

Identification of protein secretion systems in bacterial genomes

Sophie S Abby ^{a,b,#,*}, Jean Cury ^{a,b}, Julien Guglielmini ^{a,b,+}, Bertrand Néron ^c, Marie Touchon ^{a,b}, Eduardo PC Rocha ^{a,b}

^a Institut Pasteur, Microbial Evolutionary Genomics, Paris, 75015, France

^b CNRS, UMR3525, Paris, 75015, France

^c Institut Pasteur, C3BI, CIB, Paris, 75015, France

Running head: Protein secretion systems in bacteria

*Address correspondence to sophie.abby.univ@gmail.com

Present addresses:

[#] Division of Archaea Biology and Ecogenomics, Department of Ecogenomics and Systems Biology, University of Vienna, A-1090 Vienna, Austria

⁺ Bioinformatics and Biostatistics HUB, Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI), Institut Pasteur, Paris, 75015, France

KEYWORDS: MICROBIOLOGY/GENOME ANALYSIS/COMPARATIVE GENOMICS/EVOLUTIONARY GENOMICS.

Abstract

Bacteria with two cell membranes (diderms) have evolved complex systems for protein secretion. These systems were extensively studied in some model bacteria, but the characterisation of their diversity has lagged behind due to lack of standard annotation tools. We built models for accurate identification of protein secretion systems and related appendages in bacteria with LPS-containing outer membranes. They can be used with MacSyFinder (standalone program) or online (<http://mobyli.pasteur.fr/cgi-bin/portal.py#forms::txsscan>). They include protein profiles and information on the system's composition and genetic organisation. They can be used to search for T1SS-T6SS, T9SS, and accessorially for flagella, Type IV and Tad pili. We identified ~10,000 systems in bacterial genomes, where T1SS and T5SS were by far the most abundant and widespread. The recently described T6SSⁱⁱⁱ and T9SS were restricted to Bacteroidetes, and T6SSⁱⁱ to *Francisella*. T2SS, T3SS, and T4SS were frequently encoded in single-copy in one locus, whereas most T1SS were encoded in two loci. The secretion systems of diderm Firmicutes were similar to those found in other diderms. Novel systems may remain to be discovered, since some clades of environmental bacteria lacked all known protein secretion systems. Our models can be fully customized, which should facilitate the identification of novel systems.

Introduction

Proteins secreted by bacteria are involved in many important tasks such as detoxification, antibiotic resistance, and scavenging ¹. Secreted proteins also have key roles in both intra- and inter-specific antagonistic and mutualistic interactions ^{2,3}. For example, they account for many of the virulence factors of pathogens ^{4,5}. Bacteria with a LPS-containing outer-membrane (abbreviated "diderm-LPS" in this article) require specific protein secretion systems. Six types of secretion systems, numbered T1SS to T6SS, were well characterised by numerous experimental studies (for some general reviews see ⁶⁻⁸). The T9SS (PorSS) was more recently uncovered in Bacteroidetes ^{9,10}. In this study, we focus only on typical diderm-LPS protein secretion and homologous systems. A few other systems have been described in the literature. For example, the chaperone-usher pathway, sometimes named T7SS, and the T8SS were not included in this study because they are only involved, respectively, in the export of type I pili and curli ¹¹. The ESX system of *Mycobacteria*, T7SS according to some authors ¹², was also ignored because it is absent from diderm-LPS bacteria.

The role of secreted proteins has spurred interest in the production of ontologies to categorise ¹³, and computational methods to identify and characterise protein secretion systems (Table 1). This is a difficult task for a number of reasons. Firstly, protein secretion systems are large machineries with many different components, some of which are accessory and some interchangeable. Secondly, many of their key components are homologous between systems, which complicates their discrimination. For example, T2SS, T4SS and T6SS include distinct but homologous NTPases ¹⁴. Some bacterial appendages require their own secretion systems to translocate their extracellular components ^{15,16}, and these are sometimes partly homologous to classical secretion systems. For example, several components of the type IV pilus (T4P) and the Tad pilus

are homologous to components of the T2SS from *Klebsiella oxytoca*¹⁷. Thirdly, the sequences of secreted proteins, including extracellular components of the secretion systems, evolve rapidly, thereby complicating the identification of homology by sequence similarity¹⁸. Fourthly, loci encoding secretion systems are frequently horizontally transferred and lost^{19,20}, leading to the presence of partial (often inactive) systems in genomes²¹. Finally, experimental studies have focused on a small number of occurrences of each type of system, complicating the assessment of their genetic diversity. On the other hand, secretion systems are often encoded in one or a few neighbouring operons made of highly conserved, and in some cases highly specific, components. This information can facilitate the identification of genes encoding secretion systems in genome data^{22,23}.

Several programs were previously made available to identify components of some, but not all, protein secretion systems (Table 1). These programs are very useful to the biologist interested in browsing the known systems or in annotating a small set of sequences. However, they are web-based, and thus poorly adapted for the analysis of very large datasets. Most of them were not designed to categorise systems as complete or incomplete, and the few that were do not allow the identification of systems scattered in the chromosome, nor the re-definition of the categorisation parameters. Yet, the ability to re-define the type and number of components in a system facilitates the search for distant variants of the well-known experimental model systems²⁴. We have used the vast body of knowledge accumulated from experimental studies of model protein secretion systems to identify new occurrences of these systems. This was achieved by building computational models that account for the systems' composition and genetic organization. The models can be plugged in MacSyFinder²⁵ to identify protein secretion systems. This can be done using the standalone application and using local resources, or using the webserver version

90 available at <http://mobyli.pasteur.fr/cgi-bin/portal.py#forms::txsscan>. The results can be
 91 visualized with MacSyView ²⁵. In the standalone application, the users can easily modify
 92 the models to change the composition and genetic organisation of the secretion systems.
 93 Some of these parameters can also be modified in the webserver version. The accuracy of
 94 the models was quantified against an independent dataset of experimentally validated
 95 systems. Importantly, we provide models to search for an unparalleled number of protein
 96 secretion systems (and some partly homologous systems): T1SS, T2SS (Tad and T4P),
 97 T3SS (flagellum), T4SS (conjugation system), T5SS, T6SSⁱ⁻ⁱⁱⁱ, and T9SS. We used the
 98 models to search for protein secretion systems in a large panel of bacterial genomes.
 99 Previous surveys, mostly dating from a time when few genomes were available, analysed
 100 the distribution of some specific protein secretion systems in genomes or metagenomes
 101 ^{17,20,24,26-31}. Due to space limitations, we will not attempt at re-assessing all these works.
 102 Instead, we describe our models, show their accuracy, and use them to provide a broad
 103 view of the relative distribution of the different protein secretion systems.

104 Results and discussion

105 Overview of the approach

106 We defined 22 customisable models for the protein secretion systems and related
 107 appendages (File S1, Fig. 1-6, Fig. S1-S5). This was done in four steps (see Materials and
 108 Methods for details). Firstly, we searched the primary literature, reviews and books for
 109 references of well-studied systems ^{9,16,20,26,28,32-40}. We used them to define two
 110 independent datasets (*reference* and *validation*) of experimentally studied secretion
 111 systems (Tables S1 and S2). Secondly, the *reference dataset* was used to define the
 112 model for each type of system. The model includes information on the number of
 113 components that are *mandatory* (necessarily present in a system), *accessory* (not
 114 necessarily present in the system), and *forbidden* (never present in the system). Each
 115 component is associated with a protein profile that is used to search its occurrences (with
 116 HMMER ⁴¹). Our models use 204 protein profiles (included in the package), of which 194
 117 were built in our laboratory and the rest taken from public databases (Table S4). Protein
 118 profiles are more sensitive and specific than Blast-based approaches ⁴². The model
 119 indicates which genes are co-localised (at less than a given distance relative to contiguous
 120 genes in the cluster), and which genes might be encoded elsewhere in the genome
 121 (designated "loners"). Thirdly, the models were validated both in the *reference* and in the
 122 independent *validation* datasets using MacSyFinder ²⁵. Finally, the models were used with
 123 MacSyFinder to identify occurrences of each system in 1,528 complete genomes of
 124 diderm species. This procedure retrieved automatically all validly identified secretion
 125 systems. It also retrieved all hits to each component identified in the genomes whether
 126 they are part of a protein secretion system or not. In the following sections we describe the

models of each type of protein secretion system and the occurrences of the system in bacterial genomes.

T1SS

We built protein profiles for the three essential components of T1SS^{32,43,44}: the ABC-transporter (ATP-binding cassette transporter) providing an inner membrane channel, the porin (outer membrane factor, OMF) forming the outer-membrane channel, and the inner-membrane anchored adaptor protein (or membrane-fusion protein, MFP) that connects the OMF and the ABC components (Materials and Methods, Tables S4-S5). T1SS can be difficult to identify because its components have homologs involved in other machineries, e.g., in ABC transporters for the ABC and in drug efflux systems for the MFP, or can themselves be involved in other machineries, in the case of the OMF⁴⁵⁻⁴⁹. The design of the model was facilitated by the previous observation that genes encoding the ABC and MFP components are always co-localised in T1SS loci^{32,43}. The model is described in Figure 1a, and its use resulted in the correct identification of all T1SS in the reference and in the validation datasets. Overall, 20,847 proteins matched the protein profiles of the T1SS components in the bacterial genomes (Fig. 1b). The vast majority of these were not part of T1SS because they did not fit the T1SS model. We found 1,637 occurrences of the T1SS model in 821 genomes (Fig. 1b). The remaining proteins are probably associated with the numerous other systems carrying components homologous to those of the T1SS.

We found T1SS in more than half of the genomes of diderm bacteria (54%). Some genomes contained many systems; e.g., *Bradyrhizobium oligotrophicum* S58 and *Nostoc* sp. PCC 7524 encoded a record number of 9 systems (Table S3). ABC and MFP were encoded together and OMF apart in more than half (57%) of the T1SS. Hence, the co-localisation of the genes encoding the three components concerns a minority of T1SS

(Fig. 1c). We found 95 loci encoding ABC and MFP in replicons lacking OMF. Many of these systems may be functional, since 94 of these loci were found in genomes encoding at least one OMF in another replicon. Such multi-replicon functional T1SS have been previously reported⁵⁰.

T2SS, T4P and Tad pili

T2SS are encoded by 12 to 16 genes, many of which are homologous to components of the T4P and the Tad pilus^{17,51,52} (Fig. 2). We used the protein families conserved in the reference dataset to build 13 protein profiles for T2SS, 11 for type IV pili and 10 for Tad pili (Materials and Methods, Tables S4-S5). No profiles were built for GspA and GspB, because they were rarely identified in T2SS plus we could not build protein families given low sequence similarity. The most frequent components in the reference dataset were defined as mandatory in the models, and the least conserved as accessory. Some profiles built for one type of system produced matches to (homologous) components of other types of systems. Discrimination between systems was facilitated by the definition of some specific components as forbidden (e.g., GspC was declared forbidden in Tad and T4P).

Using the models, we correctly identified all of the T2SS and Tad systems of the reference dataset (Table S1). In the validation dataset we missed some components of four of the 17 T2SS (Table S2). We could retrieve three of them by modifying the default T2SS model, e.g., the "Xps-type" system could be detected by decreasing the required number components⁵³. Additionally, we missed the very atypical T2SS of *Legionella pneumophila* because it failed the co-localisation criterion (unusually, it is encoded in 5 distant loci, Table S2)⁵⁴. A default model with more relaxed parameters in terms of co-

localisation and sequence similarity would have identified all T2SS, but at the cost of less correct discrimination from the two other systems.

The quality of the default T2SS model was confirmed by the analysis of genomic data. Proteins matched by the protein profiles of T2SS were typically either highly or poorly clustered (Fig. S1a). These two extremes were respectively found to correspond to components of the T2SS and of other systems. The T2SS components co-localised much closer than the imposed distance threshold ($d \leq 5$, Fig. S1b). The vast majority (99%) of the T4P were encoded in multiple distant loci, which is accepted but not required by the model whereas most T2SS were encoded in one single locus (96.5%). To verify that the T2SS, T4P, and Tad loci were correctly classed, we compared the HMMER scores of proteins matched by protein profiles from different systems. Proteins matching profiles from two types of systems scored systematically higher for the system in which they were classed even though this is not a classification criterion in the model (Fig. 2b-c). This suggests that our classification is accurate.

We detected 400 T2SS in 360 genomes, 379 T4P in 377 genomes, and 425 Tad pili in 323 genomes. The high abundance of Tad pili is surprising given that they are much less studied than the other systems. Interestingly, we found one Tad pilus with the outer membrane channel (the secretin) in one of the rare Firmicutes with an outer membrane (Clostridia, *Acetohalobium arabaticum* DSM 5501)⁵⁵, and also in Acidobacteria, Chlorobi, and Nitrospirae. T4P, T2SS, and to a lesser extent Tad pili, were usually found in a single copy per genome, but some genomes encoded up to three systems (Fig. 2d). The observed small number of T2SS per genome reinforces previous suggestions that many T2SS might secrete several different proteins⁵⁶.

T3SS and T4SS

T3SS and T4SS are capable of secreting proteins directly into other cells. The T3SS, sometimes also termed non-flagellar T3SS or NF-T3SS, evolved from the flagellar secretion system and is encoded by 15 to 25 genes usually in a single locus^{24,57,58} (Fig. 3a). Many of the core components of this system are homologous to the distinct F-T3SS that is part of the bacterial flagellum⁵⁹⁻⁶¹. We have previously proposed models that accurately discriminate between the T3SS and the flagellum²⁴. We used the same models in this work. We identified 434 NF-T3SS in 334 genomes and 837 flagella in 762 genomes. Some genomes encode many T3SS, e.g., *Burkholderia thailandensis* MSMB121 encodes 4 T3SS. These results match experimental data showing that in *Burkholderia pseudomallei* the multiple T3SS target different types of cells⁶², and that in *Salmonella enterica* the two T3SS are expressed at different moments in the infection cycle (reviewed in⁶³). Multiplicity of T3SS is therefore likely to be associated with complex lifestyles.

T4SS are involved in protein secretion, in conjugation and in some cases in DNA release to, or uptake from, the environment⁶⁴. Here, we distinguished the protein secretion T4SS from the conjugation-related T4SS, which requires a relaxase⁶⁵, by naming them respectively pT4SS and cT4SS. It should be noted that some cT4SS are also able to secrete proteins⁶⁴. We have previously built and validated profiles and models for the pT4SS, and cT4SS⁶⁶ (Fig. 3). The latter can be divided in eight sub-types corresponding to different mating pair formation complexes (MPF)³⁰, of which six are found in diderm-LPS bacteria, and only two are known to include pT4SS (MPF_I and MPF_T). To test the specificity of the models of each T4SS sub-type, we studied the close co-occurrence of T4SS components. The results show that most protein profiles are highly specific to each T4SS sub-type (Fig. S2). Hence, our profiles are able to identify and

distinguish between these different systems. We identified 156 pT4SS (among 990 T4SS) in 130 genomes of diderm bacteria (Table S3).

T5SS

T5SS are divided in five types (reviewed in ⁶⁷⁻⁷⁰). Four types encode the translocator (pore-forming) and the passenger (secreted) domains in a single gene: the classical autotransporter (T5aSS), the trimeric autotransporter (T5cSS), the inverted autotransporter (T5eSS), and the fused two-partner system (T5dSS). In two-partner systems (T5bSS), the translocator and passenger are encoded in two separate (typically contiguous) genes. T5SS rely on the Sec machinery for inner-membrane translocation and require other cellular functions for biogenesis. Many of these functions are ubiquitous in diderm-LPS bacteria and do not facilitate the identification of T5SS. Hence, our models only included information on the conserved, mandatory translocator domain of T5SS (Fig. 4, Fig. S3). Two recently proposed families of T5SS - T5dSS and T5eSS ^{71,72} - were not matched by the T5SS profiles. We will build specific profiles for the detection of these sub-types when enough experimentally validated examples become available.

Our models were able to identify all T5SS in the *reference* and *validation* datasets, with the exception of an atypical previously described T5bSS of *Pseudomonas aeruginosa* consisting of a translocator domain fused with a component of the chaperone usher pathway ⁷³. We found 3,829 T5aSS in the genomes of diderm bacteria, which makes them by far the most abundant secretion system in our dataset. Certain *Chlamydiae* genomes contain up to 21 T5aSS. We found 1,125 T5bSS (0-8 per genome) and 849 T5cSS (0-24 per genome). T5SS were encoded in 62% of the genomes of diderm bacteria.

T6SS

T6SS secrete effectors to bacterial or eukaryotic cells. They were recently divided in three sub-types⁴⁰, among which T6SSⁱ is by far the most studied⁷⁴⁻⁸⁰. This sub-type has more than a dozen components^{77,81}. We built profiles for 14 conserved protein families (Fig. 5a, Materials and Methods, Tables S1, S4, S5), of which 13 were previously described as the most conserved components of the T6SSⁱ²⁰. The remaining profile corresponds to the PAAR-repeat-containing EvpJ protein family of the spike complex⁸², present in eight out of the nine T6SSⁱ in the reference dataset. Using this model we identified all T6SSⁱ of the reference and of the validation datasets. This suggests that the T6SSⁱ model is very accurate. We only found an inaccuracy in *Escherichia coli* O42 where two systems are adjacent in the genome and were identified as a single system. Part of the T6SSⁱ machinery is structurally homologous to the puncturing device of phages, from which it may have originated⁸³. Yet, our model did not identify a T6SSⁱ in any of the 998 phages present in GenBank, showing that it does not mistake puncturing devices for components of the T6SS.

We identified 652 T6SSⁱ in 409 bacterial genomes, with up to 6 T6SSⁱ per genome in some *Burkholderia pseudomallei* strains. Around 9% of the T6SSⁱ are encoded in multiple loci in the genome. Interestingly, 35% of the replicons encoding a T6SSⁱ encode TssI (VgrG) away from the main loci, with either a PAAR-containing component (EvpJ) or the chaperone TssD (Hcp), or both (Fig. 5b-d). PAAR-motifs promote the physical interaction between VgrG and toxins, which are often encoded in the same locus^{80,82,84}. It has recently been proposed that VgrG might also be involved in toxin export in a T6SS-independent way⁸⁰. Genomes lacking T6SSⁱ do carry some of these small *tssI*-associated clusters, although this corresponds to only 8% of the clusters we could identify. Hence, the identification of loci encoding TssI might uncover new T6SSⁱ effectors.

269 The T6SSⁱⁱ sub-type described in *Francisella tularensis*, is involved in subversion of
 270 the immune system (growth in macrophages) and virulence^{39,85-87}. Three of the
 271 components of the T6SSⁱⁱ were seldom matched by T6SSⁱ profiles (*tssBCL*), complicating
 272 the detection of T6SSⁱⁱ with the T6SSⁱ model. We built 17 protein profiles and made a
 273 specific model for T6SSⁱⁱ based on a *Francisella tularensis* system (see Fig. 5, Materials
 274 and Methods and Table S5)^{87,88}. Using HHsearch⁸⁹, we confirmed distant homology
 275 between the proteins encoded by *tssBCIL* and T6SSⁱ and/or T6SSⁱⁱⁱ components (p-
 276 value<0.001). The model detected 30 T6SSⁱⁱ in bacterial genomes. All instances were
 277 identified exclusively within the 18 genomes of *Francisella*, and all genomes on the genus
 278 contained at least one system.

279 A recent report identified T6SSⁱⁱⁱ in *Flavobacterium johnsoniae* and *Bacteroides*
 280 *fragilis* and showed it to be involved in bacterial competition⁴⁰. This sub-type included 9
 281 homologs of the 13 described core components of T6SSⁱ and lacked homologs of the
 282 "trans-envelope subcomplex" (Fig. 5). Furthermore, it had three specific components
 283 (TssN, TssO and TssP). We built 13 protein profiles based on the analysis of the reported
 284 three loci, including the nine homologs to core components of T6SSⁱ, the nearly core
 285 T6SSⁱ component EvpJ, and two specific components TssN and TssP. We could not build
 286 a protein profile for TssO because of the lack of representative sequences similar enough
 287 to constitute a protein family (Table S5). The parameters for the models were inferred from
 288 the analysis of co-localised hits of the T6SSⁱⁱⁱ components protein profiles (Fig. S4). We
 289 identified 20 T6SSⁱⁱⁱ in 18 of the 97 Bacteroidetes genomes. TssQ was not previously
 290 recognized as conserved but was found in 50% of the systems identified in genomes. It
 291 had no PFAM hit, but InterProScan predicted the presence of one secretion signal and cell
 292 localisation at the outer-membrane^{90,91}. Interestingly, we could find occurrences of EvpJ
 293 (harbouring a PAAR domain) within 6 of the 20 T6SSⁱⁱⁱ main loci, and outside of the main

locus in 6 genomes with a T6SSⁱⁱⁱ. This component co-localised with TssD (Hcp), TssE, or TssD and TssI (VgrG). This suggests it might have similar roles in both T6SSⁱ and T6SSⁱⁱⁱ. T6SSⁱⁱⁱ was only identified among Bacteroidetes.

T9SS

A novel protein secretion system, T9SS or PorSS, has been described in *F. johnsoniae* and *Porphyromonas gingivalis*^{9,92}. It is required for the secretion of components of the gliding motility apparatus, adhesins and various hydrolytic enzymes. We used eight protein profiles from TIGRFAM and PFAM for five components (some having several profiles), and built protein profiles for five other components (Tables S4-S5, Materials and Methods). One of the profiles was not specifically associated with T9SS; it is part of the gliding motility machinery (GldJ). It was included in the model because it clusters with some of the T9SS components and thus facilitates their identification. Hence, the model includes 13 protein profiles for 10 core T9SS components⁹² (Fig. 6, Fig. S5). Four components of the T9SS are scattered in the chromosome, whereas the others are encoded in two gene clusters. We detected 60 T9SS in 60 of the 97 genomes of Bacteroidetes, and none in other clades, as previously shown¹⁰. T9SS were found in 62% of the species of Bacteroidetes.

Distribution of secretion systems

To the best of our knowledge, this is the first report comparing the frequency of all well-known protein secretion systems of diderm-LPS bacteria in bacterial genomes. Therefore, we analysed the distribution of these protein secretion systems in relation to bacterial phylogeny, including clades with more than four genomes and with reliable information on their phylogenetic position (Fig. 7). Only three clades, Alpha-, Beta- and

317 Gamma-Proteobacteria, encoded all the six most-studied protein secretion systems (T1-
318 T6SSⁱ). Delta- and Epsilon-Proteobacteria showed fewer or no T2SS, T3SS and pT4SS.
319 Most other clades encode fewer types of systems. The distributions of T3SS, T4SS, T6SS,
320 and T9SS have been described recently ^{10,20,24,30,40}, so we shall focus our analysis on the
321 other systems and on their relative distribution.

322 T1SS and T5SS are the most widespread protein secretion systems (Fig. 7, see
323 below). We identified T1SS in phyla as diverse as Spirochaetes, Planctomycetes,
324 Aquificae, Bacteroidetes, and Cyanobacteria. T1SS involved in the secretion of glycolipids
325 for heterocysts formation were recently described in filamentous Cyanobacteria ^{93,94}. We
326 find that T1SS are particularly abundant in this clade, where we identified 156 systems in
327 72 genomes, with 75% of the genomes harbouring at least one T1SS. The three types of
328 T5SS showed similar taxonomic distributions, even if T5cSS are less widespread (Fig. 7).
329 Some phyla include only one type of T5SS: T5aSS in Thermodesulfobacteria and T5bSS
330 in Aquificae and Deinococcus-Thermus. These clades have too few representative
331 genomes to conclude if they lack the other T5SS.

332 We identified very few T2SS outside Proteobacteria. T2SS were also absent from the
333 98 genomes of Epsilon-proteobacteria. We did find a T2SS in a non-Proteobacteria,
334 *Desulfurispirillum indicum* S5, a free-living spiral-shaped aquatic Chrysiogenetes (also
335 encoding a T1SS). We could not find a description of the membrane architecture for this
336 species, but our analysis reinforces previous suggestions that it is a diderm ⁹⁵. Putative
337 T2SS were previously identified in clades where we failed to identify complete systems:
338 *Synechococcus elongatus* (Cyanobacteria), *Chlamydia trachomatis* (Chlamydiae) and
339 *Leptospira interrogans* (Spirochaetes) ^{28,96-98}. The cyanobacterial system, which has a role
340 in protein secretion and biofilm formation, seems to be a typical T4P encoded in multiple
341 loci. The role of T4P in secreting proteins that are not part of its structure has been

described before⁹⁹. To the best of our knowledge, the function of the *Leptospira* system was not experimentally tested. The Chlamydiae system was indeed associated with protein secretion⁹⁷. From the point of view of our models the putative T2SS from these two last clades form incomplete systems (although they could be retrieved by lowering the minimum required number of components for a valid T2SS in the model). Preliminary phylogenetic analyses did not allow conclusive assignment of these systems to T2SS or to T4P. Further experimental and computational work will be necessary for their precise characterisation.

Secretion systems and the cell envelope

The distribution of secretion systems is linked with the cell envelope. Expectedly, all genomes of monoderms lacked loci encoding diderm-like protein secretion system. Several clades of diderm bacteria have few types of protein secretion systems, but only one lacks them all: the Thermotogae. These bacteria are thermophilic, and one could speculate that high temperatures could be incompatible with protein secretion systems of mesophiles. Yet, life under high temperatures is also typical of the sister-clade Aquificae, which has T1SS and T5SS. Instead, the lack of typical protein secretion systems in Thermotogae might be caused by the peculiar sheath-like structure present in their outer cell envelope, the "toga"¹⁰⁰. This may have led to the evolution of different secretion systems. Accordingly, only a few porins have been identified so far in Thermotogae¹⁰¹. In an analogous way, *Mycobacteria* (Actinobacteria), which have a peculiar mycolate outer membrane, have specific secretion systems¹².

The cell envelope architecture is a key determinant of the evolution of secretion systems in other clades as well. The extracellular structures of T3SS are tightly linked with the type of eukaryote cell (plant vs. animal) with which the bacterium interacts²⁴. Sub-

types of T4SS are also specific to certain bacterial cell envelopes³⁰. Interestingly, some diderm bacteria from taxa dominated by monoderms have protein secretion systems homologous to those of Proteobacteria (including Clostridia, Cyanobacteria, Fusobacteria and Negativicutes). For example, we identified in Negativicutes (a clade of Firmicutes) putative pT4SS and the three types of T5SS. Some genomes of Halanaerobiales (a sub-clade of Clostridia, Firmicutes) encode T1SS and T5bSS. Similarities in the cell envelope may thus lead to the presence of similar systems in very distant bacteria.

Conclusion

Using our models, we were able to identify nearly all protein secretion systems in both the *reference* and the *validation* datasets. The few systems we have missed are very atypical (such as the scattered T2SS of *Legionella*) or have homologous components in other types of systems (several T2SS). In the latter case, the relaxation of the parameters of the T2SS model allowed their identification. We emphasize that our models are publicly available and can be modified by the user to increase sensitivity, for example by simply relaxing the parameters for components detection (HMMER i-evalue and profile coverage), or by altering the parameters of the model e.g., by decreasing the minimal number of components required for a valid system or by relaxing the co-localisation criterion. Nevertheless, this will inevitably lead to an increased number of false positives. Complementary analyses can also facilitate the identification of systems. For example, the results of Fig. 2b-c show that the comparison of the scores of the hits of the protein profiles discriminate between homologous components in different systems, even when the information on genetic organization is lacking or insufficient. This is one of the advantages of using specifically designed protein profiles, instead of generic profiles as can be found in PFAM: the system-specific profiles distinguish between homologs components in

different types of molecular systems. Comparisons of the scores for homologous proteins, or phylogenetic analyses could also be useful to analyse poorly assembled genomic or metagenomic data where co-localisation data is unavailable.

Rapid evolution of extracellular components of protein secretion systems and the paucity of experimentally studied systems outside Proteobacteria may have led us to under-estimate their presence in less studied clades. Yet, several pieces of evidence suggest that we may have not missed many homologous systems. 1) In most cases we identified all known systems in our *reference* and *validation* datasets. 2) We identified at least one type of secretion system in almost all clades of diderm bacteria. 3) We identified components of T4P and Tad (homologous to T2SS), F-T3SS (homologous to NF-T3SS), and cT4SS (homologous to pT4SS) with certain profiles for the protein secretion systems in many clades, including monoderms (Table S3). If one is able to pinpoint systems that are so distantly related, one should have been able to identify most systems of a given type. Instead, it is tempting to speculate that in clades other than Proteobacteria there are alternative and currently unknown protein secretion systems. Interestingly, the recently discovered T6SSⁱⁱⁱ and T9SS are restricted to Bacteroidetes in our dataset^{9,40}, while T6SSⁱⁱ are only found in *Francisella*. The absence of well-known protein secretion systems in certain taxa and the presence of lineage-specific systems are consistent with the hypothesis that novel systems remain to be discovered.

409 **Materials and Methods**

410 **Data**

411 The genomes of bacteria (2,484) and archaea (159) were downloaded from NCBI
412 RefSeq (<ftp://ftp.ncbi.nih.gov/genomes/>, November 2013). From these, we sorted out
413 1,528 genomes of bacteria classed as diderm in the literature ¹⁰²⁻¹⁰⁴ (Table S3). A total of
414 998 genomes of phages were downloaded from Genbank (last access, February 2013).
415 The sequences of the reference protein secretion systems were retrieved from Genbank or
416 from complete genomes (Tables S1 and S2).

417 **Systems definition and identification**

418 We built a dataset of experimentally studied secretion systems (T1SS-T6SS, T9SS)
419 and related assemblages (Tad, T4P and the bacterial flagellum) from the analysis of
420 published data. This reference dataset was used to build the models of each type of
421 system (Table S1) using MacSyFinder. This software is publicly available and was
422 described in detail previously ²⁵. Here, we focus on the features of the program that are
423 pertinent for the secretion systems models. A model in MacSyFinder defines the
424 components of the secretion system, the minimal acceptable number of components and
425 their genetic organisation. Among other things (see <http://macsyfinder.readthedocs.org> for
426 full documentation), one can specify the following relevant information. 1) Systems are
427 encoded in a single locus (*single-locus* system) or in several loci (*scattered* system). 2)
428 Core components (ubiquitous and essential) are defined as *mandatory*. 3) Components
429 that are accessory or poorly conserved in sequence are defined as *accessory*. These
430 components are accessory for the computational model, but their function may be
431 essential. This happens when different proteins have analogous functions or when

proteins evolve so fast that distant homologs are not recognisable. 4) Some genes are ubiquitous and specific to a system and can be defined as forbidden in models of other systems. This facilitates discrimination between systems with homologous components. For example, the NF-T3SS-specific secretin may be declared as forbidden in the F-T3SS. 5) An occurrence of a system is validated when a pre-defined number (*quorum*) of mandatory components and/or sum of mandatory and accessory components is found²⁵. 6) Components can be defined as reciprocally *exchangeable* in the quorum (which prevents them from being counted twice). 7) Two components are co-localised when they are separated by less than a given number of genes (parameter $d=inter_gene_max_space$). 8) A component defined with the *loner* attribute does not need to be co-localised with other components to be part of a system (e.g., OMF in T1SS). 9) A component that can participate in several instances of a system (e.g., OMF in T1SS) receives the *multi_system* attribute. These different properties can be combined when necessary.

The models for the different protein secretion systems were described in files following a dedicated XML grammar²⁵ and named after the system (e.g., T1SS.xml, File S1). Models can be easily modified on the standalone version of MacSyFinder. The webserver allows the modification of the most important search parameters.

MacSyFinder was used to identify protein secretion systems in bacterial genomes in several steps (for a full description of the software see²⁵). The program did the search in three steps. Firstly, components were identified using protein profile searches with HMMER⁴¹. The hits were kept when the alignments covered more than 50% of the protein profile and the $i\text{-value} < 10^{-3}$ (default parameters). Secondly, the components were clustered according to their proximity in the genome using the parameter d . Finally, the clusters were validated if they passed the criteria in the models of the systems.

Definition of protein profiles

The models included 204 protein profiles. Among these, the two profiles for T5aSS and T5cSS were extracted from PFAM^{67,90}, and eight profiles for T9SS were extracted from PFAM or TIGRFAM^{90,105}. The remaining 194 were the result of our previous work^{24,66,106} or this study (84 protein profiles for T1SS, T2SS, Tad, type IV pilus, T5bSS, T6SSⁱ, T6SSⁱⁱ, T6SSⁱⁱⁱ and T9SS, listed in Table S4). To build these profiles we sampled the experimentally studied systems for proteins representative of each component of each system. We selected examples within these to maximise sequence diversity (Table S1). Protein families were constructed by clustering homologous proteins. The details of the methods and parameters used to build each protein profile are described in Table S5. In the case of the T9SS, where only two systems were experimentally characterised, we used components from the well-studied system of *F. johnsoniae* (or *P. gingivalis* when the gene was absent from *F. johnsoniae*) for Blastp searches against our database of complete genomes, and retained the best sequence hits (e-value < 10⁻²⁰) to constitute protein families. A similar approach was taken to build protein profiles for the T6SSⁱⁱ, based on the *Francisella tularensis subsp. tularensis* SCHU S4 FPI system displayed in Table 1 of³⁹. The largest families obtained were aligned and manually curated to produce hidden Markov model profiles with HMMER 3.0⁴¹.

Availability

Detection and visualization of all systems described in this paper can be performed online on the Mobyly-based¹⁰⁷ webserver TXSScan: <http://mobyly.pasteur.fr/cgi-bin/portal.py#forms::txsscan>. Detection can also be performed locally using the standalone program MacSyFinder²⁵, and the sets of models and profiles described here. MacSyFinder is freely available for all platforms at <https://github.com/gem>

481 [pasteur/macsyfinder](https://research.pasteur.fr/en/tool/txsscan-models-and-profiles-for-protein-secretion-systems). Models and required protein profiles are available as supplemental
 482 material (File S1) at [https://research.pasteur.fr/en/tool/txsscan-models-and-profiles-for-](https://research.pasteur.fr/en/tool/txsscan-models-and-profiles-for-protein-secretion-systems)
 483 [protein-secretion-systems](https://research.pasteur.fr/en/tool/txsscan-models-and-profiles-for-protein-secretion-systems) ([https://researchpullzone-yhello.netdna-ssl.com/wp-](https://researchpullzone-yhello.netdna-ssl.com/wp-content/uploads/2015/10/research.pasteur.fr_txsscan-models-and-profiles-for-protein-secretion-systems.gz)
 484 [content/uploads/2015/10/research.pasteur.fr_txsscan-models-and-profiles-for-protein-](https://research.pasteur.fr/en/tool/txsscan-models-and-profiles-for-protein-secretion-systems.gz)
 485 [secretion-systems.gz](https://research.pasteur.fr/en/tool/txsscan-models-and-profiles-for-protein-secretion-systems.gz)) The models are provided as simple text (XML) files, so they can be
 486 easily modified and extended by the user. The results of MacSyFinder can be visualized
 487 with MacSyView, that is available online at <http://macsyview.web.pasteur.fr> or for
 488 download at <https://github.com/gem-pasteur/macsyview> (also included in the release of
 489 MacSyFinder).

Acknowledgments

We are grateful to Elie Dassa, Olivera Francetic, and Marc Garcia-Garcerà for fruitful discussions, and Hervé Ménager for its contribution to MacSyView. We thank Eric Duchaud for discussions and comments on T9SS. This work was supported by the CNRS, the Institut Pasteur and the European Research Council (grant EVOMOBILOME, number 281605). JC is a member of the French doctoral school "Ecole Doctorale Interdisciplinaire Européenne Frontières du Vivant ED474".

Competing financial interests

The authors declare no competing financial interests.

Authors' contributions

SSA and EPCR designed the analyses. SSA designed the secretion systems models and profiles, and performed the analyses. JC and JG contributed to the T4SS models, the T4SS protein profiles and the corresponding analyses. BN contributed to the MacSyFinder and MacSyView software, and created the online interface for TXSScan and MacSyView. MT contributed to the T9SS models and the T9SS protein profiles. SSA and EPCR wrote the manuscript with the help of the other authors. All authors read and approved the full manuscript.

Abbreviations

ESX – ESAT-6 secretion system

HMM – hidden Markov model

LPS – Lipopolysaccharide

OMF – outer-membrane porin of T1SS

- 512 MFP – membrane-fusion protein from T1SS.
- 513 MFS – major facilitator superfamily
- 514 MPF – mating pair formation complex
- 515 RND – resistance nodulation cell division
- 516 T1SS – Type I secretion system
- 517 T2SS – Type II secretion system
- 518 T3SS – Type III secretion system
- 519 T4P – Type IV pilus
- 520 T4SS – Type IV secretion system
- 521 cT4SS – conjugative T4SS
- 522 pT4SS – protein delivery T4SS
- 523 T5SS – Type V secretion system
- 524 T6SS – Type VI secretion system
- 525 T9SS – Type IX secretion system
- 526 Tad – Tight adherence
- 527 XML – Extensible Markup Language

References

1. Wandersman C & Delepelaire P. Bacterial iron sources: from siderophores to hemophores. *Annu. Rev. Microbiol.* **58**, 611 (2004).
2. Ruhe ZC, Low DA & Hayes CS. Bacterial contact-dependent growth inhibition. *Trends Microbiol.* **21**, 230 (2013).
3. Viprey V, Del Greco A, Golinowski W, Broughton WJ & Perret X. Symbiotic implications of type III protein secretion machinery in Rhizobium. *Mol. Microbiol.* **28**, 1381 (1998).
4. Ma W & Guttman DS. Evolution of prokaryotic and eukaryotic virulence effectors. *Curr. Opin. Plant Biol.* **11**, 412 (2008).
5. Raymond B *et al.* Subversion of trafficking, apoptosis, and innate immunity by type III secretion system effectors. *Trends Microbiol.* **21**, 430 (2013).
6. Bleves S *et al.* Protein secretion systems in Pseudomonas aeruginosa: A wealth of pathogenic weapons. *Int. J. Med. Microbiol.* **300**, 534 (2010).
7. Dalbey RE & Kuhn A. Protein traffic in Gram-negative bacteria--how exported and secreted proteins find their way. *FEMS Microbiol. Rev.* **36**, 1023 (2012).
8. Chang JH, Desveaux D & Creason AL. The ABCs and 123s of bacterial secretion systems in plant pathogenesis. *Annu Rev Phytopathol* **52**, 317 (2014).
9. Sato K *et al.* A protein secretion system linked to bacteroidete gliding motility and pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 276 (2010).
10. McBride MJ & Zhu Y. Gliding motility and Por secretion system genes are widespread among members of the phylum bacteroidetes. *J. Bacteriol.* **195**, 270 (2013).
11. Desvaux M, Hebraud M, Talon R & Henderson IR. Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends Microbiol.* **17**, 139 (2009).
12. Abdallah AM *et al.* Type VII secretion--mycobacteria show the way. *Nat. Rev. Microbiol.* **5**, 883 (2007).
13. Tseng TT, Tyler BM & Setubal JC. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiol.* **9 Suppl 1**, S2 (2009).
14. Planet PJ, Kachlany SC, DeSalle R & Figurski DH. Phylogeny of genes for secretion NTPases: identification of the widespread tadA subfamily and development of a diagnostic key for gene classification. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 2503 (2001).
15. Minamino T & Namba K. Self-assembly and type III protein export of the bacterial flagellum. *J. Mol. Microbiol. Biotechnol.* **7**, 5 (2004).
16. Pelicic V. Type IV pili: e pluribus unum? *Mol. Microbiol.* **68**, 827 (2008).
17. Peabody CR *et al.* Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. *Microbiology* **149**, 3051 (2003).

- 563 18. Nogueira T, Touchon M & Rocha EP. Rapid evolution of the sequences and gene repertoires of
564 secreted proteins in bacteria. *PLoS ONE* **7**, e49403 (2012).
- 565 19. Gophna U, Ron EZ & Graur D. Bacterial type III secretion systems are ancient and evolved by
566 multiple horizontal-transfer events. *Gene* **312**, 151 (2003).
- 567 20. Boyer F, Fichant G, Berthod J, Vandenbrouck Y & Attree I. Dissecting the bacterial type VI secretion
568 system by a genome wide in silico analysis: what can be learned from available microbial
569 genomic resources? *BMC Genomics* **10**, 104 (2009).
- 570 21. Ren CP *et al.* The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia*
571 *coli*, is present in the majority of strains but has undergone widespread mutational attrition. *J.*
572 *Bacteriol.* **186**, 3547 (2004).
- 573 22. Huynen M, Snel B, Lathe W, 3rd & Bork P. Predicting protein function by genomic context:
574 quantitative evaluation and qualitative inferences. *Genome Res.* **10**, 1204 (2000).
- 575 23. Wolf YI, Rogozin IB, Kondrashov AS & Koonin EV. Genome alignment, evolution of prokaryotic
576 genome organization, and prediction of gene function using genomic context. *Genome Res.* **11**,
577 356 (2001).
- 578 24. Abby SS & Rocha EP. The non-flagellar type III secretion system evolved from the bacterial
579 flagellum and diversified into host-cell adapted systems. *PLoS Genet.* **8**, e1002983 (2012).
- 580 25. Abby SS, Neron B, Menager H, Touchon M & Rocha EP. MacSyFinder: A Program to Mine Genomes
581 for Molecular Systems with an Application to CRISPR-Cas Systems. *PLoS ONE* **9**, e110726
582 (2014).
- 583 26. Yen MR *et al.* Protein-translocating outer membrane porins of Gram-negative bacteria. *Biochim.*
584 *Biophys. Acta* **1562**, 6 (2002).
- 585 27. Pallen MJ, Beatson SA & Bailey CM. Bioinformatics, genomics and evolution of non-flagellar type-III
586 secretion systems: a Darwinian perspective. *FEMS Microbiol. Rev.* **29**, 201 (2005).
- 587 28. Cianciotto NP. Type II secretion: a protein secretion system for all seasons. *Trends Microbiol.* **13**,
588 581 (2005).
- 589 29. Saier MH, Ma CH, Rodgers L, Tamang DG & Yen MR. Protein secretion and membrane insertion
590 systems in bacteria and eukaryotic organelles. *Adv. Appl. Microbiol.* **65**, 141 (2008).
- 591 30. Guglielmini J, de la Cruz F & Rocha EP. Evolution of conjugation and type IV secretion systems. *Mol.*
592 *Biol. Evol.* **30**, 315 (2013).
- 593 31. Barret M, Egan F & O'Gara F. Distribution and diversity of bacterial secretion systems across
594 metagenomic datasets. *Environmental microbiology reports* **5**, 117 (2013).
- 595 32. Delepelaire P. Type I secretion in gram-negative bacteria. *Biochim. Biophys. Acta* **1694**, 149 (2004).
- 596 33. Bouige P, Laurent D, Piloyan L & Dassa E. Phylogenetic and functional classification of ATP-binding
597 cassette (ABC) systems. *Curr. Protein Pept. Sci.* **3**, 541 (2002).
- 598 34. Dassa E & Bouige P. The ABC of ABCS: a phylogenetic and functional classification of ABC systems
599 in living organisms. *Res. Microbiol.* **152**, 211 (2001).

35. Jacob-Dubuisson F, Fernandez R & Coutte L. Protein secretion through autotransporter and two-partner pathways. *Biochim. Biophys. Acta* **1694**, 235 (2004).
36. Henderson IR, Navarro-Garcia F, Desvaux M, Fernandez RC & Ala'Aldeen D. Type V protein secretion pathway: the autotransporter story. *Microbiol. Mol. Biol. Rev.* **68**, 692 (2004).
37. Linke D, Riess T, Autenrieth IB, Lupas A & Kempf VA. Trimeric autotransporter adhesins: variable structure, common function. *Trends Microbiol.* **14**, 264 (2006).
38. Tomich M, Planet PJ & Figurski DH. The tad locus: postcards from the widespread colonization island. *Nat. Rev. Microbiol.* **5**, 363 (2007).
39. Broms JE, Sjostedt A & Lavander M. The Role of the Francisella Tularensis Pathogenicity Island in Type VI Secretion, Intracellular Survival, and Modulation of Host Cell Signaling. *Front. Microbiol.* **1**, 136 (2010).
40. Russell AB *et al.* A Type VI Secretion-Related Pathway in Bacteroidetes Mediates Interbacterial Antagonism. *Cell Host Microbe* **16**, 227 (2014).
41. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
42. Eddy SR. Profile hidden Markov models. *Bioinformatics* **14**, 755 (1998).
43. Holland IB, Schmitt L & Young J. Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway. *Mol. Membr. Biol.* **22**, 29 (2005).
44. Kanonenberg K, Schwarz CK & Schmitt L. Type I secretion systems - a story of appendices. *Res. Microbiol.* **164**, 596 (2013).
45. Paulsen IT, Park JH, Choi PS & Saier MH, Jr. A family of gram-negative bacterial outer membrane factors that function in the export of proteins, carbohydrates, drugs and heavy metals from gram-negative bacteria. *FEMS Microbiol. Lett.* **156**, 1 (1997).
46. Dinh T, Paulsen IT & Saier MH, Jr. A family of extracytoplasmic proteins that allow transport of large molecules across the outer membranes of gram-negative bacteria. *J. Bacteriol.* **176**, 3825 (1994).
47. Dassa E. Natural history of ABC systems: not only transporters. *Essays Biochem.* **50**, 19 (2011).
48. Davidson AL, Dassa E, Orelle C & Chen J. Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol. Mol. Biol. Rev.* **72**, 317 (2008).
49. Koronakis V, Eswaran J & Hughes C. Structure and function of TolC: the bacterial exit duct for proteins and drugs. *Annu. Rev. Biochem.* **73**, 467 (2004).
50. Burland V *et al.* The complete DNA sequence and analysis of the large virulence plasmid of Escherichia coli O157:H7. *Nucleic Acids Res.* **26**, 4196 (1998).
51. Korotkov KV, Sandkvist M & Hol WG. The type II secretion system: biogenesis, molecular architecture and mechanism. *Nat. Rev. Microbiol.* **10**, 336 (2012).
52. Nivaskumar M & Francetic O. Type II secretion system: a magic beanstalk or a protein escalator. *Biochim. Biophys. Acta* **1843**, 1568 (2014).

- 636 53. Karaba SM, White RC & Cianciotto NP. *Stenotrophomonas maltophilia* Encodes a Type II Protein
637 Secretion System That Promotes Detrimental Effects on Lung Epithelial Cells. *Infect. Immun.*
638 **81**, 3210 (2013).
- 639 54. Cianciotto NP. Many substrates and functions of type II secretion: lessons learned from *Legionella*
640 *pneumophila*. *Future Microbiol.* **4**, 797 (2009).
- 641 55. Zhilina TN & Zavarzin GA. Extremely halophilic, methylotrophic, anaerobic bacteria. *FEMS*
642 *Microbiol. Lett.* **87**, 315 (1990).
- 643 56. Rondelet A & Condemine G. Type II secretion: the substrates that won't go away. *Res. Microbiol.*
644 **164**, 556 (2013).
- 645 57. Galan JE & Wolf-Watz H. Protein delivery into eukaryotic cells by type III secretion machines.
646 *Nature* **444**, 567 (2006).
- 647 58. Cornelis GR. The type III secretion injectisome. *Nat. Rev. Microbiol.* **4**, 811 (2006).
- 648 59. Ginocchio CC, Olmsted SB, Wells CL & Galan JE. Contact with epithelial cells induces the formation
649 of surface appendages on *Salmonella typhimurium*. *Cell* **76**, 717 (1994).
- 650 60. Van Gijsegem F *et al.* The *hrp* gene locus of *Pseudomonas solanacearum*, which controls the
651 production of a type III secretion system, encodes eight proteins related to components of the
652 bacterial flagellar biogenesis complex. *Mol. Microbiol.* **15**, 1095 (1995).
- 653 61. Young GM, Schmiel DH & Miller VL. A new pathway for the secretion of virulence factors by
654 bacteria: the flagellar export apparatus functions as a protein-secretion system. *Proc. Natl.*
655 *Acad. Sci. U. S. A.* **96**, 6456 (1999).
- 656 62. Sun GW & Gan YH. Unraveling type III secretion systems in the highly versatile *Burkholderia*
657 *pseudomallei*. *Trends Microbiol.* **18**, 561 (2010).
- 658 63. Hansen-Wester I & Hensel M. *Salmonella* pathogenicity islands encoding type III secretion systems.
659 *Microbes Infect.* **3**, 549 (2001).
- 660 64. Alvarez-Martinez CE & Christie PJ. Biological diversity of prokaryotic type IV secretion systems.
661 *Microbiol. Mol. Biol. Rev.* **73**, 775 (2009).
- 662 65. de la Cruz F, Frost LS, Meyer RJ & Zechner E. Conjugative DNA Metabolism in Gram-negative
663 Bacteria. *FEMS Microbiol. Rev.* **34**, 18 (2010).
- 664 66. Guglielmini J *et al.* Key components of the eight classes of type IV secretion systems involved in
665 bacterial conjugation or protein secretion. *Nucleic Acids Res.* (2014).
- 666 67. Dautin N & Bernstein HD. Protein Secretion in Gram-Negative Bacteria via the Autotransporter
667 Pathway. *Annu. Rev. Microbiol.* **61**, 89 (2007).
- 668 68. Mazar J & Cotter PA. New insight into the molecular mechanisms of two-partner secretion. *Trends*
669 *Microbiol.* **15**, 508 (2007).
- 670 69. Leyton DL, Rossiter AE & Henderson IR. From self sufficiency to dependence: mechanisms and
671 factors important for autotransporter biogenesis. *Nat. Rev. Microbiol.* **10**, 213 (2012).
- 672 70. Leo JC, Grin I & Linke D. Type V secretion: mechanism(s) of autotransport through the bacterial
673 outer membrane. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 1088 (2012).

- 674 71. Salacha R *et al.* The *Pseudomonas aeruginosa* patatin-like protein PlpD is the archetype of a novel
675 Type V secretion system. *Environ. Microbiol.* **12**, 1498 (2010).
- 676 72. Oberhettinger P *et al.* Intimin and invasins export their C-terminus to the bacterial cell surface using
677 an inverse mechanism compared to classical autotransport. *PLoS ONE* **7**, e47069 (2012).
- 678 73. Ruer S, Ball G, Filloux A & de Bentzmann S. The 'P-usher', a novel protein transporter involved in
679 fimbrial assembly and TpsA secretion. *EMBO J.* **27**, 2669 (2008).
- 680 74. Mougous JD *et al.* A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion
681 apparatus. *Science* **312**, 1526 (2006).
- 682 75. Hood RD *et al.* A type VI secretion system of *Pseudomonas aeruginosa* targets a toxin to bacteria.
683 *Cell Host Microbe* **7**, 25 (2010).
- 684 76. Schwarz S *et al.* Burkholderia type VI secretion systems have distinct roles in eukaryotic and
685 bacterial cell interactions. *PLoS Pathog.* **6**, e1001068 (2010).
- 686 77. Silverman JM, Brunet YR, Cascales E & Mougous JD. Structure and regulation of the type VI
687 secretion system. *Annu. Rev. Microbiol.* **66**, 453 (2012).
- 688 78. Basler M, Ho BT & Mekalanos JJ. Tit-for-tat: type VI secretion system counterattack during bacterial
689 cell-cell interactions. *Cell* **152**, 884 (2013).
- 690 79. Brunet YR, Espinosa L, Harchouni S, Mignot T & Cascales E. Imaging type VI secretion-mediated
691 bacterial killing. *Cell reports* **3**, 36 (2013).
- 692 80. Hachani A, Allsopp LP, Oduko Y & Filloux A. The VgrG proteins are "A la carte" delivery systems for
693 bacterial type VI effectors. *J. Biol. Chem.* **289**, 17872 (2014).
- 694 81. Cascales E & Cambillau C. Structural biology of type VI secretion systems. *Philos. Trans. R. Soc. Lond.*
695 *B. Biol. Sci.* **367**, 1102 (2012).
- 696 82. Shneider MM *et al.* PAAR-repeat proteins sharpen and diversify the type VI secretion system spike.
697 *Nature* **500**, 350 (2013).
- 698 83. Pukatzki S, Ma AT, Revel AT, Sturtevant D & Mekalanos JJ. Type VI secretion system translocates a
699 phage tail spike-like protein into target cells where it cross-links actin. *Proc. Natl. Acad. Sci. U.*
700 *S. A.* **104**, 15508 (2007).
- 701 84. Whitney JC *et al.* Genetically distinct pathways guide effector export through the type VI secretion
702 system. *Mol. Microbiol.* **92**, 529 (2014).
- 703 85. Nano FE *et al.* A *Francisella tularensis* pathogenicity island required for intramacrophage growth. *J.*
704 *Bacteriol.* **186**, 6430 (2004).
- 705 86. Ludu JS *et al.* The *Francisella* pathogenicity island protein PdpD is required for full virulence and
706 associates with homologues of the type VI secretion system. *J. Bacteriol.* **190**, 4584 (2008).
- 707 87. Barker JR *et al.* The *Francisella tularensis* pathogenicity island encodes a secretion system that is
708 required for phagosome escape and virulence. *Mol. Microbiol.* **74**, 1459 (2009).
- 709 88. Camacho C *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
- 710 89. Soding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951 (2005).

- 711 90. Finn RD *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281 (2008).
- 712 91. Quevillon E *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116 (2005).
- 713 92. Kharade SS & McBride MJ. Flavobacterium johnsoniae PorV is required for secretion of a subset of
714 proteins targeted to the type IX secretion system. *J. Bacteriol.* **197**, 147 (2015).
- 715 93. Moslavac S *et al.* A TolC-like protein is required for heterocyst development in Anabaena sp. strain
716 PCC 7120. *J. Bacteriol.* **189**, 7887 (2007).
- 717 94. Staron P, Forchhammer K & Maldener I. Structure-function analysis of the ATP-driven glycolipid
718 efflux pump DevBCA reveals complex organization with TolC/HgdD. *FEBS Lett.* **588**, 395
719 (2014).
- 720 95. Rauschenbach I, Yee N, Haggblom MM & Bini E. Energy metabolism and multiple respiratory
721 pathways revealed by genome sequencing of Desulfurispirillum indicum strain S5. *Environ.*
722 *Microbiol.* **13**, 1611 (2011).
- 723 96. Zeng L *et al.* Extracellular proteome analysis of Leptospira interrogans serovar Lai. *Omics : a*
724 *journal of integrative biology* **17**, 527 (2013).
- 725 97. Nguyen BD & Valdivia RH. Virulence determinants in the obligate intracellular pathogen Chlamydia
726 trachomatis revealed by forward genetic approaches. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1263
727 (2012).
- 728 98. Schatz D *et al.* Self-suppression of biofilm formation in the cyanobacterium Synechococcus
729 elongatus. *Environ. Microbiol.* **15**, 1786 (2013).
- 730 99. Hager AJ *et al.* Type IV pili-mediated secretion modulates Francisella virulence. *Mol. Microbiol.* **62**,
731 227 (2006).
- 732 100. Huber R *et al.* Thermotoga maritima sp. nov. represents a new genus of unique extremely
733 thermophilic eubacteria growing up to 90 C. *Arch. Microbiol.* **144**, 324 (1986).
- 734 101. Petrus AK *et al.* Genes for the major structural components of Thermotogales species' togas
735 revealed by proteomic and evolutionary analyses of OmpA and OmpB homologs. *PLoS ONE* **7**,
736 e40236 (2012).
- 737 102. Sutcliffe IC. A phylum level perspective on bacterial cell envelope architecture. *Trends Microbiol.*
738 **18**, 464 (2010).
- 739 103. Francke C *et al.* Comparative analyses imply that the enigmatic Sigma factor 54 is a central
740 controller of the bacterial exterior. *BMC Genomics* **12**, 385 (2011).
- 741 104. Vesth T *et al.* Veillonella, Firmicutes: Microbes disguised as Gram negatives. *Stand Genomic Sci* **9**,
742 431 (2013).
- 743 105. Haft DH *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.* **41**, D387 (2013).
- 744 106. Guglielmini J, Quintais L, Garcillan-Barcia MP, de la Cruz F & Rocha EP. The Repertoire of ICE in
745 Prokaryotes Underscores the Unity, Diversity, and Ubiquity of Conjugation. *PLoS Genet.* **7**,
746 e1002222 (2011).
- 747 107. Neron B *et al.* Mobyle: a new full web bioinformatics framework. *Bioinformatics* **25**, 3005 (2009).

- 748 108. Nunn DN & Lory S. Product of the *Pseudomonas aeruginosa* gene pilD is a prepilin leader
749 peptidase. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 3281 (1991).
- 750 109. Pepe CM, Eklund MW & Strom MS. Cloning of an *Aeromonas hydrophila* type IV pilus biogenesis
751 gene cluster: complementation of pilus assembly functions and characterization of a type IV
752 leader peptidase/N-methyltransferase required for extracellular protein secretion. *Mol.*
753 *Microbiol.* **19**, 857 (1996).
- 754 110. Marsh JW & Taylor RK. Identification of the *Vibrio cholerae* type 4 prepilin peptidase required for
755 cholera toxin secretion and pilus formation. *Mol. Microbiol.* **29**, 1481 (1998).
- 756 111. Christie PJ. Type IV secretion: the *Agrobacterium* VirB/D4 and related conjugation systems.
757 *Biochim. Biophys. Acta* **1694**, 219 (2004).
- 758 112. Nagai H & Kubori T. Type IVB Secretion Systems of *Legionella* and Other Gram-Negative Bacteria.
759 *Front. Microbiol.* **2**, 136 (2011).
- 760 113. Schroder G, Schuelein R, Quebatte M & Dehio C. Conjugative DNA transfer into human cells by the
761 VirB/VirD4 type IV secretion system of the bacterial pathogen *Bartonella henselae*. *Proc. Natl.*
762 *Acad. Sci. U. S. A.* **108**, 14643 (2011).
- 763 114. Abby SS, Tannier E, Gouy M & Daubin V. Lateral gene transfer as a support for the tree of life. *Proc.*
764 *Natl. Acad. Sci. U. S. A.* **109**, 4962 (2012).
- 765 115. Yutin N, Puigbo P, Koonin EV & Wolf YI. Phylogenomics of prokaryotic ribosomal proteins. *PLoS*
766 *ONE* **7**, e36972 (2012).
- 767 116. Wu D *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**,
768 1056 (2009).
- 769 117. Boussau B, Gueguen L & Gouy M. Accounting for horizontal gene transfers explains conflicting
770 hypotheses regarding the position of aquificales in the phylogeny of Bacteria. *BMC Evol. Biol.* **8**,
771 272 (2008).
- 772 118. Souza RC *et al.* AtlasT4SS: a curated database for type IV secretion systems. *BMC Microbiol.* **12**,
773 172 (2012).
- 774 119. Bi D *et al.* SecReT4: a web-based bacterial type IV secretion system resource. *Nucleic Acids Res.*
775 **41**, D660 (2013).
- 776 120. Li J *et al.* SecReT6: a web-based resource for type VI secretion systems found in bacteria. *Environ.*
777 *Microbiol.* **17**, 2196 (2015).
- 778 121. Pundhir S & Kumar A. SSPred: A prediction server based on SVM for the identification and
779 classification of proteins involved in bacterial secretion systems. *Bioinformatics* **6**, 380 (2011).
- 780 122. Martinez-Garcia PM, Ramos C & Rodriguez-Palenzuela P. T346Hunter: a novel web-based tool for
781 the prediction of type III, type IV and type VI secretion systems in bacterial genomes. *PLoS ONE*
782 **10**, e0119317 (2015).
- 783 123. Wang Y, Huang H, Sun M, Zhang Q & Guo D. T3DB: an integrated database for bacterial type III
784 secretion system. *BMC Bioinformatics* **13**, 66 (2012).

785

Figure legends

Figure 1. Model and results for the T1SS.

a. Schema of the structure (left panel) and model of the genetic organisation (right panel) of T1SS. We built protein profiles for the three components and modelled the two possible genetic architectures of the T1SS: one with the three components encoded in a single locus (*inter_gene_max_space* parameter in MacSyFinder: $d \leq 5$ genes), another with the ABC transporter and the MFP encoded in a locus while the OMF is further away (*loner* attribute). A single OMF can also be used by different T1SS⁴⁵ and this is noted by the attribute *multi_system*. **b.** Distribution of hits for the protein profiles of the T1SS components, separated in two groups: hits effectively part of a T1SS main locus (*i.e.*, containing at least ABC and MFP, grey) and hits found elsewhere (“Other”, blue). Even if encoded outside of “main loci” (grey area of the bar), OMF might be involved in T1SS (*loner* property), whereas it is not the case for ABC and MFP. **c.** T1SS encoded in one single locus (ABC, MFP and OMF co-localise) (3) or in two (OMF encoded away from the other components) (2+1).

Figure 2. Model of the T2SS for detection and discrimination from the T4 and Tad pili.

a. Schema of the structure (left panel) of T2SS, and model of its genetic organisation (right panel), indicating components with homologies with T4P and Tad pilus. We built protein profiles for all these components (Tables S4 and S5). Protein families represented by the same colour are homologous, and their profiles often match proteins from the other systems (except for the Flp and TadE/F families that are less similar). Some prepilin peptidases of T2SS and T4P are defined as functionally interchangeable¹⁰⁸⁻¹¹⁰ (curved double-headed arrow, exchangeable attribute). Boxes represent components: mandatory (plain), accessory (dashed) and forbidden (red crosses). **b.** Scores of proteins matched with the profiles of T2SS and T4P. The components of actual T2SS (dark blue) and actual

812 T4P (in red) are well separated, indicating that in each case the best match corresponds to
 813 the profile of the correct model system. The exceptions (blue points surrounded by a black
 814 ellipse) concern the prepilin peptidases (light blue squares, circled in blue), which are
 815 effectively inter-changeable. **c.** Representation similar to **b**, but for the comparison
 816 between T2SS (blue) and Tad (grey) systems. In this case, the separation is perfect: the
 817 proteins always match better the protein profile of the correct system. **d.** Number of
 818 detected systems per genome among the 1,528 genomes of diderm bacteria.

819 Figure 3. Model of T3SS and T4SS.

820 **a.** The models of T3SS and flagellum were built based on a previous study ²⁴
 821 (representation conventions as in Fig. 2). Of the nine mandatory components for the T3SS
 822 only the secretin is forbidden in the model of the flagellum. Conversely, three flagellum-
 823 specific components are forbidden in the T3SS model. Three different types of secretins
 824 are found in T3SS derived from different appendages, which are thus defined as
 825 exchangeable in the model. **b.** Models of the T4SS were built based on a previous study
 826 ⁶⁶. Two different proteins have been described as type 4 coupling proteins (T4CP: VirD4
 827 and TcpA) and two as the major ATPases (VirB4 and TraU, which are homologous). Some
 828 pT4SS lack a T4CP and secrete proteins from the periplasm ⁶⁴. The relaxase (*rel*), is
 829 necessary for conjugation but not for protein secretion, although some relaxase-encoding
 830 T4SS are both cT4SS and pT4SS ¹¹¹⁻¹¹³. Only two MPF types are associated with protein
 831 secretion - pT4SS_I and pT4SS_T, corresponding to MPF_I and MPF_T types. The specificity of
 832 type-specific profiles is assessed in Fig. S2.

833 Figure 4. Model of the T5aSS, T5bSS and T5cSS.

834 The left panel shows simplified schemas of the T5SS, and the right panel displays the
 835 respective genetic model (only one component that is classed as *loner*). The translocator,

836 pore-forming domains were searched using PFAM domains for T5aSS and T5cSS (resp.
837 PF03797 and PF03895), and a profile built for this work for the T5bSS (Tables S1, S4,
838 S5).

839 Figure 5. Model and results for the detection of T6SS.

840 **a.** The left panel shows the schema of the structure of T6SSⁱ, and the right panel displays
841 the genetic model of the three sub-types of T6SS (representation conventions as in Fig.
842 2). For T6SSⁱ, we built profiles for the 14 mandatory components, which were clustered if
843 at a distance of $d \leq 20$ (see Fig. S4). For T6SSⁱⁱ and T6SSⁱⁱⁱ, we built 17 and 13 profiles
844 respectively. All components were set as mandatory, except for TssQ, which is found in
845 half of the T6SSⁱⁱ. Homologies between components that are displayed by the mean of the
846 same colours of boxes between the different sub-types are based on previous studies.
847 *Putative type-specific genes are displayed in grey boxes that do not represent
848 homologies. However, several putative homologies were retrieved using Hhsearch (e-
849 value<1 and p-value<0.05) on T6SSⁱⁱ components: iglC (tssG), iglG (tssF), iglH (tssE), iglJ
850 (tssH) and pdpD (tssH). **b.** Number of different components per cluster of T6SSⁱ. Following
851 this analysis, we set the quorum parameter of T6SSⁱ to 11. **c.** Frequency of hits for each
852 type of T6SSⁱ components in the genomes. Hits matching a single-locus T6SSⁱ are in blue.
853 The other hits match outside the T6SSⁱ loci. **d.** Frequency of each component within
854 single-locus T6SSⁱ. The components EvpJ and TssA were detected in less than 45% of
855 the T6SSⁱ, while the other components are found in most T6SSⁱ loci (>89%).

856 Figure 6. Genetic model of the T9SS.

857 The representation follows the conventions of Fig. 2. The model includes 11 components
858 for which 13 protein profiles were obtained from PFAM (SprA and SprA-2), TIGRFAM
859 (GldJ, GldK, GldL, GldM, GldN and SprA-3) or designed for this study (PorU, PorV, PorQ,

860 SprE, SprT). Four components were declared as *loners*. The co-localisation distance for
 861 the others was set at $d \leq 5$ (see Fig. S5). As several profiles were available for SprA, we
 862 included them all in the models, and declared them as exchangeable homologs in the
 863 model. GldJ is not part of the secretion system, but of the gliding motility system. It was
 864 included in the model as it facilitates the detection of T9SS that co-localises with it.

865 Figure 7. Phylogenetic distribution of protein secretion systems in bacteria.

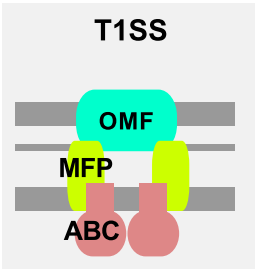
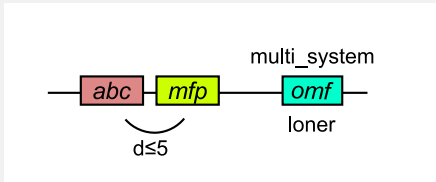
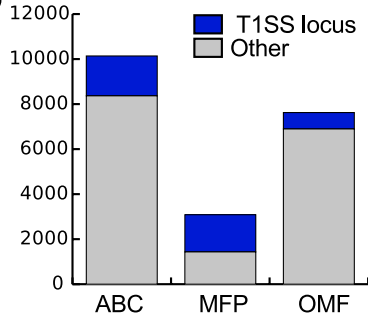
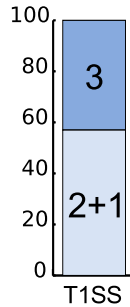
866 Within each clade, the proportion of genomes harbouring each system is indicated in
 867 boxes whose colours follow a gradient from full red (100%) to white (0%) (see legend).
 868 Clades were classed as monoderms (grey or "M" symbol), diderms with LPS-containing
 869 outer membranes (Diderm-LPS in bold, no symbol), diderms with homologs of LPS
 870 pathway that putatively have LPS (D[?]) and diderms with no LPS (D-). The peculiar
 871 envelope of the Thermotogae is indicated (T). The Firmicutes are typically monoderms, but
 872 some of their members are diderms (the Negativicutes, some Clostridia, Mycobacteria).
 873 The bar plot shows the number of detected systems. Bars are split in two categories to
 874 separate on one side Alpha- Beta- and Gamma-proteobacteria, and on the other genomes
 875 from other bacteria. We display the number of occurrences of systems occurring rarely in
 876 our dataset on top of the bars. Clades with less than 4 genomes and/or with unreported
 877 phylogenetic position are not shown (*i.e.*, Chrysiogenetes, Gemmatimonadetes,
 878 Nitrospirae and Thermodesulfobacteria). This sketch tree was drawn from the compilation
 879 of different published phylogenetic analyses ¹¹⁴⁻¹¹⁷.

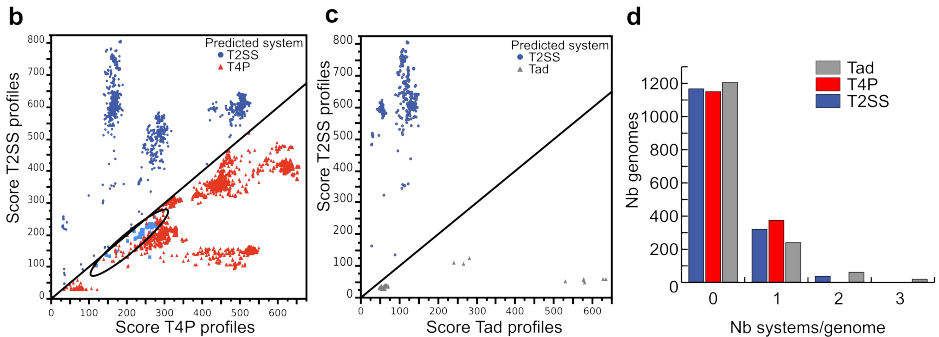
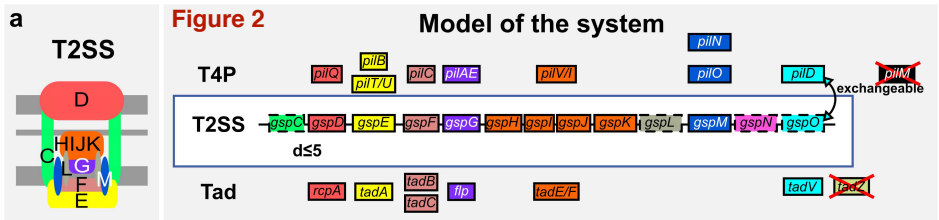
880

881 Table 1- Public webserver (Web) and downloadable applications (App) to identify
 882 components (C), clusters of components (CC), or complete (eventually scattered, S)
 883 bacterial protein secretion systems.

Name	System	Web	App	C (method)	CC	S	URL and Reference
AtlasT4SS	T4	Yes	No	Yes (Blast)	No	No	http://www.t4ss.incc.br 118
CONJscan	T4	Yes	No	Yes (HMM)	No	No	http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::CONJscan-T4SSscan 66
SecReT4	T4	Yes	No	Yes (Blast or HMM)	Yes	No	http://db-mm1.sjtu.edu.cn/SecReT4 119
SecReT6	T6 ^{II}	Yes	No	Yes (Blast or HMM)	Yes	No	http://db-mm1.sjtu.edu.cn/SecReT6 120
SSPred	T1, T2, T3, T4	Yes	No	Yes (other)*	No	No	http://www.bioinformatics.org/sspred 121
T346Hunter	T3, T4, T6 ^I	Yes	No	Yes (Blast and HMM)	Yes	No	http://bacterial-virulence-factors.cbgp.upm.es/T346Hunter 122
T3DB	T3	Yes	No	Yes (Blast)	No	No	http://biocomputer.bio.cuhk.edu.hk/T3DB/browse 123
T3SSscan	T3	Yes	No	Yes (HMM)	No	No	http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::T3SSscan-FLAGscan 24
TXSScan	T1, T2, T3, T4, T5 (a,b,c), T6 ⁱ⁻ⁱⁱⁱ , T9, T4P, Tad flagellum.	Yes	Yes	Yes (HMM)	Yes	Yes	http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::txsscan This work

884 * Predicts proteins based on amino acid composition not sequence similarity.

a**Figure 1****Model of the system****b****c**



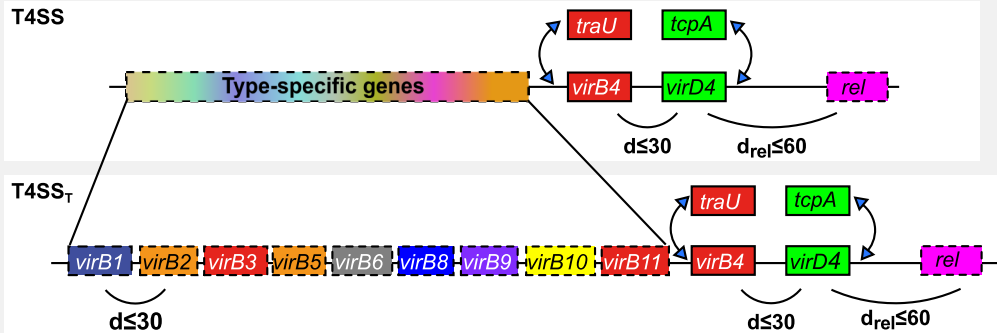
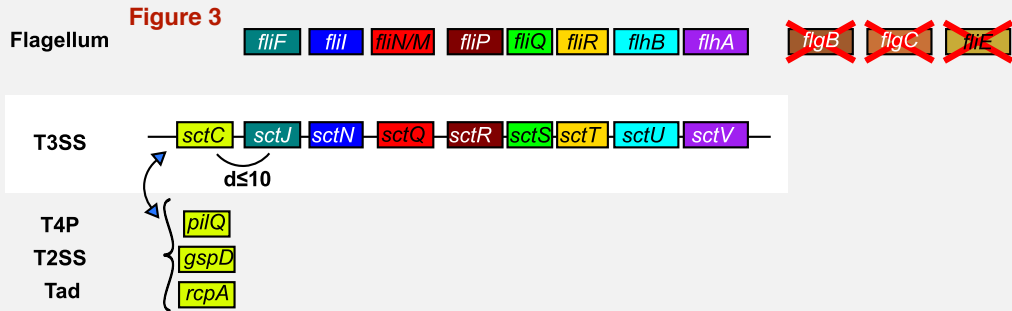
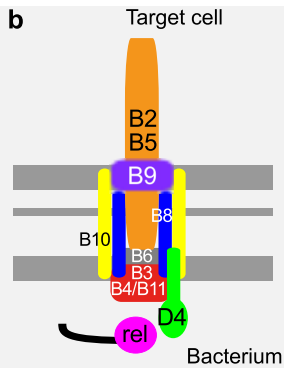
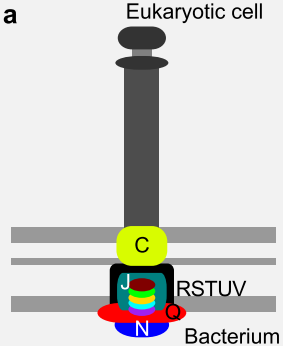
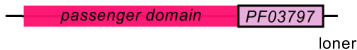
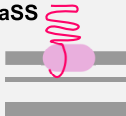


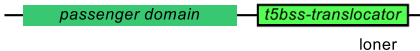
Figure 4

Models of the systems

T5aSS



T5bSS



T5cSS

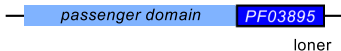


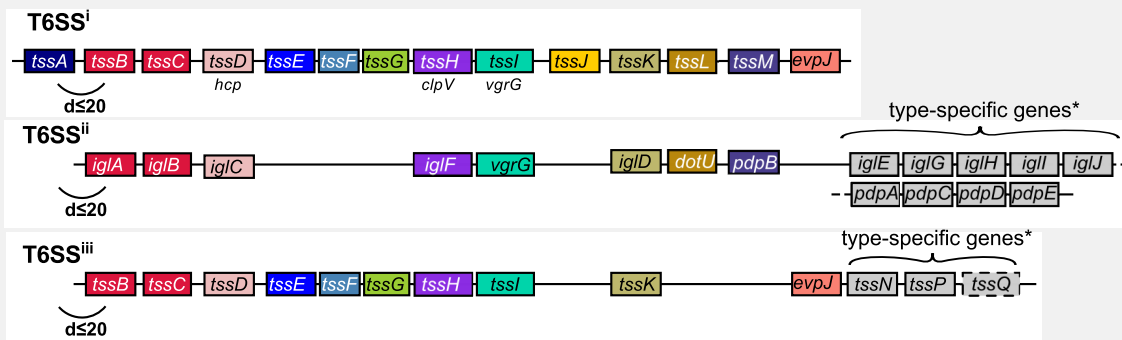
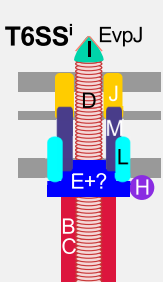
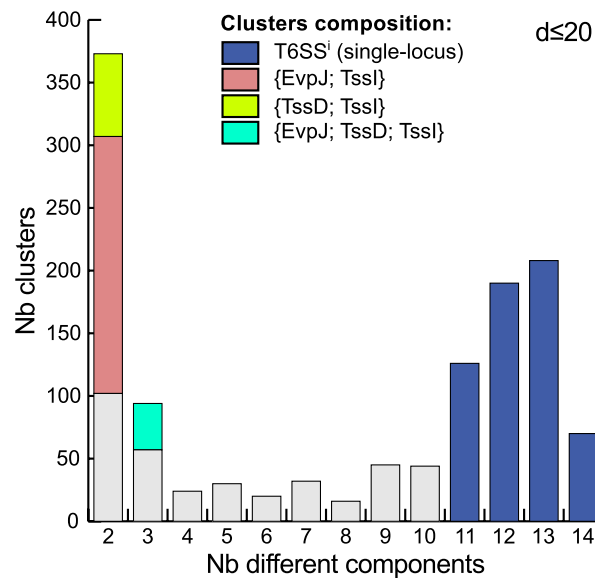
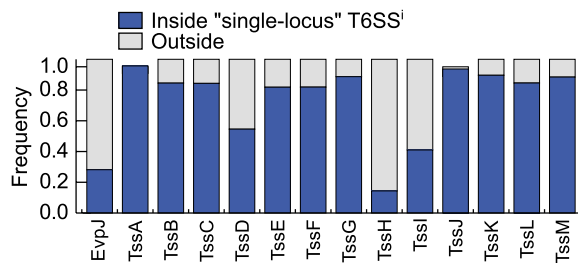
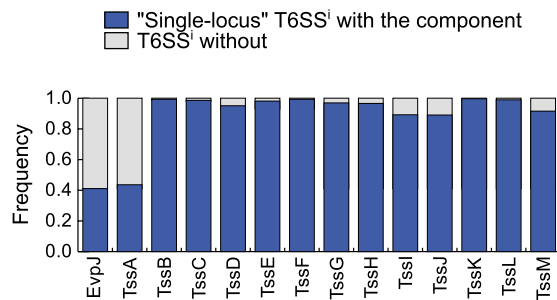
Figure 5**Model of the systems****b****c****d**

Figure 6

Model of the system



Membrane architecture:

Monoderms (M)

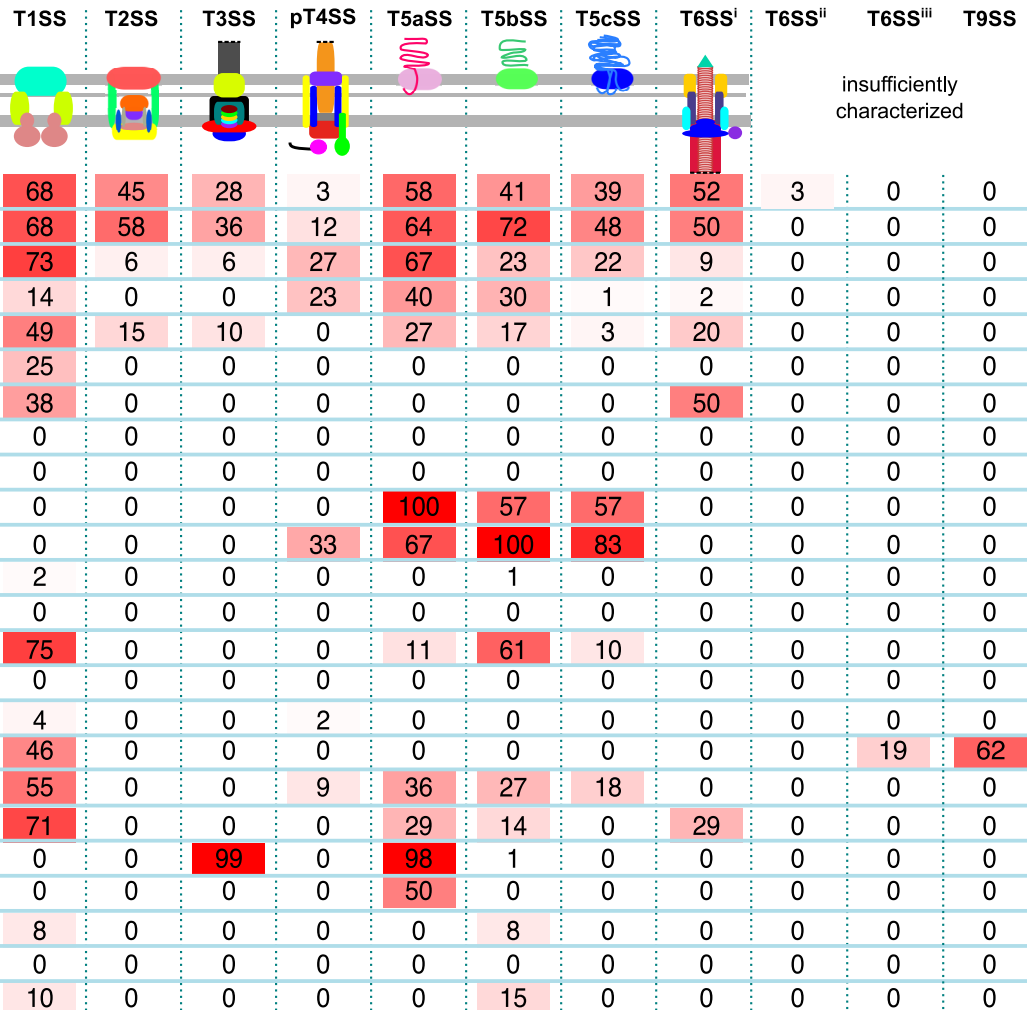
Diderm-LPS

Diderm, putatively with LPS (D⁺)

Diderm with no LPS (D⁻)

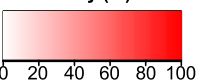
Toga envelope (T)

Figure 7



insufficiently characterized

Colour key (%)



α, β, γ-proteobacteria (946 genomes)

rest of the diderm dataset (582 genomes)

