through key amino acid substitutions that we detect to be under FD.

In conclusion, the distinct FD patterns observed here between TatCo and TatCt have played an important role in the adaptation of halophilic archaea to environments with high salt concentrations, which is a clear example of how FD of the Tat pathway has contributed to the ecological adaptation of halophilic archaea.

Functional divergence of thermophilic archaea *Sulfolobus* TatA

FD analyses of TatA highlighted thermophilic archaea *Sulfolobus* to be the clade with the strongest profile of FD (Figure 1B, amino acid sites under FD are mapped to TatA alignment shown in Figure S7). The other two clades containing non-pathogenic *Mycobacteria* and pathogenic *E. coli* strains have the same amount of amino acid sites under FD as *Sulfolous* (Figure 1B, Supporting information: Table S5). In TatA we identified a main cluster joining domains APH, NT and H (Figure 1B). Importantly, the C-terminal domain showed strong signal of FD compared to the other domains (Figure 1B). In contrast to TatC, however, the amount of functional divergence in the different domains of TatA was low and we decided therefore to look at the phylogeny that included the entire set of clades to gain insight on FD in TatA. Clustering analyses based on functional divergence approach and taking all clades together for TatA allowed identifying a more general pattern with three well-distinguishable clusters (Supporting information: Figure S2). The first cluster contained the C-terminal domain, being this the one most affected by functional divergence in TatA, the second cluster was almost exclusively represented by the amphipathic helix domain and the third cluster contained the transmembrane, the hinge and the N-terminal domains. This clustering pattern is consistent with previous functional studies that reported mutations in the first 42 residues of TatA, particularly in the amphipathic helix region, to affect significantly Tat-dependent secretion (Hicks et al. 2005). The

C-terminal domain was suggested to play an important functional role in lineage specific substrate transport (Warren et al. 2009) . The low FD observed in TatA transmembrane domain points to the conserved role of this domain, mainly responsible for the formation and structural support of the pore containing complexes for protein transport (Leake et al. 2008; Warren et al. 2009).

Extensive functional divergence in TatB of pathogenic bacteria

Analysis of TatB showed extensive FD affecting clades mostly containing pathogenic bacteria. For example, the top five clades in terms of the amount of amino acid sites under functional divergence included *Corynebacterium*, *Neisseria*, *Bartonella* and *Salmonella*, well-known bacteria for their pathogenic lifestyle (Supporting information: Table S6). Moreover, two human pathogens (*Neisseria gonorrhoeae* and *Neisseria meningitidis*) from the *Neisseria* clade were the second most functionally divergent (11 amino acid sites, supporting information Table S6, amino acid sites under FD are mapped to TatB alignment, as shown in Figure S8). FD affecting TatB was heterogeneously distributed amongst clades and protein domains: the amphipathic helix and the C-terminal domains were the most affected. Interestingly, we observed the same three-cluster pattern as in TatA (Supporting information: Figure S3), suggesting that both proteins may share similar evolutionary histories. In support of this, a fusion protein consisting of the N-terminal region of TatA and the amphipathic helix and C-terminal domain of TatB can maintain a low level of Tat-dependent secretion (Lee et al. 2002). Moreover, the secretion of a tat substrate in *E. coli* can still be detected after mutating TatA in a TatB knock-out strain (Blaudeck et al. 2005). These data indicate that FD of TatB may serve to determine the specificity of some Tat-

dependent substrates through the modulation of the interaction of TatB with TatA, which is important to bacterial virulence (De Buck, Lammertyn, Anne 2008).

Differential protein transport in functionally divergent prokaryotic lineages

We identified the Sec- and Tat-dependent secretome of three halophilic archaeal species, *Halobacterium sp. NCR-1*, *Haloferax volvanii* and *Halomicrobium mukohataei* using SignalP and TatP approaches, respectively (only the data for Tat, and not for Sec, are shown here). Tat substrates in the two secretomes were grouped based on COG functional categories, which were then plotted and ranked according to their numbers of substrates within each of the categories. These numbers were statistically tested for category enrichment in functional divergence by comparing the observed number of secreted proteins within each category to the expected number. Expected number of secreted proteins within each of the categories was calculated by assuming that the percentage of secreted proteins is proportional to the total number of proteins within COG categories for that organism. Interestingly, in *Halobacterium sp. NCR-1*, the functional category P--annotated as Inorganic ion transport and metabolism—was significantly enriched for secreted proteins and was ranked the largest group amongst all the COG functional categories (Figure 2A). Similar patterns were also observed in *Haloferax volvanii* (Figure 2B) and *Halomicrobium mukohataei* (Figure 2C) and in Sec-dependent secretomes (data not shown). It is known that halophilic archaea accumulate potassium ions to counter balance the high salt concentration in their environment (Albers, Szabo, Driessen 2006). It is not surprising therefore to see that a large fraction of Tat-dependent substrates are related to inorganic ion transport and metabolism (Supporting information: Table S7). To determine if FD in TatC from haloarchaea may be correlated with its adaptation to halophilic environments and not to other indirect causes,

we also identified the Sec- and Tat-dependent secretome of three species from the other three non-halophilic lineages under FD (Figure 1A): *Leptospira interrogans Lai*, *Bartonella quintana* and *Rickettsia rickettsii*. Conversely to the case of haloarchaea, Tat-dependent secretome analysis of these bacteria did not rank the COG P category amongst the best represented in the secretome (Figure 2D-F, Supporting information: Table S7).

Sulfolobus clade was ranked among the top ones regarding the enrichment for FD in TatA (Figure 1B). *Sulfolobus* is known to live in hot environments with temperatures ranging between 60 and 90 ℃ (Huber, Prangishvili 2006). Ribosomal proteins (RPs) are shown to bind zinc (Makarova, Ponomarev, Koonin 2001) and mounting evidence point to their alternative extra-cellular location (Trost et al. 2005; Tjalsma et al. 2008; Joshi et al. 2010), in where they likely perform non-ribosomal functions (Warner, McIntosh 2009). Interestingly, prediction of Tat-dependent substrate proteins from one of the *S. islandicus* strains (M.14.25) and subsequent COG grouping of the secreted proteins (Supporting information: Table S8) identified the functional category J—that comprises proteins involved in translation, ribosomal structure and biogenesis—as the most enriched for secreted proteins from the secretome (χ^2 =7.81, P<0.01). 13 ribosomal proteins (RPs) were predicted to be Tat substrate proteins, 6 out of which were predicted to contain zinc binding sites. Moreover, 4 of the 6 RPs were predicted to have multiple zinc binding sites (one has 4 zinc binding sites; three have 5 zinc binding sites) (Shu, Zhou, Hovmoller 2008). To test the significance of Tat-dependent zinc containing RPs, we analysed the 64 RPs from *S. islandicus* M.14.25. We found that there are 19 RPs having at least one zinc-binding site, 4 of which have at least 5 zinc binding sites. Interestingly, 3 out of the 4 RPs are

predicted to be Tat-dependent substrates (χ^2 =5.89, P<0.05). The same analysis was performed in other clades to determine whether this observation was restricted to *Sulfolobus* or was general to thermophiles. Our results suggest that this pattern was unique to *Sulfolobus*. RP with binding zinc was proposed to be more stable, and therefore they are more frequently used in thermophilic bacteria (Makarova, Ponomarev, Koonin 2001). In a thermophilic environment proteins are likely to go denaturation. Zn-binding proteins are more stable in such environments because binding Zn, as other metals, stabilizes proteins (Vetriani et al. 1998; Li, Solomon 2001; Watanabe, Kodaki, Makino 2005). Maintaining the homeostasis of cytosolic Zn(II) is critical to cell physiology as excess metal ions in the cytosol are toxic to the cell. Metallochaperones were shown to maintain homeostasis of Copper in eukaryotic cells (Robinson, Winge 2010). However, no cytoplasmic metallochaperones for Zn(II) have been identified in any organism. A role for Zinc binding RPs in maintaining cytosolic Zn(II) homeostasis has recently been proposed (Gunasekera et al. 2009). In their model, Gunasekera and colleagues also proposed that Zinc binding RPs may participate in Zn(II) export. Interestingly, TatA was shown to have proofreading function of cofactor (Fe, Zn, etc) containing substrate proteins (Matos, Robinson, Di Cola 2008). FD of *S. islandicus* TatA may also contribute to the export of Zn(II) that is bound to Tat-dependent RPs, which can subsequently mediate the homeostasis of Zn(II) in the cytosol.

In pathogenic bacteria, COG functional category J was the largest, although not always significantly enriched, Tat-dependent substrate group with FD (Figure 2D-F). This was the case for all the pathogenic bacterial strains we analysed, such as *Leptospira borgpetersenii JB197* (TatC, Supporting information: Figure S9A), *L. interrogans* (TatC, Figure 2D), *B. quintana* (TatB and TatC, Figure 2E), *Rickettsia prowazekii Madrid E* (TatC, Supporting information:

Figure S9B) and *R. rickettsii* (TatC, Figure 2F)(Supporting information: Table S7), *Corynebacterium diphtheriae* (TatB, Supporting information: Figure S9C), *Neisseria meningitides Z2491(*TatB, Supporting information: Figure S9D). As shown above, zinc containing proteins are known to play vital role in bacterial physiology, including DNA polymerases, proteases and ribosomal proteins, among others. During pathogenesis, bacterial pathogen can experience zinc starvation, which can interfere with the normal function of zinc-containing proteins. A feasible solution would be using a homologous protein which does not need, or that would require very little, zinc to function properly (Panina, Mironov, Gelfand 2003). We found 3 out of 8 secreted RPs from the obligate intracellular bacterial pathogen *R. rickettsii* bearing predicted zinc binding sites, although prediction scores were weak. Among the identified Tat-dependent secreted proteins in *L. borgpetersenii* and *L. interrogans*, we found the 30S ribosomal protein S3 (RPS3). Recently, a study demonstrated that RPS3 is an active component of NF-*k*B transcription factor complex (Wan et al. 2007). NF-*k*B plays an important role in the host immune response to pathogens by regulating the expression of genes involved in different immune pathways (Vallabhapurapu, Karin 2009). Bacterial secreted Tat substrate RPS3 can be important to interfere with the normal function of host RPS3, which would impair NF-*k*B dependent gene regulation and expression. These pathogenic bacteria may therefore evade host immune response by the Tat-dependent secretion of RPS3 and other ribosomal proteins to the immune cells. This conclusion, however, should be taken with caution and requires further investigation.

Acknowledgement

This work was supported by Science Foundation Ireland, under the Research Frontiers Program (10/RFP/Gen2685) and a grant from Ministerio de Ciencia e Innovación (BFU2009-12022) to MAF. XJ is supported by IRCSET (Irish Research Council for Science, Engineering and Technology) Government of Ireland Postgraduate Scholarship in Science, Engineering and Technology. We would like to thank both the editor and reviewers for their important contribution in helping us to improve the clarity and presentation of this manuscript. Finally, we are thankful to all our colleagues who contributed to improve the quality of the writing of this manuscript.

References

Abascal, F, R Zardoya, D Posada. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-2105.

Alami, M, I Luke, S Deitermann, G Eisner, HG Koch, J Brunner, M Muller. 2003. Differential interactions between a twin-arginine signal peptide and its translocase in Escherichia coli. Molecular Cell 12:937-946.

Albers, SV, Z Szabo, AJM Driessen. 2006. Protein secretion in the Archaea: multiple paths towards a unique cell surface. Nature Reviews Microbiology 4:537-547.

Behrendt, J, U Lindenstrauss, T Bruser. 2007. The TatBC complex formation suppresses a modular TatB-multimerization in Escherichia coli. Febs Letters 581:4085-4090.

Bendtsen, JD, H Nielsen, D Widdick, T Palmer, S Brunak. 2005. Prediction of twin-arginine signal peptides. Bmc Bioinformatics 6:167.

Blaudeck, N, P Kreutzenbeck, M Muller, GA Sprenger, R Freudl. 2005. Isolation and characterization of bifunctional Escherichia coli TatA mutant proteins that allow efficient Tat-dependent protein translocation in the absence of TatB. Journal of Biological Chemistry 280:3426-3432.

Bryson, K, LJ McGuffin, RL Marsden, JJ Ward, JS Sodhi, DT Jones. 2005. Protein structure prediction servers at university college london. Nucleic Acids Research 33:W36-W38.

Buchanan, G, E de Leeuw, NR Stanley, M Wexler, BC Berks, F Sargent, T Palmer. 2002. Functional complexity of the twin-arginine translocase TatC component revealed by site-directed mutagenesis. Molecular Microbiology 43:1457-1470.

De Buck, E, E Lammertyn, J Anne. 2008. The importance of the twin-arginine translocation pathway for bacterial virulence. Trends in Microbiology 16:442-453.

Dilks, M, MI Gimenez, M Pohlschroder. 2005. Genetic and biochemical analysis of the twin-arginine translocation pathway in halophilic archaea. Journal of Bacteriology 187:8104-8113.

Edgar, RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Eijlander, RT, JDH Jongbloed, OP Kuipers. 2009. Relaxed Specificity of the Bacillus subtilis TatAdCd Translocase in Tat-Dependent Protein Secretion. Journal of Bacteriology 191:196-202.

Eijlander, RT, MA Kolbusz, EM Berendsen, OP Kuipers. 2009. Effects of altered TatC proteins on protein secretion efficiency via the twin-arginine translocation pathway of Bacillus subtilis. Microbiology-Sgm 155:1776-1785.

Elofsson, A, G von Heijne. 2007. Membrane protein structure: Prediction versus reality. Annual Review of Biochemistry 76:125-140.

Emanuelsson, O, S Brunak, G von Heijne, H Nielsen. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nature Protocols 2:953-971.

Eser, M, M Ehrmann. 2003. SecA-dependent quality control of intracellular protein localization. Proceedings of the National Academy of Sciences of the United States of America 100:13231-13234.

Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication. Molecular Biology and Evolution 16:1664-1674.
Gunasekera, T, JA Easton, A Sugerbaker Stacy, L Klingbeil, W Crowder Michael. 2009. Zn(II) Homeostasis in *E. coli*. Bioinorganic Chemistry: American Chemical Society. p. 81-95.
Henikoff, S, JG Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America 89:10915-10919.
Hicks, MG, PA Lee, G Georgiou, BC Berks, T Palmer. 2005. Positive selection for loss-of-function tat mutations identifies critical residues required for TatA activity. Journal of Bacteriology 187:2920-2925.
Holzapfel, E, G Eisner, M Alami, et al. 2007. The entire N-terminal half of TatC is involved in twin-arginine precursor binding. Biochemistry 46:2892-2898.
Huber, H, D Prangishvili. 2006. Sulfolobales. Prokaryotes: A Handbook on the Biology of Bacteria, Vol 3, Third Edition:Archaea. Bacteria: Firmicutes, Actinomycetes. p. 23-51.
Irihimovitch, V, J Eichler. 2003. Post-translational secretion of fusion proteins in the halophilic archaea Haloferax volcanii. Journal of Biological Chemistry 278:12881-12887.
Jongbloed, JDH, U Martin, H Antelmann, M Hecker, H Tjalsma, G Venema, S Bron, JM van Dijl, J Muller. 2000. TatC is a specificity determinant for protein secretion via the twin-arginine translocation pathway. Journal of Biological Chemistry 275:41350-41357.
Joshi, MV, SG Mann, H Antelmann, et al. 2010. The twin arginine protein transport pathway exports multiple virulence proteins in the plant pathogen Streptomyces scabies. Molecular Microbiology 77:252-271.
Kanehisa, M, M Araki, S Goto, et al. 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Research 36:D480-D484.
Kreutzenbeck, P, C Kroger, F Lausberg, N Blaudeck, GA Sprenger, R Freudl. 2007. Escherichia coli twin arginine (Tat) mutant translocases possessing relaxed signal peptide recognition specificities. Journal of Biological Chemistry 282:7903-7911.
Leake, MC, NP Greene, RM Godun, T Granjon, G Buchanan, S Chen, RM Berry, T Palmer, BC Berks. 2008. Variable stoichiometry of the TatA component of the twin-arginine protein transport system observed by in vivo single-molecule imaging. Proceedings of the National Academy of Sciences of the United States of America 105:15376-15381.
Lee, PA, G Buchanan, NR Stanley, BC Berks, T Palmer. 2002. Truncation analysis of TatA and TatB defines the minimal functional units required for protein translocation. Journal of Bacteriology 184:5871-5879.
Lee, PA, D Tullman-Ercek, G Georgiou. 2006. The bacterial twin-arginine translocation pathway. Annual Review of Microbiology 60:373-395.
Letunic, I, T Doerks, P Bork. 2009. SMART 6: recent updates and new developments. Nucleic Acids Research 37:D229-D232.
Li, X, B Solomon. 2001. Zinc-mediated thermal stabilization of carboxypeptidase A. Biomolecular engineering 18:179-183.
Makarova, KS, VA Ponomarev, EV Koonin. 2001. Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. Genome Biology 2.

Matos, C, C Robinson, A Di Cola. 2008. The Tat system proofreads FeS protein substrates and directly initiates the disposal of rejected molecules. Embo Journal 27:2055-2063.
McFarland, L, O Francetic, CA Kumamoto. 1993. A mutation of Escherichia coli SecA protein that partially compensates for the absence of SecB. Journal of Bacteriology 175:2255-2262.
Mendel, S, A McCarthy, JP Barnett, RT Eijlander, A Nenninger, OP Kuipers, C Robinson. 2008. The Escherichia coli TatABC system and a Bacillus subtilis TatAC-type system recognise three distinct targeting determinants in twin-arginine signal peptides. Journal of Molecular Biology 375:661-672.
Panahandeh, S, E Holzapfel, M Mueller. 2009. The Twin-arginine Translocation Pathway. In: K Wooldridge, editor. Bacterial Secreted Proteins: Secretory Mechanisms and Role in Pathogenesis. Norfolk, UK: Caister Academic Press. p. 23-43.
Panina, EM, AA Mironov, MS Gelfand. 2003. Comparative genomics of bacterial zinc regulons: Enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. Proceedings of the National Academy of Sciences of the United States of America 100:9912-9917.
Popot, J. 2010. Amphipols, Nanodiscs, and Fluorinated Surfactants: Three Nonconventional Approaches to Studying Membrane Proteins in Aqueous Solutions. Annual Review of Biochemistry 79.
Punginelli, C, B Maldonado, S Grahl, R Jack, M Alami, J Schroder, BC Berks, T Palmer. 2007. Cysteine scanning mutagenesis and topological mapping of the Escherichia coli twin-arginine translocase TatC component. Journal of Bacteriology 189:5482-5494.
Reynolds, M, L Bogomolnaya, J Guo, L Aldrich, D Bokhari, C Santiviago, M McClelland, H Andrews-Polymenis, M Hensel. 2011. Abrogation of the Twin Arginine Transport System in Salmonella enterica Serovar Typhimurium Leads to Colonization Defects during Infection. PLoS One 6:18003-18006.
Robinson, NJ, DR Winge. 2010. Copper metallochaperones. Annual Review of Biochemistry 79:537-562.
Rose, RW, T Bruser, JC Kissinger, M Pohlschroder. 2002. Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway. Molecular Microbiology 45:943-950.
Shu, NJ, TP Zhou, S Hovmoller. 2008. Prediction of zinc-binding sites in proteins from sequence. Bioinformatics 24:775-782.
Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.
Strauch, EM, G Georgiou. 2007. Escherichia coli tatC mutations that suppress defective twin-arginine transporter signal peptides. Journal of Molecular Biology 374:283-291.
Tarry, MJ, E Schafer, SY Chen, G Buchanan, NP Greene, SM Lea, T Palmer, HR Saibil, BC Berks. 2009. Structural analysis of substrate binding by the TatBC component of the twin-arginine protein transport system. Proceedings of the National Academy of Sciences of the United States of America 106:13284-13289.
Tatusov, RL, ND Fedorova, JD Jackson, et al. 2003. The COG database: an updated version includes eukaryotes. Bmc Bioinformatics 4:41.

Thomas, JR, A Bolhuis. 2006. The tatC gene cluster is essential for viability in halophilic archaea. Fems Microbiology Letters 256:44-49.
Tjalsma, H, L Lambooy, PW Hermans, DW Swinkels. 2008. Shedding & shaving: Disclosure of proteomic expressions on a bacterial face. Proteomics 8:1415-1428.
Toft, C, TA Williams, MA Fares. 2009. Genome-Wide Functional Divergence after the Symbiosis of Proteobacteria with Insects Unraveled through a Novel Computational Approach. Plos Computational Biology 5.
Trost, M, D Wehmhoner, U Karst, G Dieterich, J Wehland, L Jansch. 2005. Comparative proteome analysis of secretory proteins from pathogenic and nonpathogenic Listeria species. Proteomics 5:1544-1557.
Vallabhapurapu, S, M Karin. 2009. Regulation and Function of NF-kappa B Transcription Factors in the Immune System. Annual Review of Immunology 27:693-733.
Vetriani, C, DL Maeder, N Tolliday, KSP Yip, TJ Stillman, KL Britton, DW Rice, HH Klump, FT Robb. 1998. Protein thermostability above 100 degrees C: A key role for ionic interactions. Proceedings of the National Academy of Sciences of the United States of America 95:12300-12305.
Wan, FY, DE Anderson, RA Barnitz, et al. 2007. Ribosomal protein S3: A KH domain subunit in NF-kappa B complexes that mediates selective gene regulation. Cell 131:927-939.
Warner, JR, KB McIntosh. 2009. How Common Are Extraribosomal Functions of Ribosomal Proteins? Molecular Cell 34:3-11.
Warren, G, J Oates, C Robinson, AM Dixon. 2009. Contributions of the Transmembrane Domain and a Key Acidic Motif to Assembly and Function of the TatA Complex. Journal of Molecular Biology 388:122-132.
Watanabe, S, T Kodaki, K Makino. 2005. Complete reversal of coenzyme specificity of xylitol dehydrogenase and increase of thermostability by the introduction of structural zinc. The Journal of biological chemistry 280:10340-10349.
Waterhouse, AM, JB Procter, DMA Martin, M Clamp, GJ Barton. 2009. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189-1191.
Widdick, DA, K Dilks, G Chandra, A Bottrill, M Naldrett, M Pohlschroder, T Palmer. 2006. The twin-arginine translocation pathway is a major route of protein export in Streptomyces coelicolor. Proceedings of the National Academy of Sciences of the United States of America 103:17927-17932.
Williams, TA, FM Codoner, C Toft, MA Fares. 2010. Two chaperonin systems in bacterial genomes with distinct ecological roles. Trends in Genetics 26:47-51.
Wooldridge, K, editor. 2009. Bacterial Secreted Proteins: Secretory Mechanisms and Role in Pathogenesis. Norfolk, UK: Caister Academic Press.
Yuan, JJ, JC Zweers, JM van Dijl, RE Dalbey. 2010. Protein transport across and into cell membranes in bacteria and archaea. Cellular and Molecular Life Sciences 67:179-199.

Tables

Table 1. Amino acid sites under functional divergence in haloarchaea TatCo. Amino acid residues are represented with one letter code. Numbers of the same amino acid residues are also calculated in each clade. Amino acid sites under functional divergence are also mapped to TatC alignment shown in Figure S4.

Col[1].	Z[2]	Residues (count) for **C1**/C2/Outgroup[3]
226	8.633***	**V(4) M(1) I(1) A(2)**/ R(4)/ R(1)
242	3.321***	**A(3) T(2) S(3)**/ I(2) M(2)/ V(1)
245	2.878*	**V(6) L(1) A(1)**/ A(2) S(1) I(1)/ S(1)
539	5.549***	**H(2) G(2) Q(2) N(2**)/ S(3) A(1)/ S(1)
543	198.49***	**L(7) T(1)**/ W(3) M(1)/ W(1)
551	34.199***	**V(8)**/ L(3) M(1)/ F(1)
553	5.26***	**T(3) G(1) S(3) A(1)**/ L(3) F(1)/ L(1)
569	10.404***	**T(6) S(2)**/ L(3) A(1)/ L(1)
577	197***	**M(8)**/ A(4)/ A(1)
596	73.907***	**Y(8)**/ F(3) L(1)/ L(1)
599	10.538***	**S(3) A(5)**/ I(1) V(3)/ I(1)
604	116.59***	**L(7) V(1)**/ F(3) L(1)/ Y(1)
607	19.379***	**G(2) S(2) A(4)**/ F(4)/ F(1)
611	2.798*	**V(5) I(2) L(1)**/ A(2) T(1) G(1)/ T(1)
629	43.133***	**Y(8)**/ Y(3) I(1)/ I(1)
709	4.966***	**I(3) V(3) L(1) T(1)**/ A(3) P(1)/ P(1)
808	34.199***	**L(8)**/ V(4)/ A(1)
811	3.125***	**T(3) A(1) G(2) R(1) S(1)**/ Q(3) G(1)/ E(1)
815	3.376***	**D(2) N(6)**/ S(4)/ T(1)
817	34.199***	**I(4) L(2) M(2)**/ V(4)/ A(1)
823	34.199***	**L(2) F(3) Y(3)**/ G(4)/ S(1)
839	115.6***	**M(8)**/ K(4)/ R(1)
844	13.071***	**T(7) S(1)**/ E(1) K(2) Q(1)/ K(1)
845	131.98***	**R(8)**/ Y(3) R(1)/ Y(1)
847	135.95***	**W(8)**/ T(4)/ T(1)
859	4.404***	**G(7) A(1)**/ V(1) A(2) L(1)/ L(1)
860	6.419***	**F(3) S(2) A(3)**/ I(3) L(1)/ I(1)
879	237.7***	**P(8)**/ Q(4)/ Q(1)
883	43.133***	**A(7) T(1)**/ A(3) L(1)/ L(1)
885	167.22***	**T(8)**/ L(3) V(1)/ L(1)
893	34.199***	**T(8)**/ S(4)/ G(1)
902	54.715***	**W(8)**/ L(3) F(1)/ I(1)

[1] Relative position of amino acid sites in TatC multiple sequence alignment
[2] Score of Functional Divergence: *, ** and *** indicate P<0.05, P<0.01, P<0.001, respectively
[3] C1 is the clade under functional divergence, C2 is the clade against which comparative analyses are made.

Table 2. Amino acid sites under functional divergence in haloarchaea TatCt. Amino acid residues are represented with one letter code. Numbers of the same amino acid residues are also calculated in each clade. Amino acid sites under functional divergence are also mapped to TatC alignment shown in Figure S5.

Col[1].	Z[2]	Residues (count) for **C1**/C2/Outgroup[3]
190	4.33***	**A(4) G(1) S(1) T(1)**/ E(3) P(1) M(1)/ P(1)
225	3.919***	**A(6) I(1)**/ L(4) F(1)/ L(1)
230	52.645***	**Q(7)**/ L(4) T(1)/ L(1)
236	4.798***	**F(6) W(1)**/ L(3) I(1) V(1)/ L(1)
244	4.53***	**G(5) A(2)**/ V(1) I(2) L(2)/ I(1)
536	61.242***	**A(6) T(1)**/ Y(3) Q(1) L(1)/ Y(1)
570	16.101***	**D(3) R(3) N(1)**/ Y(4) L(1)/ Y(1)
617	2.439*	**L(4) A(2) V(1)**/ L(3) F(1) M(1)/ F(1)
618	42.122***	**F(7)**/ M(5)/ V(1)
620	33.667***	**F(7)**/ L(4) F(1)/ V(1)
631	78.648***	**A(7)**/ L(3) Y(1) M(1)/ L(1)
707	5.134***	**F(5) L(2)**/ V(4) A(1)/ A(1)
815	4.52***	**Q(3) E(3) G(1)**/ S(3) N(2)/ T(1)
817	4.737***	**V(2) I(5)**/ I(2) A(2) V(1)/ A(1)
828	27.752***	**A(4) S(2) G(1)**/ F(5)/ F(1)
839	68.695***	**Y(7)**/ R(4) K(1)/ R(1)
844	2.905*	**P(4) A(1) Q(1) R(1)**/ K(3) P(1) Q(1)/ K(1)
853	56.751***	**W(7)**/ R(5)/ R(1)
857	173.78***	**V(5) I(2)**/ Y(5)/ Y(1)
861	7.867***	**F(5) A(1) Y(1)**/ M(1) L(3) F(1)/ L(1)
868	26.379***	**F(7)**/ V(2) I(2) T(1)/ T(1)
891	19.033***	**G(5) V(1) A(1)**/ E(5)/ E(1)
895	3.558***	**A(1) Y(4) G(1) Q(1)**/ V(3) L(2)/ L(1)
897	6.241***	**A(3) S(3) T(1)**/ V(4) L(1)/ L(1)

[1] Relative position of amino acid sites in TatC multiple sequence alignment

[2] Score of Functional Divergence: *, ** and *** indicate P<0.05, P<0.01, P<0.001, respectively

[3] C1 is the clade under functional divergence, C2 is the clade against which comparative analyses are made.

Figure legends

Figure 1. Clustering analysis of phylogenetic clades and protein domains according to functional divergence analyses. We examined enrichment for amino acid sites with functional divergence in the proteins of the Tat secretion pathway TatC (A), TatA (B) and TatB (C) under functional divergence. The relative position of the enriched clades for functional divergence is shown in red in the maximum-likelihood phylogenetic tree next to each of the heatmaps. The high resolution trees are available as supporting information (TatC: Figure S10; TatA: Figure S11; TatB: Figure S12).

Figure 2. Analysis of Tat-dependent secreted proteins (secretome) of Halophilic archaea and pathogenic bacteria. Substrates are grouped functionally according to the classification of proteins into the Cluster of Orhtologous Groups (COG). These functional categories were ranked and plotted according to the numbers of their substrates. (A) *Halobacterium sp. NCR*-1 (B) *Haloferax volvanii* (C) *Halomicrobium mukohataei* (D) *Leptospira interrogans* (E) *Bartonella quintana* (F) *Rickettsia rickettsii*. COG functional categories showing significant enrichment (black stars) or impoverishment (grey stars) of predicted substrate proteins are labelled by (*: $P<0.05$; **: $P<0.01$; ***: $P<0.001$).

Supporting information tables

Table S1. Functional divergence analysis of TatC. We identified 204 clades with at least one amino acid sites under functional divergence. NA denotes Host or disease unknown.

Table S2. Functional divergence analysis of TatA. We identified 145 clades with at least one amino acid sites under functional divergence. NA denotes Host or disease unknown.

Table S3. Functional divergence analysis of TatB. We identified 119 clades with at least one amino acid sites under functional divergence. NA denotes Host or disease unknown.

Table S4. Amino acid sites under functional divergence in TatC for haloarchaea, *Leptospira*, *Bartonella* and *Rickettsia*.

Table S5. Amino acid sites under functional divergence in TatA for *Sulfolobus*, non-pathogenic *Mycobacteria* and pathogenic *Escherichia coli* .

Table S6. Amino acid sites under functional divergence in TatB for *Corynebacterium*, *Neisseria*, *Bartonella* and *Salmonella*.

Table S7. Predicted Tat-dependent substrates annotated with COG for *Bartonella henselae str. Houston-1, Bartonella quintana* str. Toulouse, *Rickettsia prowazekii* str. Madrid E and *Rickettsia rickettsii* str. 'Sheila Smith'.

Table S8. Predicted Tat-dependent substrates annotated with COG for *Sulfolobus islandicus* M.14.25, *Escherichia coli* O157:H7 str. EDL933 and *Mycobacterium smegmatis* str. MC2 155.

Table S9. Predicted Tat-dependent substrates annotated with COG for *Neisseria meningitidis Z2491, Bartonella henselae str. Houston-1, Bartonella quintana* str. Toulouse and *Salmonella enterica subsp. enterica serovar Typhi str. CT18*.

Table S10. Predicted Sec-dependent substrates annotated with COG for two halophilic bacteria *Halorhodospira halophila* and *Salinibacter ruber*.

Table S11. Predicted Tat-dependent substrates annotated with COG for two halophilic bacteria *Halorhodospira halophila* and *Salinibacter ruber*.

Supporting information figure legends

Figure S1. Clustering analysis of all clades with functional divergence for TatC.

Figure S2. Clustering analysis of all clades with functional divergence for TatA.

Figure S3. Clustering analysis of all clades with functional divergence for TatB.

Figure S4. Amino acid sites of haloarchaea TatCo under functional divergence mapped to TatC alignment. Sites here from haloarchaea TatCo correspond to the sites in Table 1. Protein sequences in the alignment forming clade 1, clade 2 and outgroup are labelled accordingly.

Figure S5. Amino acid sites haloarchaea of TatCt under functional divergence mapped to TatC alignment. Sites here from haloarchaea TatCt correspond to the sites in Table 2. Protein sequences in the alignment forming clade 1, clade 2 and outgroup are labelled accordingly.

Figure S6. Sec- and Tat-dependent secretome analysis of two extreme halophilic bacteria *Halorhodospira halophila* SL1 and *Salinibacter ruber* DSM 13855. COG functional categories showing significant enrichment (black stars) or impoverishment (grey stars) of predicted substrate proteins are labelled by (*: $P<0.05$; **: $P<0.01$; ***: $P<0.001$).

Figure S7. Amino acid sites of *Sulfolobus* TatA under functional divergence mapped to TatA alignment. Sites here from *Sulfolobus* TatA correspond to the sites in Table S5. Protein sequences in the alignment forming clade 1, clade 2 and outgroup are labelled accordingly.

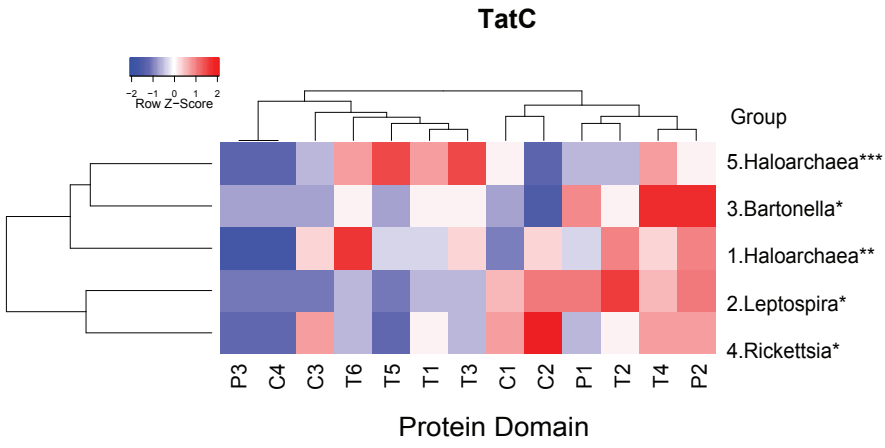
Figure S8. Amino acid sites of *Neisseria* TatB under functional divergence mapped to TatB alignment. Sites here from *Neisseria* TatA correspond to the sites in Table S6. Protein sequences in the alignment forming clade 1, clade 2 and outgroup are labelled accordingly.

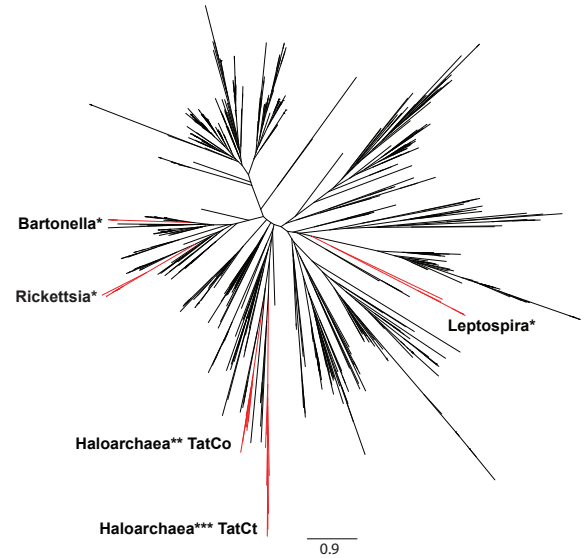
Figure S9. Tat-dependent secretome analysis of pathogenic bacteria. Substrates are grouped by COG functional category in each secretome, these functional categories are then ranked and plotted according to numbers of their substrates. (A) *Leptospira borgpetersenii* JB197, (B) *Rickettsia prowazekii* Madrid E, (C) *Corynebacterium diphtheriae,* (D) *Neisseria meningitides* Z2491. COG functional categories showing significant enrichment (black stars) or impoverishment (grey stars) of predicted substrate proteins are labelled by (*: $P<0.05$; **: $P<0.01$; ***: $P<0.001$).

Figure S10. The maximum likelihood phylogenetic tree for TatC, it is a high-resolution version of Figure 1A.

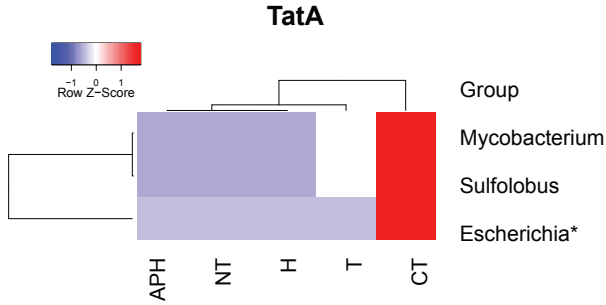
Figure S11. The maximum likelihood phylogenetic tree for TatA, it is a high-resolution version of Figure 1B.

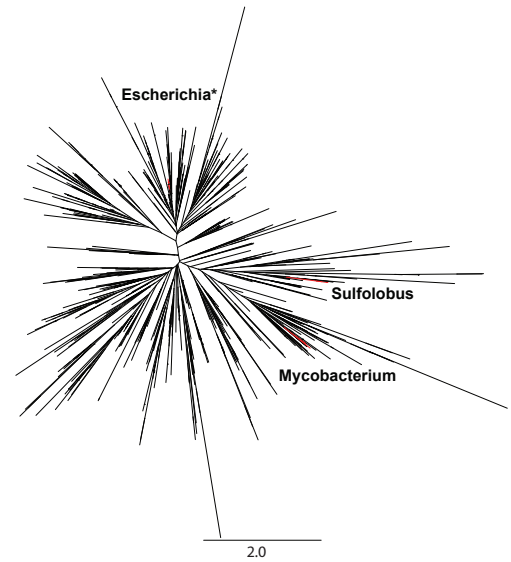
Figure S12. The maximum likelihood phylogenetic tree for TatB, it is a high-resolution version of Figure 1C.



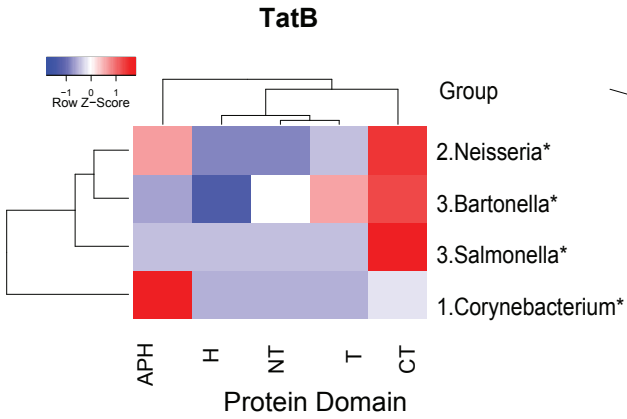


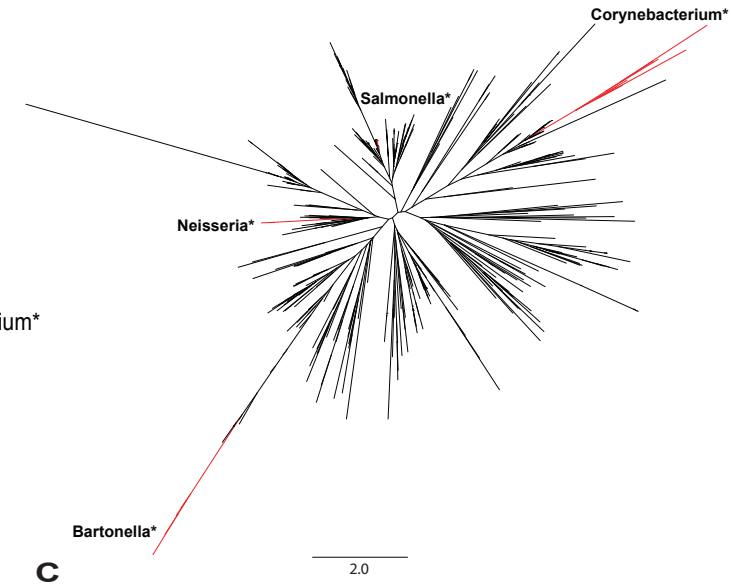
A





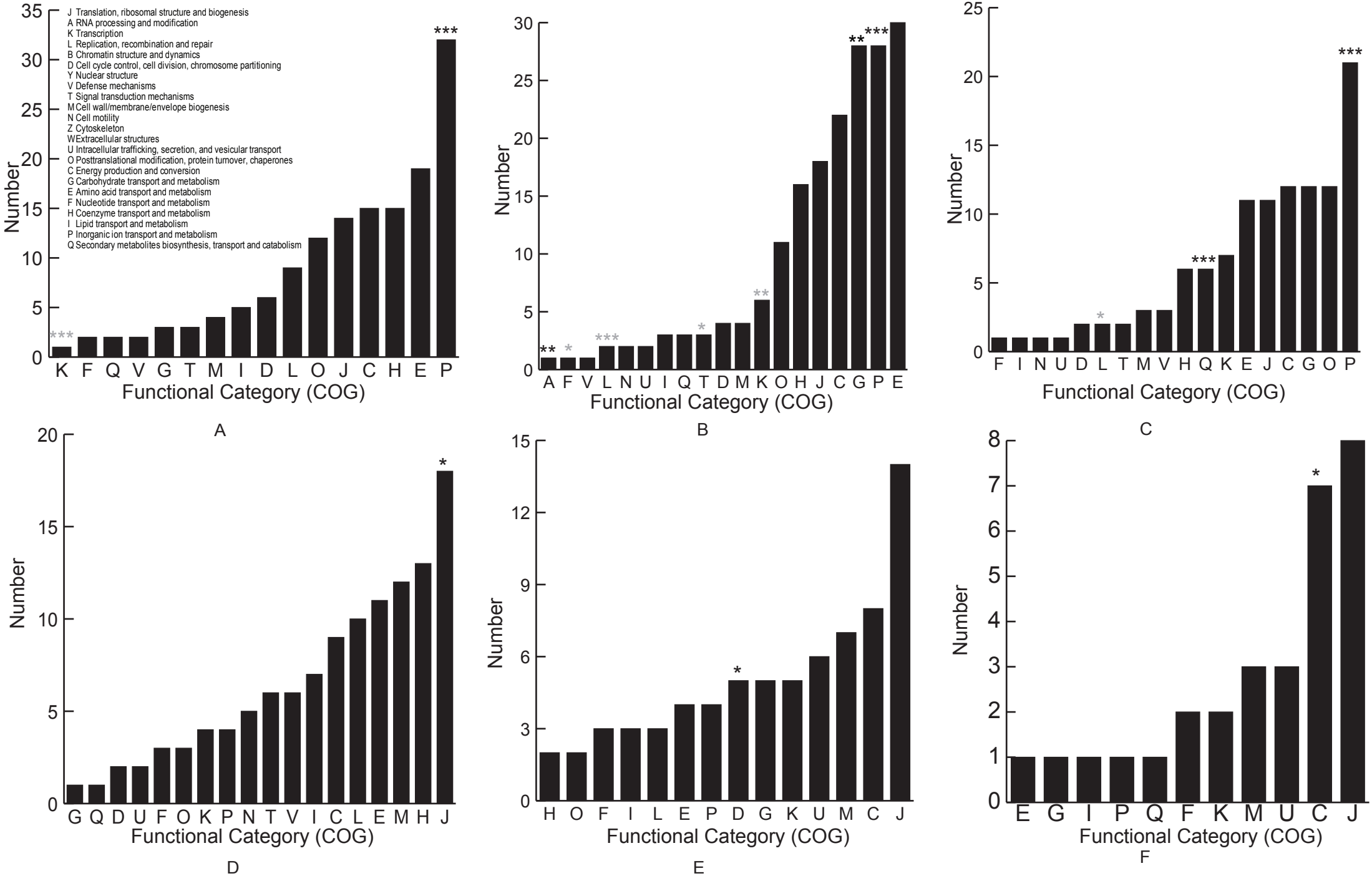
B





C

C	Cytoplasmic domain
APH	Amphipathic helix
CT	C-terminal domain
H	Hinge
P	Periplasmic domain
T	Transmembrane domain
NT	N-terminal domain
*	Pathogenic bacteria
**	TatCo
***	TatCt


J Translation, ribosomal structure and biogenesis
A RNA processing and modification
K Transcription
L Replication, recombination and repair
B Chromatin structure and dynamics
D Cell cycle control, cell division, chromosome partitioning
Y Nuclear structure
V Defense mechanisms
T Signal transduction mechanisms
M Cell wall/membrane/envelope biogenesis
N Cell motility
Z Cytoskeleton
W Extracellular structures
U Intracellular trafficking, secretion, and vesicular transport
O Posttranslational modification, protein turnover, chaperones
C Energy production and conversion
G Carbohydrate transport and metabolism
E Amino acid transport and metabolism
F Nucleotide transport and metabolism
H Coenzyme transport and metabolism
I Lipid transport and metabolism
P Inorganic ion transport and metabolism
Q Secondary metabolites biosynthesis, transport and catabolism
Number
Functional Category (COG)
A
B
C
D
E
F