

SecReT4: a web-based bacterial type IV secretion system resource

Dexi Bi¹, Linmeng Liu¹, Cui Tai¹, Zixin Deng¹, Kumar Rajakumar^{2,*} and Hong-Yu Ou^{1,*}

¹State Key Laboratory of Microbial Metabolism, Shanghai Jiaotong University, Shanghai 200030, China and

²Department of Infection, Immunity and Inflammation, University of Leicester, Leicester LE1 9HN, UK

Received August 12, 2012; Revised October 10, 2012; Accepted November 4, 2012

ABSTRACT

SecReT4 (<http://db-mml.sjtu.edu.cn/SecReT4/>) is an integrated database providing comprehensive information of type IV secretion systems (T4SSs) in bacteria. T4SSs are versatile assemblages that promote genetic exchange and/or effector translocation with consequent impacts on pathogenesis and genome plasticity. T4SSs have been implicated in conjugation, DNA uptake and release and effector translocation. The effectors injected into eukaryotic target cells can lead to alteration of host cellular processes during infection. SecReT4 offers a unique, highly organized, readily exploreable archive of known and putative T4SSs and cognate effectors in bacteria. It currently contains details of 10752 core components mapping to 808 T4SSs and 1884 T4SS effectors found in representatives of 289 bacterial species, as well as a collection of more than 900 directly related references. A broad range of similarity search, sequence alignment, phylogenetic, primer design and other functional analysis tools are readily accessible via SecReT4. We propose that SecReT4 will facilitate efficient investigation of large numbers of these systems, recognition of diverse patterns of sequence-, gene- and/or functional conservation and an improved understanding of the biological roles and significance of these versatile molecular machines. SecReT4 will be regularly updated to ensure its ongoing maximum utility to the research community.

INTRODUCTION

Bacteria transport numerous substrates across cellular membranes via secretion systems that are essential for virulence and survival. Seven general classes of secretion systems, termed type I to type VII, have been described in

bacteria (1). Type IV secretion systems (T4SSs) are versatile, bacterial membrane-spanning apparatuses, composed of diverse structural units, which mediate both genetic exchange and the delivery of effector proteins to target eukaryotic cells (2,3). Hence, these secretory organelles play key roles in bacterial genome plasticity and pathogenesis. Indeed, T4SSs have been directly implicated in the horizontal transfer of genes coding for virulence determinants, antibiotic resistance and other bacterial adaptation traits (4,5).

T4SSs have been classified into three subfamilies by function (2). ‘Conjugation’ associated T4SSs, often encoded on self-transmissible plasmids (6) and integrative and conjugative elements (ICEs) (7), mediate contact-dependent DNA transfer into diverse bacterial species and, in selected instances, even into fungal, plant or human cells. ‘DNA uptake and release’ T4SSs function independently of contact with a target cell and instead promote genetic exchange by a different mechanism. Finally, ‘effector translocator’ T4SSs inject effectors into target eukaryotic cells during host–bacterium interaction processes to mediate bacterium-directed subversion of a myriad of host cell functions. T4SSs are multiunit assemblages which are typically coded for by loci exhibiting differing gene contents, gene orders and numbers of shared homologues. Therefore, T4SSs have been divided into subgroups based on gene order, content and evolutionary relationship (8). Originally, T4SSs were classified into types F, P and I, based on the incompatibility groups IncF, IncP and IncI, respectively, of the plasmids that coded for archetypal members of each of these classes (9). An alternative T4SS classification scheme comprise types IVA, IVB and an ‘others’ category to encompass those that bear little or no discernible ancestral relatedness to the types IVA or IVB systems (10). Type IVA systems are composed of subunits similar in composition and number to those of the archetypal *Agrobacterium tumefaciens* VirB/D system, whereas type IVB systems are assembled from subunits related to the archetypal *Legionella pneumophila* Dot/Icm system. Type F and P

*To whom correspondence should be addressed. Tel: +86 21 62932943; Fax: +86 21 62932418; Email: hyou@sjtu.edu.cn
Correspondence may also be addressed to Kumar Rajakumar. Tel: +44 116 2231498; Fax: +44 116 2525030; Email: kr46@le.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

systems map to the type IVA group, whereas type I systems map to the type IVB category. In addition, the newly described conjugative T4SSs that are encoded by a wide variety of syntenically related genomic islands (GIs), such as ICE*Him1056*, but that are poorly related to earlier described systems have been categorized as GI type T4SSs (5). Finally, a class of T4SSs that mediate the conjugative transfer of plasmids and ICEs and thus promote dissemination of multiple-antibiotic resistance and other virulence-associated traits have recently been described in an increasing number of Gram-positive bacteria (4,8,11).

T4SS components and effectors are being identified in increasing numbers as the available bacterial genome sequence data expands exponentially (12,13). However, few published T4SS-focused web-based resources are available. The VFDB database (14) of bacterial virulence factors captures some relevant information but to date only contains details of annotated T4SSs from 23 bacterial strains. AtlasT4SS is a manual curated database that carries details of 1617 T4SS proteins from 58 Bacteria and one Archaea and provides a hierarchical classification scheme with three levels to classify these proteins into orthologous groups (15); data derived from 11 plasmids are also included. Notably, AtlasT4SS is focused on individual T4SS component proteins rather than entire T4SSs, with component proteins presented by orthology and not as organized collective functional systems. The dataset size of AtlasT4SS is still relatively modest. Besides, neither of the above two databases include adequate information on T4SS effectors. As a well-organized, functionally flexible and comprehensive database of T4SSs is not available, we are collating available experimental and bioinformatics analyses data and the literature on known and putative T4SSs and cognate effectors in bacteria as an open-access database called SecReT4 (Type IV Secretion system Resource). SecReT4 presents 10424 reported and predicted essential components that make up T4SS instruments in Gram-negative and Gram-positive bacteria. All the predicted data have been curated and re-annotated according to a set of reference systems. Components are not only organized by entire T4SSs but can also be analyzed alongside other components in SecReT4 belonging to the same bit part. Thus, as an example, users can access a list of all VirB1 proteins within SecReT4. SecReT4 has classified all archived T4SSs based on above mentioned classification schemes. SecReT4 also includes data on effectors. Furthermore, more than 900 references have been collected, mined and mapped to T4SSs and effectors. The database provides functional descriptions for T4SSs, T4SS components and cognate effectors according to findings reported in the literature. To organize and display these data in an efficient manner, we designed a customized PostgreSQL schema and developed a user-friendly web-interface. Besides, SecReT4 offers easily accessible tools such as BLAST, HMMER, Primer3 and MUSCLE to permit a wide range of analyses. SecReT4 exclusively provides a web tool to identify T4SS components encoded by user-supplied DNA sequences, thus promoting the efficient identification of novel putative T4SSs in newly sequenced

genomes. We envisage that SecReT4 will facilitate efficient, multi-disciplinary and innovative exploration of the large numbers of these systems, and in the process lead to a better understanding of the biological roles and significance of these versatile molecular machines. We expect that SecReT4 will prove to be of major interest to a broad community of researchers.

MATERIALS AND METHODS

Complete sequences of 1537 prokaryotic chromosomes, 1094 plasmids sequenced with chromosomes and 1362 plasmids sequenced alone available as of 13 April 2012 were retrieved from the NCBI RefSeq project. Data on T4SSs identified among these replicons were derived from reports of experimentally validated T4SSs, published information related to T4SSs and computational analyses of bacterial genome sequences. The approach used to predict T4SS-encoding loci is shown in the Supplementary Figure S1. Information on 18 representative T4SSs spanning the recognized genetic and function diversity was collected as an initial query dataset (Supplementary Table S1). The selected representative systems had been well-studied and mapped to a wide range of T4SSs in terms of gene and/or protein sequence similarity and/or gene order. The inferred phylogenetic relationships of the VirB4 components were illustrated in the Supplementary Figure S2. We then extracted the verified or proposed core T4SS components from this query dataset (Supplementary Table S2). Likely, T4SS-associated lytic transglycosylases (16), such as VirB1, which can degrade cell walls are also included. Co-occurrence of genes coding for protein homologues of individual core components has allowed for the identification of putative T4SSs (12) as T4SSs contain one or more regions with clustered components. Therefore, we searched the full set of bacterial replicons using BLASTp for homologues ($e\text{-value} \leq 0.0001$) of all components listed in the Supplementary Table S2 and identified instances of co-localization of encoding genes. In addition, the obtained partial T4SSs were further manually curated based on the RefSeq genome annotation (Supplementary Figure S1). Subsequently, newly identified T4SSs were loaded into the query dataset in turn and the search was performed reiteratively for more T4SS-encoding loci. Using this approach, 808 T4SSs that are assembled from 10424 component proteins have been identified and archived in SecReT4. These 808 T4SSs systems are encoded by 285 chromosomes and 351 plasmids of 524 bacterial strains belonging to 289 species.

More than 900 references were collected from PubMed using the search terms 'type IV secretion', and 'type IV secretion AND effector', and by capturing T4SS-related citations within these sources. All references were manually screened for details of T4SSs. Thus, we performed validation of the predicted results by checking all identified putative T4SSs against information contained within these references. Ninety five of the identified T4SSs exhibited good matches to those reported in experimental literatures, while 200 were in accord with genome sequencing reports. Notably, our predictions were

consistent with a previous bioinformatics study which identified specific components of T4SSs but with an emphasis on the conjugation machinery of ICEs (12).

Data on 1638 experimentally verified and/or *in silico* predicted T4SS effectors were also collected from the literature mentioned earlier. These effectors were subsequently used to identify an additional 245 homologues by using BLASTp searches against the sequenced bacterial genomes with the strict cutoffs (e -value ≤ 0.0001 and identity $\geq 70\%$) and manual examination for the co-presence of T4SSs in the same strain. In addition, the *Helicobacter pylori* peptidoglycan, a non-protein effector delivered by the Cag system (17), was included. As such, SecReT4 currently contains 1884 (1638 + 245 + 1 = 1884) T4SS effectors.

The two classification schemes mentioned earlier were utilized to classify T4SSs archived within SecReT4. For consistency with the scheme which uses type IVA and type IVB categories, we propose that currently defined GI type T4SSs and those present in Gram-positive bacteria be designated type IVC and type IVD, respectively (Supplementary Table S1). In addition, it is important to note that T4SSs encoded by GIs are not restricted to type IVC and that many map to the other categories, such as the well-studied SXT ICE whose conjugation is mediated by a type IVA system (7).

The SecReT4 database is run on a high-performance four-slot, four-way server (Inspur NF8560), which has been equipped with four six-core XEON E7-4807 1.86 GHz processors and 64 GB memory. SecReT4 employs the relational database management system PostgreSQL as its back-end with a customized schema designed to organize all uploaded information, including data on T4SSs and cognate effectors, as well as related references. SecReT4 runs on a Linux platform with an Apache web server. Web interfaces were developed using HTML, CSS and JavaScript. Data pipelines were developed with PHP and Perl. In addition, the database was enhanced by use of the CGview circular genome visualization tool (18), as well as the MUSCLE (19) and Jalview (20) multiple sequence alignment and visualization tools.

RESULTS AND DISCUSSION

PostgreSQL-based SecReT4 offers a flexible and friendly web-interface. Users can create maps of T4SS gene clusters, investigate gene loci of interest with the embedded graphic display and search for T4SSs by name, function or host organism. Users can also search a query sequence against SecReT4 to identify potential homologues of T4SS components or effectors. Multiple sequence alignment analyses can easily be performed using the embedded MUSCLE and Jalview tools for user-directed investigations focused on diverse T4SS components, thus facilitating varied individualized directions of research. The SecReT4 homepage contains the following interfaces: 'Home', 'Browse', 'Search', 'Tools', 'Download' (nucleotide/protein sequences), 'References' and 'Introduction' (description of T4SSs and SecReT4).

SecReT4 browse module

The SecReT4 browse module contains detailed information on all archived T4SSs, matching essential component parts and predicted T4SS effector molecules. These data include unique identifiers, host species details, sequences, functions and hyperlink paths to other public databases, such as NCBI, UniprotKB, PDB and KEGG. The 'Browse by replicon' page provides a hyperlinked organized list of bacterial replicons in which T4SSs have been identified. For each T4SS-encoding replicon, SecReT4 provides whole genome maps with the locations and sizes of genes coding for the identified T4SS components and any identified candidate effectors flagged and hyperlinked to allow for T4SS-centered zoom-out genome-scale views. Users can further access individual pages dedicated to each T4SS as required. Here, the VirB T4SS system encoded by the Ti plasmid (NCBI Refseq accession no. NC_003065) of *A. tumefaciens* C58 is taken as an example (Figure 1). *Agrobacterium tumefaciens* C58 contains two plasmids coding for three T4SSs, two of which are on the nopaline-agrocinopine-type tumor-inducing Ti plasmid. *Agrobacterium tumefaciens* causes crown gall tumor in plants via VirB-mediated delivery of transferred DNA (T-DNA) and virulent effectors into target plant cells. The introduced 'parasitic' T-DNA is then able to randomly integrate into the genome of its new eukaryotic host. The ability of *A. tumefaciens* to facilitate inter-kingdom genetic exchange has also made it a powerful tool for transgenic plant production. Tabulated (Figure 1A) and graphically displayed (Figure 1B) outputs of the two experimentally verified T4SSs and five characterized effectors encoded on the Ti plasmid are as shown. Further details pertaining to and the genomic context of the VirB system are also highlighted (Figure 1C). The VirB system consists of 12 components including the mating pair formation proteins VirB1–VirB11, and the coupling protein VirD4 that recruits substrates. Users can access detailed information on each component by clicking the gene name (Figure 1D), and can also retrieve a list of proteins from diverse T4SSs belonging to the same component part by clicking on the assigned component name in the gene list, with the readily available option of follow-up multiple sequence alignment (Supplementary Figure S3). The *A. tumefaciens* VirB system is implicated in plant pathogenesis via translocation of oncogenic DNA and protein into target cells. Hyperlinked T-DNA coordinates and protein effectors are present in the database (Figure 1E). VirD2 nicks the T-DNA region at the border repeats and thereby becomes covalently linked to the 5'-end of the nick site. The VirD2-T-strand complex is then translocated through the secretion channel. In contrast, VirE2, VirE3, VirF and VirD5 are secreted effectors which contribute directly to virulence in plant cells. For example, VirE2 is a DNA-binding protein and protect the T-strand from nucleolytic degradation in the plant cell. In addition to the above features, SecReT4 automates primer design for PCR amplification of sequences coding for both component parts and effectors via Primer3Plus (21).

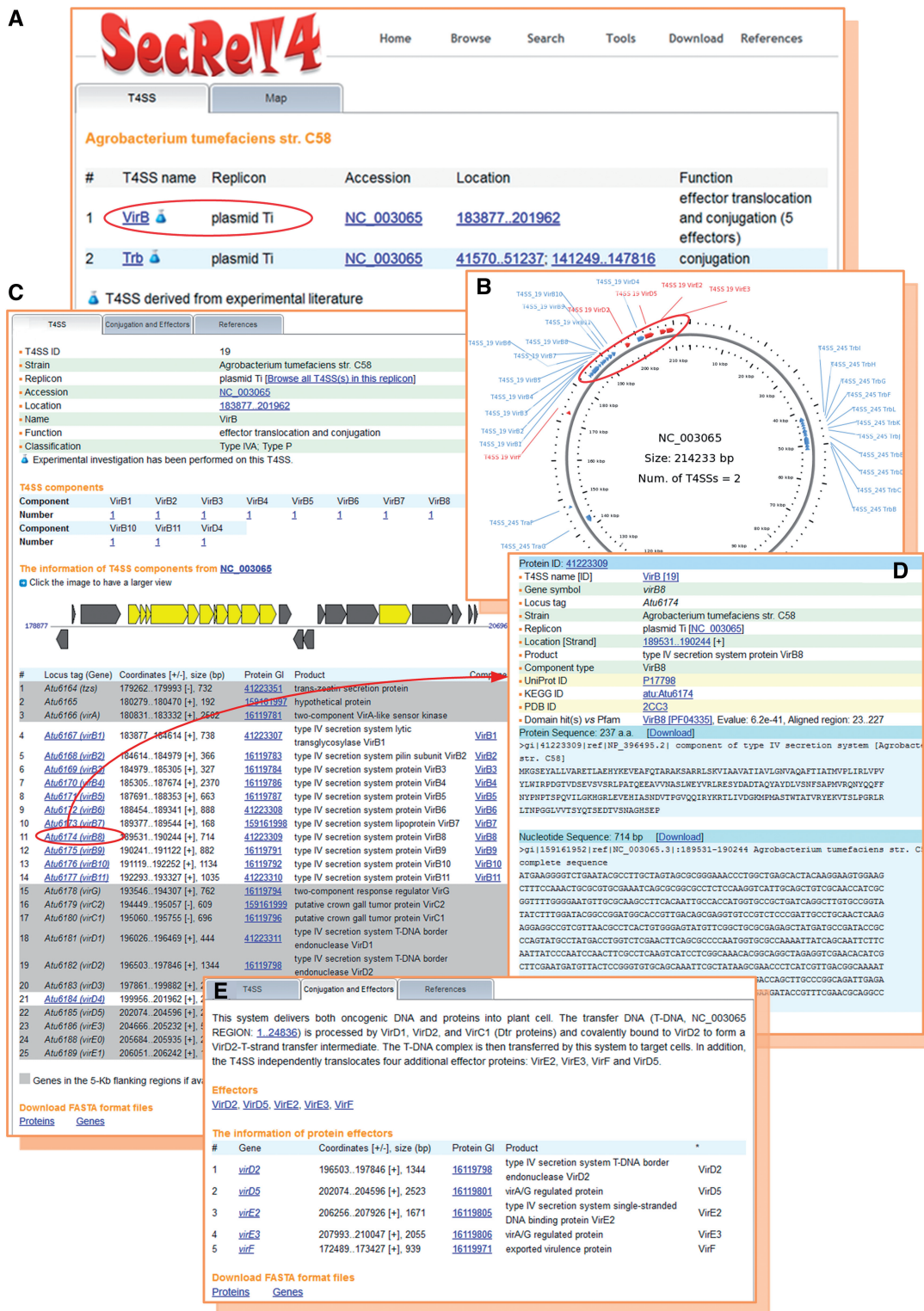


Figure 1. An overview of SecRet4 outputs using the VirB system encoded by the *Agrobacterium tumefaciens* C58 Ti plasmid as an example. (A) List of two identified T4SSs encoded by the *A. tumefaciens* C58 Ti plasmid, both of which have been experimentally verified. (B) A scaled representation of the circular Ti plasmid generated by the SecRet4-integrated CGview (18) utility showing locations and sizes of genes coding for components of both T4SSs (blue) and VirB-associated effectors (red) within this replicon. (C) An overview of the features of the VirB system: replicon, location, functions, classification, number of components and a complete gene list of this system with links to component-related information. Hyperlinks to NCBI are provided as appropriate, as is a link that allows for visualization of the gene content of the system facilitated by the embedded graphic display. (D) SecRet4-enabled tools for examining component-related features are exemplified by the VirB8 entry. In addition to sequence information, direct links to NCBI, UniprotKB, KEGG and PDB are provided. Significant hits against the Pfam database are also presented. (E) Data on substrates transported by the VirB system are illustrated by information on T-DNA and the characterized cognate protein effectors.

The 'Browse by classification' page allows users to retrieve T4SSs classified by two schemes (Supplementary Table S1). According to the scheme proposed by Christie *et al.* (10), we further classified 'the other' systems into type IVC and type IVD. Types IVA and IVB have been previously defined. Type IVC includes the newly described GI type T4SS, and type IVD systems are related to those recently characterized in Gram-positive bacteria. Similarly, for an alternative classification scheme, we have added type G for GI type and type GP for Gram-positive T4SSs to the earlier defined type F, type P and type I grouping scheme (6,9). In addition, the component part VirB4, which is conserved in all characterized T4SSs, has been proposed as a signature for classification of these secretions (8,12). Indeed, our phylogenetic analysis of 837 proteins assigned as the VirB4 component (Supplementary Figure S2) matches the current classification schemes, and suggests VirB4 can be used as a classification marker once appropriate quantitative phylogenetic criteria are established.

The 'Browse by function' page catalogs T4SSs by function including 'conjugation', 'DNA uptake and release', 'effector translocation', 'other' and 'unknown' for those whose function has not been characterized or proposed yet. Currently, 211 T4SSs have been proved or proposed to partake in conjugation and 8 in DNA uptake and release, 47 have been implicated in effector translocation while 5 participate in virulence via other ways, such as by adhesion to erythrocytes. Functions of 537 (66%) T4SSs await further investigation. T4SSs encoded by ICEs have been hyperlinked to corresponding elements in ICEberg (22).

SecReT4 search options and tools

SecReT4 enables text, BLAST and HMMER3 searches with varied options. Through the 'Search' page, users can retrieve details of T4SSs, components or effectors from SecReT4 by the following categories: species, replicon RefSeq accession, T4SS name, T4SS function, component name, effector name and/or NCBI protein ID. The 'Tools' page allows users to search a query sequence against SecReT4 to obtain and visualize potential homologous matches using WU-BLAST 2.0 (Gish, W., personal communication) and HMMER3 (23). We also provide a web tool to identify T4SS components in a user-supplied DNA sequence. For raw sequences, the tool predicts genes using Glimmer3 (24) or Prodigal (25) based on users' option. Proteins are searched against SecReT4 for putative T4SS components by BLASTp. Significant hits within a user-defined distance can be clustered. Graphic output displaying genome context is available. For short sequences (30–10 000 bp), if no gene is detected, BLASTx will be initiated to search the DNA sequence directly against SecReT4.

SecReT4 reference module

The SecReT4 reference section offers information on papers relating to T4SSs that have been identified as described earlier, and further pertinent references relating to entries within SecReT4. Direct links to the

matching PubMed entries are also provided. At present SecReT4 contains records of more than 800 directly relevant scientific publications. This reference collection will be updated with new entries being subject to manual curation and organization in a timely manner. These references have been assigned with the following tags: experimental studies, *in silico* analyses, genome sequencing and reviews. SecReT4 links the examined literature to relevant T4SSs and effectors as indicated by corresponding thumbnail icons. The reference collection is also searchable by author, title, journal, year and PubMed ID, and matching abstracts can be subjected to standard word searches. This provides an easily accessible literature resource that has been subjected to both text mining, manual curation and subset categorization.

CONCLUSION

Newly available data on T4SSs and cognate effectors will be uploaded regularly to keep pace with the rapidly expanding bacterial genome database. New records of publications will be mined to update the status of currently archived T4SSs, component parts and effectors, and to supplement the current collection, as when new data arises. As T4SSs are multiunit instruments, we plan to capture data on component interaction and endeavor to establish protein–protein interaction prediction pipelines to stimulate research on structural models of T4SSs based on sequence information. Furthermore as bioinformatics efforts at identifying and classifying T4SS effectors have been limited, we will aim to further analyze the currently verified and predicted T4SS effector sets, establish biologically relevant schemes for organizing these proteins and exploit machine learning strategies to expand the set of known likely T4SS effectors. In addition, as homologues of T4SSs have been identified in Archaea (4), we will include archaeal T4SSs after a complete such system has been characterized. We envisage an evolving resource that captures a growing variety of T4SS-related data extracted and curated from experimental literature and derived by increasingly sophisticated bioinformatics analyses. Ultimately, we propose a T4SS-specific resource to facilitate efficient investigation of large numbers of these systems, recognition of diverse patterns of sequence-, gene- and/or functional conservation, and an improved understanding of the biological roles and significance of these versatile molecular machines.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2 and Supplementary Figures 1–3.

ACKNOWLEDGEMENTS

We are grateful to three anonymous referees for their constructive suggestions.

FUNDING

National Natural Science Foundation of China [31170082]; the 973 program, Ministry of Science and Technology, China [2012CB721002]; Program for New Century Excellent Talents in University, Ministry of Education, China [NCET-10-0572]; Chen Xing Young Scholars Programme, Shanghai Jiaotong University (to H.Y.O.); a Royal Society—National Natural Science Foundation of China International Joint Project grant [to K.R. and Z.D.; 2007/R3]. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement. None declared.

REFERENCES

1. Tseng, T.T., Tyler, B.M. and Setubal, J.C. (2009) Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiol.*, **9**(Suppl. 1), S2.
2. Cascales, E. and Christie, P.J. (2003) The versatile bacterial type IV secretion systems. *Nat. Rev. Microbiol.*, **1**, 137–149.
3. Fronzes, R., Christie, P.J. and Waksman, G. (2009) The structural biology of type IV secretion systems. *Nat. Rev. Microbiol.*, **7**, 703–714.
4. Alvarez-Martinez, C.E. and Christie, P.J. (2009) Biological diversity of prokaryotic type IV secretion systems. *Microbiol. Mol. Biol. Rev.*, **73**, 775–808.
5. Juhas, M., Crook, D.W. and Hood, D.W. (2008) Type IV secretion systems: tools of bacterial horizontal gene transfer and virulence. *Cell Microbiol.*, **10**, 2377–2386.
6. Smillie, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E.P. and de la Cruz, F. (2010) Mobility of plasmids. *Microbiol. Mol. Biol. Rev.*, **74**, 434–452.
7. Wozniak, R.A. and Waldor, M.K. (2010) Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat. Rev. Microbiol.*, **8**, 552–563.
8. Wallden, K., Rivera-Calzada, A. and Waksman, G. (2010) Type IV secretion systems: versatility and diversity in function. *Cell Microbiol.*, **12**, 1203–1212.
9. Lawley, T.D., Klimke, W.A., Gubbins, M.J. and Frost, L.S. (2003) F factor conjugation is a true type IV secretion system. *FEMS Microbiol. Lett.*, **224**, 1–15.
10. Christie, P.J., Atmakuri, K., Krishnamoorthy, V., Jakubowski, S. and Cascales, E. (2005) Biogenesis, architecture, and function of bacterial type IV secretion systems. *Annu. Rev. Microbiol.*, **59**, 451–485.
11. Grohmann, E., Muth, G. and Espinosa, M. (2003) Conjugative plasmid transfer in gram-positive bacteria. *Microbiol. Mol. Biol. Rev.*, **67**, 277–301, table of contents.
12. Guglielmini, J., Quintais, L., Garcillan-Barcia, M.P., de la Cruz, F. and Rocha, E.P. (2011) The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.*, **7**, e1002222.
13. Burstein, D., Zusman, T., Degtyar, E., Viner, R., Segal, G. and Pupko, T. (2009) Genome-scale identification of *Legionella pneumophila* effectors using a machine learning approach. *PLoS Pathog.*, **5**, e1000508.
14. Chen, L., Xiong, Z., Sun, L., Yang, J. and Jin, Q. (2012) VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.*, **40**, D641–D645.
15. Souza, R.C., Quispe Saji, G.D., Costa, M.O., Netto, D.S., Lima, N.C., Klein, C.C., Vasconcelos, A.T. and Nicolas, M.F. (2012) AtlasT4SS: a curated database for type IV secretion systems. *BMC Microbiol.*, **12**, 172.
16. Koraimann, G. (2003) Lytic transglycosylases in macromolecular transport systems of Gram-negative bacteria. *Cell. Mol. Life Sci.*, **60**, 2371–2388.
17. Tegtmeyer, N., Wessler, S. and Backert, S. (2011) Role of the cag-pathogenicity island encoded type IV secretion system in *Helicobacter pylori* pathogenesis. *FEBS J.*, **278**, 1190–1202.
18. Stothard, P. and Wishart, D.S. (2005) Circular genome visualization and exploration using CGView. *Bioinformatics*, **21**, 537–539.
19. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
20. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. and Barton, G.J. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
21. Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R. and Leunissen, J.A. (2007) Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.*, **35**, W71–74.
22. Bi, D., Xu, Z., Harrison, E.M., Tai, C., Wei, Y., He, X., Jia, S., Deng, Z., Rajakumar, K. and Ou, H.Y. (2012) ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Res.*, **40**, D621–D626.
23. Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
24. Delcher, A.L., Bratke, K.A., Powers, E.C. and Salzberg, S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.
25. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.