SCIENTIFIC REPORTS

OPEN

Received: 18 October 2016
Accepted: 14 December 2016
Published: 23 January 2017

SecretEPDB: a comprehensive web-based resource for secreted effector proteins of the bacterial types III, IV and VI secretion systems

Yi An[1,2,*], Jiawei Wang[3,*], Chen Li[4], Jerico Revote[5], Yang Zhang[1], Thomas Naderer[6], Morihiro Hayashida[7], Tatsuya Akutsu[7], Geoffrey I. Webb[2], Trevor Lithgow[4] & Jiangning Song[2,6]

Bacteria translocate effector molecules to host cells through highly evolved secretion systems. By definition, the function of these effector proteins is to manipulate host cell biology and the sequence, structural and functional annotations of these effector proteins will provide a better understanding of how bacterial secretion systems promote bacterial survival and virulence. Here we developed a knowledgebase, termed SecretEPDB (Bacterial Secreted Effector Protein DataBase), for effector proteins of type III secretion system (T3SS), type IV secretion system (T4SS) and type VI secretion system (T6SS). SecretEPDB provides enriched annotations of the aforementioned three classes of effector proteins by manually extracting and integrating structural and functional information from currently available databases and the literature. The database is conservative and strictly curated to ensure that every effector protein entry is supported by experimental evidence that demonstrates it is secreted by a T3SS, T4SS or T6SS. The annotations of effector proteins documented in SecretEPDB are provided in terms of protein characteristics, protein function, protein secondary structure, Pfam domains, metabolic pathway and evolutionary details. It is our hope that this integrated knowledgebase will serve as a useful resource for biological investigation and the generation of new hypotheses for research efforts aimed at bacterial secretion systems.

In the course of pathogenesis, bacteria utilize highly evolved secretion systems to translocate (secrete) proteins into host cells. A majority of these secreted proteins are enzymes, toxins or "effectors"; with effector proteins functioning to subvert the pathways of host cells to facilitate bacterial pathogenicity[1,2]. A growing number of bacterial secretion systems have been identified to date, from type I to type IX[2–7]. They play important roles in mediating the interactions of bacteria with their host cells, and thus determine infection outcomes[8]. For example, bacteria are able to degrade the extracellular matrix and cell walls of host niches using secreted enzymes[3,9]. These enzymes are exported to the environment and their secretion is mainly through the secretion systems of type I (T1SS), type II (T2SS) or type V (T5SS)[10].

Effector proteins are translocated into host cells predominantly by the type III secretion system (T3SS), type IV secretion system (T4SS) or type VI secretion system (T6SS)[1,11–13]. Of these, the T3SS has been most extensively studied both structurally and functionally and has been shown to exist in diverse bacterial species[6,7]. Both animals and plants can be infected by pathogens that use T3SS effectors (T3SEs)[3,6,7,14]. The T4SS is regarded as one of the most functionally diverse bacterial secretion system, both in terms of transported substrates and targeted

[1]College of Information Engineering, Northwest A&F University, Yangling 712100, China. [2]Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia. [3]School of Electronic and Computer Engineering, Peking University, Beijing 100871, China. [4]Infection and Immunity Program, Biomedicine Discovery Institute and Department of Microbiology, Monash University, Melbourne, VIC 3800, Australia. [5]Monash Bioinformatics Platform, Monash University, Melbourne, VIC 3800, Australia. [6]Infection and Immunity Program, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia. [7]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to Y.Z. (email: zhangyang@nwsuaf.edu.cn) or G.I.W. (email: Geoff.Webb@monash.edu) or J.S. (email: Jiangning.Song@monash.edu)

recipients[15]. The T4SS is characterized as a large family of macromolecule transporter systems that incorporates three recognized sub-families: bona fide effector protein transport systems (e.g. Dot/Icm from *Legionella*; CagPAI from *Helicobacter*), machinery for DNA uptake/release (e.g. Tra from *Neisseria*) and conjugation systems for the transfer of genetic material between bacteria as well as from bacteria to eukaryotic cells (e.g. VirB and Trw from *Bartonella*)[16,17]. Only more recently discovered, necessarily less is known about the T6SS which is composed of a contractile, needle-tube puncturing apparatus to deliver effectors into host or other bacterial cells[13].

Through secretion of effector proteins and interaction with host factors, protein secretion represents an important aspect of bacterial physiology, and a crucial means for adaptation and survival within host niches. With the functional importance of bacteria secretion systems in mediating the mutualistic symbiotic or pathogenic relationships[18], experimental and computational studies have been aimed at understanding the role of effector proteins in host-pathogen interactions[19–30]. Web servers for predicting T3SS, T4SS or T6SS effector proteins from genome sequence data have been established[31–34]. However, to the best of our knowledge, there are currently no available knowledge-bases or resources that document and curate the annotations for effector proteins of the T3SS, T4SS and T6SS. Considering the importance of bacteria secretion systems, comprehensive sequence, structural and functional annotations of their effector proteins will provide a better understanding of their importance.

To bridge this knowledge gap, we developed a new web-based resource termed SecretEPDB (Bacterial Secreted Effector Protein DataBase), for a comprehensive annotation of effector proteins secreted by the bacterial T3SS, T4SS and T6SS. SecretEPDB provides detailed annotations for the three types of effector proteins, through manual extraction and integration using currently available databases or the literature from PubMed. Importantly, the database has been strictly curated to ensure that all effector protein entries in SecretEPDB are supported by experimental evidence published in the scientific literature of being secreted by T3SS, T4SS or T6SS. In addition, several key features of the developed SecretEPDB are as follows:

(1) Protein 3D structural information is available in SecretEPDB. For each entry with available structural information, the corresponding Protein Data Bank (PDB)[35] accession numbers, experimental structural determination methods, and the 3D structures were extracted and made available.
(2) For the entries with UniProt (http://www.uniprot.org/)[36] accession numbers, their functional sites and domains were assembled and can be visualized with the IBS (Illustrator of Biological Sequences) program[37] to provide an enhanced visualization of the sequence context information.
(3) Data visualization is also available for multiple sequence alignment (MSA) of each entry with Alignment-to-Html[38], a third-party JavaScript tool. Alignment-to-Html enables SecretEPDB to visualize MSAs with overlapped functional domains and/or sites. In addition, SecretEPDB allows users to search protein motifs with a user-friendly interface and an option of exporting multiple retrieved sequences in the FASTA format as plain text or directly to MS Word.
(4) SecretEPDB provides annotations of metabolic/signaling pathway for each entry by cross-referencing the KEGG database[39] where such information is available. Pathway annotations are important for understanding the functional roles of the effector proteins within the host cells.
(5) SecretEPDB includes single point mutations and their pathogenicity annotations of each protein entry. For each mutation, detailed annotations including disease type and the corresponding reference papers are provided.
(6) Post-translational modification sites are crucial for protein function. SecretEPDB includes kinase-specific phosphorylation site annotations predicted by the Group-based Prediction System (GPS) program[40], which was employed to provide the annotations of predicted kinase-specific phosphorylation sites in hierarchy. Identifying phosphorylation sites in partner with their cognate protein kinases provides important information for understanding a variety of related cellular processes that are potentially associated with the effector proteins.
(7) In an effort to keep up with the rapid accumulation of experimental data, SecretEPDB allows researchers to submit their most-recent experimental findings of novel effector proteins via an online submission webpage. Please refer to 'Database utility' for more information.

Database construction and content

Data collection.

Figure 1 presents the flowchart describing construction of SecretEPDB. Current database entries were extracted from three major resources: UniProt, Datasets from published studies, and the relevant literature (entries were collected from the literature via keyword search in NCBI Protein). A three-step procedure of the entry retrieval and collection is described as follows.

Firstly, keywords including 'bacterial secretion system', 'bacterial secretion effectors', 'T3SS', 'T4SS' and 'T6SS' were each used to search the entire Swiss-Prot database (i.e. the manually annotated and reviewed dataset of the UniProt database). Expectedly, the search returned a huge number of redundant (and sometimes irrelevant) secreted effector protein candidates. After being carefully reviewed, those proteins that did not belong to any of the three classes (i.e. T3SS, T4SS or T6SS) were disregarded. It is important to note that the obtained entries were required to have accurate and unambiguous descriptions and evidence (such as "secreted by T3SS", or "translocated into the host cell via the type IV secretion system"). As a result, a list of 169 entries was obtained, including 161 type III secretion system effectors (termed as T3SEs), 4 type IV secretion system effectors (T4SEs) and 4 type VI secretion system effectors (T6SEs).

Secondly, a number of effector proteins were collected from datasets[14,20,21,27,28,41] or databases[32] published in the literature. Note that these proteins were collected from NCBI Protein or UniProt database, which may or may not be in their full-length form. During this step, we extracted complete sequences by searching the accession numbers in NCBI Protein or UniProt database. In addition, a number of entries unavailable for extraction, having

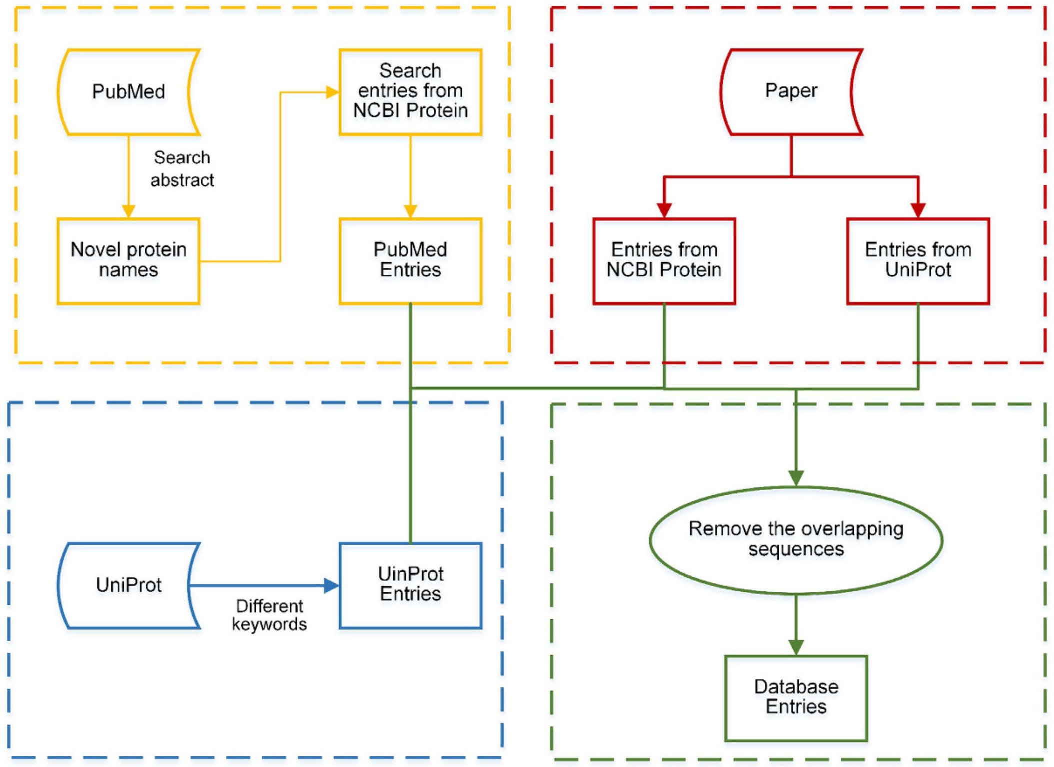


Figure 1. Flowchart of the data collection process in SecretEPDB.

Type	Reference	Number of entries in the reference	Number of entries included in SecretEPDB
T3SE	Dong, X. *et al.*[21]; Wang, Y. *et al.*[28]	150	150
	Arnold, R. *et al.*[20]	100	94
	Tay, D. M. *et al.*[41]	504	334
	Yang, X. *et al.*[14]	283	260
	Dong, X. *et al.*[32]	1215	704
T4SE	Bi, D. *et al.*[31]	239	239
	Zou, L. *et al.*[29]	340	340
	Wang, Y. *et al.*[42]	347	347
	Lifshitz, Z. *et al.*[43]	290	290
T6SE	Li, J. *et al.*[33]	107	88
	Salomon, D. *et al.*[24]	6	6
	Russell, A. *et al.*[44]	50	40
	Russell, A. *et al.*[45]	61	38

Table 1. Statistical summary of the three types of effector proteins collected from the literature.

been removed from NCBI Protein or UniProt database. After careful curation, we obtained 2538 entries: 1150 T3SEs, 1216 T4SEs and 172 T6SEs. Among these, 1090 T3SEs and 254 T4SEs were derived from the UniProt database (i.e. the TrEMBL database). Table 1 provides lists of the number for each type which were obtained from the various data sources.

Finally, we searched the PubMed abstracts for relevant literature to retrieve experimentally reported T3SS, T4SS and T6SS effector proteins. These entries represent newly discovered effector proteins that may not yet be included in the current databases or datasets. In particular, the abstract of each paper in PubMed was mined using a text-mining technique called Scrapy (http://scrapy.org/), a fast and powerful web crawling tool. We extracted T3SS, T4SS and T6SS proteins and their associated information including their names and accession numbers. The collected information was then used to search against the NCBI protein database to retrieve proteins sequences in the FASTA format. Note, some effectors mentioned in the literature would not be included in SecretEPDB until such time as sequence information is available on these effectors. After this step, a total number of 44 entries, including 27 T3SEs, 8 T4SEs and 9 T6SEs were extracted and added into SecretEPDB.

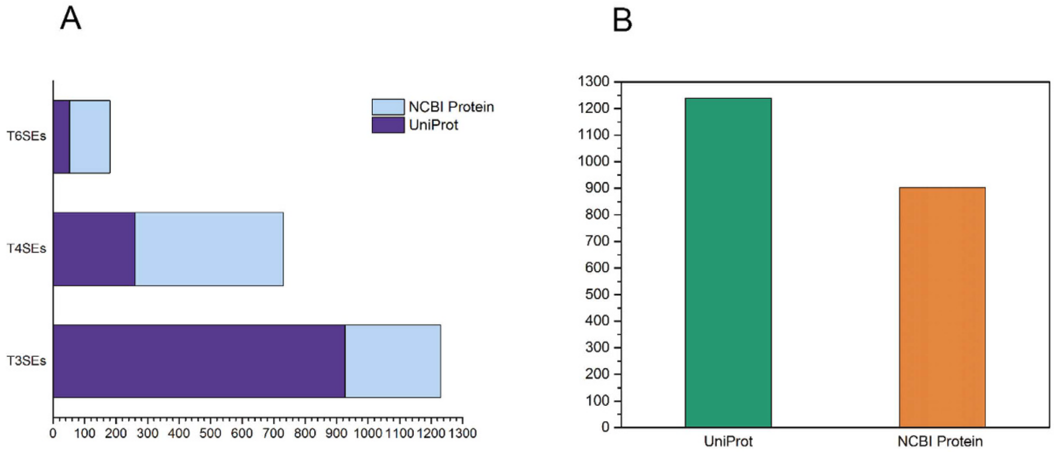


Figure 2. Statistical summary of collected entries currently in SecretEPDB. (**A**) Distribution of effector protein entries according to the original resources used; (**B**) Distribution of entries from UniProt and NCBI protein database.

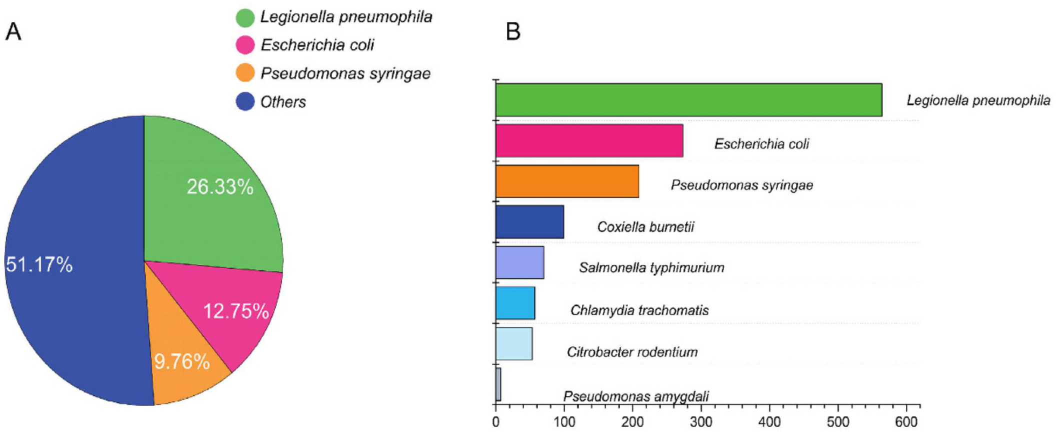


Figure 3. Distribution of collected entries according to bacterial species. (**A**) Distribution of the entries from the three dominant bacterial species; (**B**) Statistical analysis of the entries from the top eight bacterial species.

These steps generated a total of 1338 T3SEs, 1228 T4SEs and 185 T6SEs. We then reduced the sequence redundancy of the collected effector proteins by comparing their UniProt and NCBI Protein accession numbers. Altogether SecretEPDB collected 2142 experimentally verified effectors (1239 entries exist in UniProt and 903 entries exist in the NCBI Protein database). Figure 2 shows the numbers of T3SEs, T4SEs and T6SEs and the distribution of these entries from different resources.

With the collected effector protein entries in SecretEPDB we conducted a statistical analysis of their distribution across the bacterial species: the most abundant source with 26.33% of entries was *Legionella pneumophila*, followed by *Escherichia coli* (12.75%) and *Pseudomonas syringae* (9.76%). The distribution of collected effector proteins across different species is shown in Fig. 3. These effectors were either published or have been previously used in positive data sets for training machine learning models in past computational studies. This distribution across species is currently biased, which is probably due to two factors. The first one is the biased historical research interest. For example, much of the early work on the T3SS focused on *Salmonella enterica* serovars to establish this organism as the model for T3SE discovery[6,46]. The second "bias" derives from the prevalence of effector proteins in some species. For example, recent comprehensive surveys suggest more than three hundred T4SEs are encoded in the genome of some strains of *Legionella pneumophila*[43,47–49].

For all these three secretion systems (T3SS, T4SS and T6SS), the targeting information in the effectors has remained nebulous. Several previous studies using genetics and biochemistry to analyse specific effector proteins of interest suggest that the N- or C-terminal region may carry the targeting information. In the case of *Legionella*, some of the T4SEs depend on a C-terminal targeting signal, a dedicated molecular chaperone (icmS and W), or both in order to be secreted[43]. By way of demonstrating the utility of having a large, experimentally validated set of effector protein sequences to raise hypotheses about T4SS function, the composition and putative preference for conserved amino acids located at the both N- and C-termini of the collected entries was addressed.

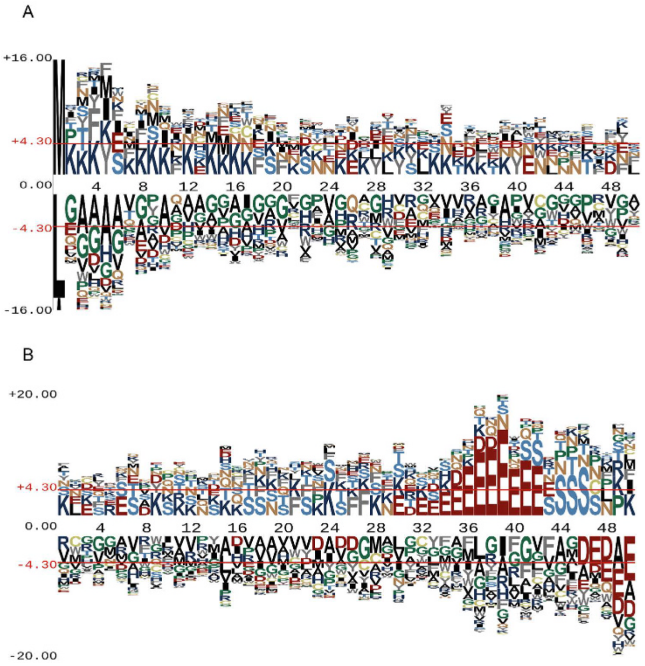
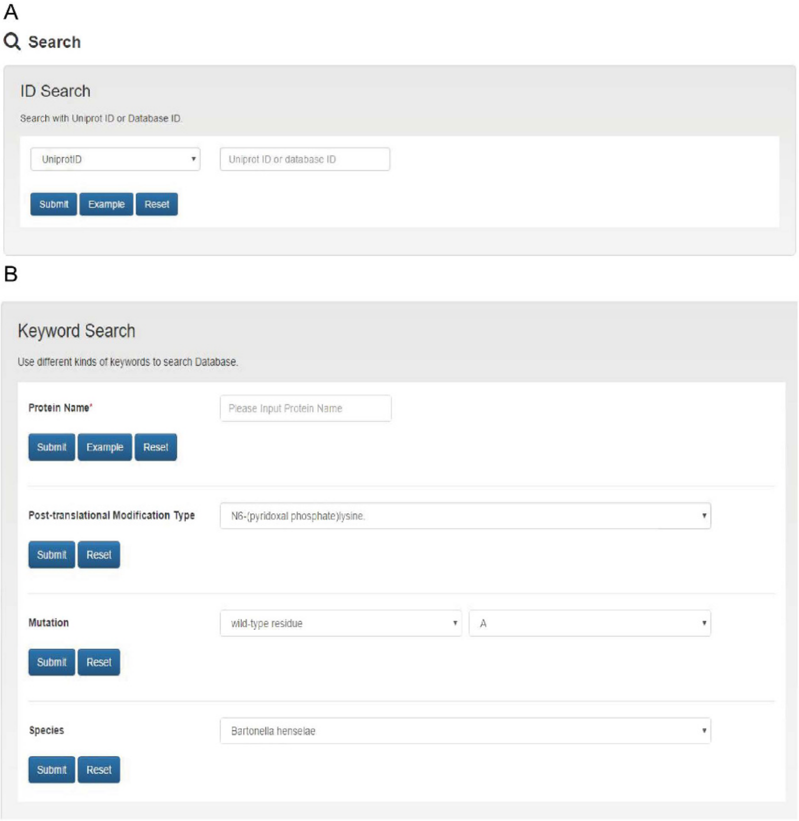


Figure 4. Sequence logos showing the amino acid conservation and preference in T4SEs. Sequence Logo plots of the indicated number of residues in the N-terminal (**A**) and C-terminal (**B**) regions of the collected sequences of T4SEs from *Legionella pneumophila*. The x-axis represents residue numbers, and Amino acids above the *x*-axis are favoured while those underneath the *x*-axis are disfavoured at the corresponding positions. Note that because of the mechanism of protein synthesis, the N-terminal position of a bacterial protein can only ever be methionine (M), isoleucine (I) or leucine (L), with M being vastly the most common.

The collection of sequences offered the opportunity to look for dominant residues in an effector collection (e.g. T4SEs) and within a given species (e.g. the T4SEs in *Legionella*).

The background dataset for statistical analysis was based on protein sequences obtained by searching UniProt with "Legionella protein" as the keywords. For each type of effector proteins, the motifs from N- and C-termini were extracted using a window size of 50 amino acids[28,29]. The sequence conservation for T4SEs is depicted using pLogo[50] for the N- terminal motifs and C-terminal motifs of the single dominant species *Legionella* (Fig. 4). Excluding the translation-initiating N-terminal methionine (M) from position 1 (Fig. 4A), two observations become apparent. Firstly, in the case of the C-terminal motifs, there is a striking confirmation of the preponderance of glutamate (E) at positions -9 to -16 for the T4SEs. Furthermore, it becomes clear that there is a strong dis-favoring of glutamate and the other acidic amino acid aspartate (D) from the final five positions at the C-terminus of the sequences (Fig. 4B). These signatures impact on protein translocation[43,51]. Secondly, there is a preponderance of lysine (K) residues for 3–4 positions, reoccurring through the N-terminal segment (Fig. 4A). For practical reasons, the alignments are made from position 1, which is an artificial means to register the sequences. Given this, the observed distribution would occur if a periodical presence of lysine e.g. occurring on an aligned face of an alpha-helical segment, were part of a consensus sequence important for recognition and/or translocation. Glycine (G), alanine (A) and proline (P) residues tend to be dis-favored in the N-terminal segments (Fig. 4A), which would be consistent with a helical structure being an important feature of the T4SE. As several effectors rely on the dot/icm chaperones IcmS and W for efficient translocation by T4SS, a reasonable hypothesis would be that these conserved sequence features in a secondary structure context serve as binding sites for chaperones such as IcmS and W. To provide an overview of sequence preferences for N- and C-terminal segment of all the three types of effectors, we also generated sequence logo representations for all the collected entries of T3SS, T4SS and T6SS (Supplementary Figure S1).

Database contents. For all entries in SecretEPDB, we extracted, manually checked and integrated their annotations from several publicly available databases, including UniProt[36], NCBI Protein database (http://www.ncbi.nlm.nih.gov/protein), Pfam[52], KEGG (Kyoto Encyclopedia of Genes and Genomes)[39] and PDB[53]. From the UniProt database information was extracted forprotein accession number, protein name, bacterial species and functional annotations. We also annotated protein secondary structures, by mapping each entry onto the PDB database using BLAST search. For each structure included in SecretEPDB, an overview snapshot is provided for the structure. In addition to conserved structural elements, protein disordered regions can also be crucial for protein function[54], with some protein regions being intrinsically disordered and lacking structural information[55]. Given that protein disordered regions may be functionally important[54,56], we used the popular bioinformatics tool VSL2B[57] to predict natively disordered regions. Where available, disordered region prediction results from the Database of Disordered Protein Prediction (D^2P^2)[58] are provided for protein entries. This provides a general



ID Search

Search with Uniprot ID or Database ID.

UniprotID | Uniprot ID or database ID

Keyword Search

Use different kinds of keywords to search Database.

Protein Name*: Please Input Protein Name

Post-translational Modification Type: N6-(pyridoxal phosphate)lysine.

Mutation: wild-type residue | A

Species: Bartonella henselae

Figure 5. Examples of search options available in SecretEPDB. (**A**) Search option with UniProt ID or SecretEPDB ID; (**B**) Search option with a number of keywords, including protein name, mutation and species.

overview of ordered and disordered regions in those proteins. To better display the protein context information, we employed IBS[37] to present and visualize functional sites and domains in an integrative manner. These functional sites and domains were retrieved using the UniProt accession number for each entry and then used for plotting figures by IBS.

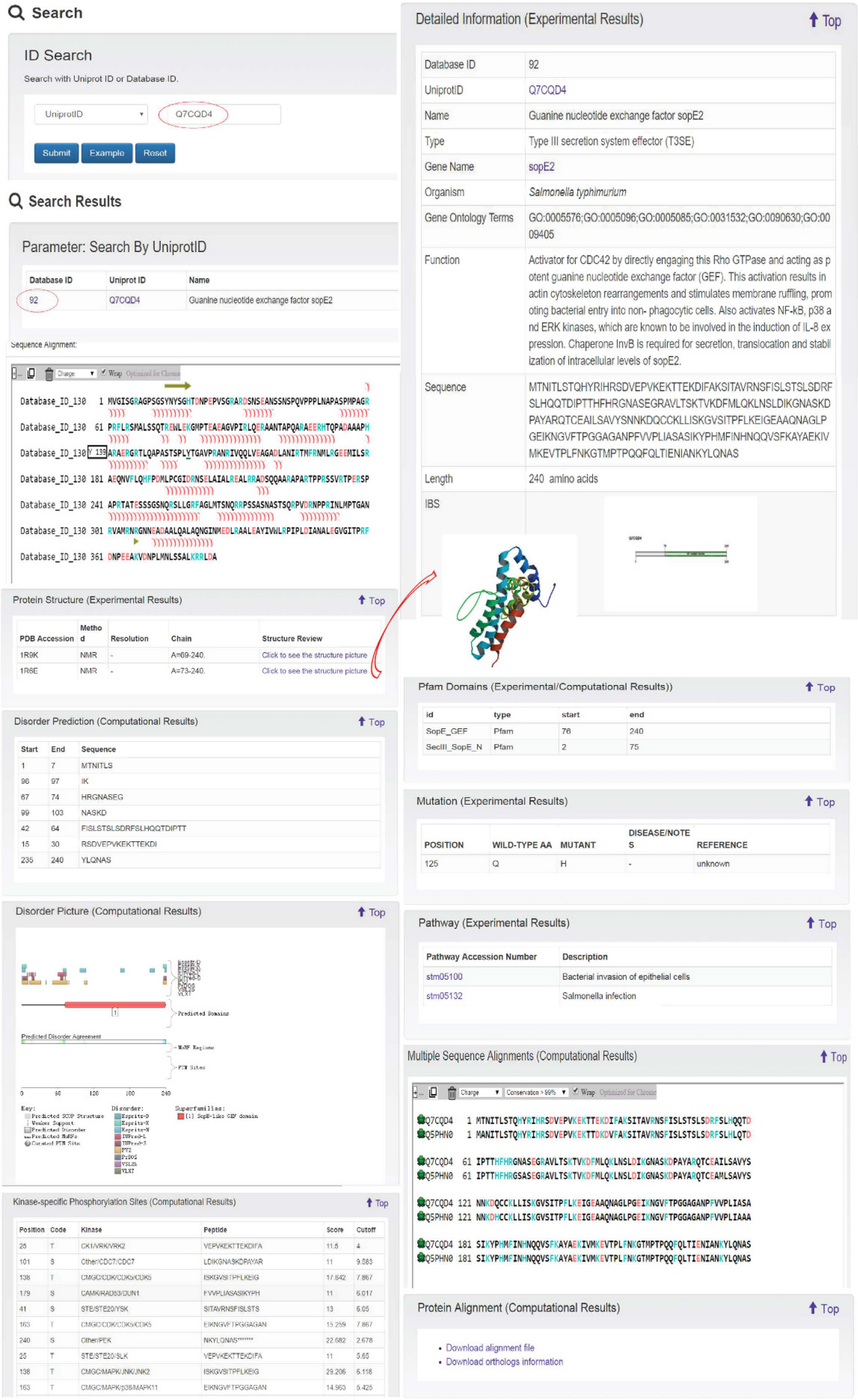
To capture related but non-identical sequence relationships between effectors, both sequence modules and MSAs of each entry were generated and annotated using Strap[59] in an interactive manner in SecretEPDB. The sequence module provides the amino acid sequence augmented by predicted secondary structure inferred by the SSpro program[60] included in the SCRATCH suite. The MSAs were generated by Clustal Omega[61] based on the homologous sequences of each entry, which were retrieved using PSI-BLAST search against the Swiss-Prot database (with an e-value < 0.0001 and sequence identity > 0.8).

The entries in SecretEPDB are deposited in a MySQL relational database. Several website development techniques (including jQuery, Bootstrap, JAVA, Structs and Hibernate) were utilized to implement SecretEPDB, enabling the design of a user-friendly interface, multiple functionalities and enhanced data visualization.

Database utility. SecretEPDB provides a number of functionalities to optimize the user experience including database searches, browsing, download and new entry submission. A webpage is also available to provide a statistical overview of the current entries in SecretEPDB in terms of the secretion system types, bacterial species (http://secretepdb.erc.monash.edu/statistics.jsp).

There are in total currently 2142 proteins in SecretEPDB. The search webpage (http://secretepdb.erc.monash.edu/getDropDownList.action) allows users to search these entries in SecretEPDB in two different ways, i.e. search with the ID and keyword (Fig. 5). For search with the ID, SecretEPDB provides two alternative IDs: UniProt ID and SecretEPDB ID. The former is composed of 6 letters and digits, whereas the latter is drawn from a range of consecutive integers. Searching with the keyword is also straightforward. Several types of keywords are provided including protein name, mutation and bacterial species. For these different search options, a corresponding example is available for guiding users to search the database. By clicking the 'Example' button, users can promptly get the example keyword provided by SecretEPDB. After selecting the "Submit" button, the corresponding search results will be displayed at the result webpage. Users can click a database ID to visualize the detailed information of the current entry. Note that if a protein entry was originally extracted from the UniProt database, then a link pointing to the corresponding UniProt webpage is also provided at the search result webpage. Users click the UniProt ID to transfer to the corresponding UniProt webpage of this entry.

By way of example, binding of *Salmonella* to the surface of intestinal epithelial cells activates a T3SS that secretes several T3SEs including SopE, SopE2 and SopB: these effectors mimic host cell proteins and thereby



Parameter: Search By UniprotID

Database ID	Uniprot ID	Name
92	Q7CQD4	Guanine nucleotide exchange factor sopE2

Sequence Alignment:

```
Database_ID_130   1 MVGISGRAGPSGSYNYSGHTDNPEPVSGRARDSNSEANSSNSPQVPPPLNAPASPMPAGR
Database_ID_130  61 PRFLRSMALSSQTREWLEKGMPTEAEAGVPIRLQERAANTAPQARAEERHTQPADAAAPH
Database_ID_130Y 139ARAERGRTLQAPASTSPLYTGAVPRANRIVQQLVEAGADLANIRTMFRNMLRGEEMILSR
Database_ID_130 181 AEQNVFLQHFPDMLPCGIDRNSELAIALREALRRADSQQAARAPARTPPRSSVRTPERSP
Database_ID_130 241 APRTATESSSGSNQRSLLGRFAGLMTSNQRRPSSASNASTSQRPVDRNPPRINLMPTGAN
Database_ID_130 301 RVAMRNRGNNEADAALQALAQNGINMEDLRAALEAYIVWLRPIPLDIANALEGVGITPRF
Database_ID_130 361 DNPEEAKVDNPLMNLSSALKRRLDA
```

Protein Structure (Experimental Results) ↑ Top

PDB Accession	Method	Resolution	Chain	Structure Review
1R9K	NMR	-	A=69-240.	Click to see the structure picture
1R6E	NMR	-	A=73-240.	Click to see the structure picture

Disorder Prediction (Computational Results) ↑ Top

Start	End	Sequence
1	7	MTNITLS
96	97	IK
67	74	HRGNASEG
99	103	NASKD
42	64	FISLSTSLSDRFSLHQQTDIPTT
15	30	RSDVEPVKEKTTEKDI
235	240	YLQNAS

Disorder Picture (Computational Results) ↑ Top

Kinase-specific Phosphorylation Sites (Computational Results) ↑ Top

Position	Code	Kinase	Peptide	Score	Cutoff
25	T	CK1/VRK/VRK2	VEPVKEKTTEKDIFA	11.5	4
101	S	Other/CDC7/CDC7	LDIKGNASKDPAYAR	11	9.583
138	T	CMGC/CDK/CDK5/CDK5	ISKGVSITPFLKEIG	17.642	7.867
179	S	CAMK/RAD53/DUN1	FVVPLIASASIKYPH	11	6.017
41	S	STE/STE20/YSK	SITAVRNSFISLSTS	13	6.05
163	T	CMGC/CDK/CDK5/CDK5	EIKNGVFTPGGAGAN	15.259	7.867
240	S	Other/PEK	NKYLQNAS*******	22.682	2.678
25	T	STE/STE20/SLK	VEPVKEKTTEKDIFA	11	5.65
138	T	CMGC/MAPK/JNK/JNK2	ISKGVSITPFLKEIG	29.206	6.118
163	T	CMGC/MAPK/p38/MAPK11	EIKNGVFTPGGAGAN	14.963	5.425

Detailed Information (Experimental Results) ↑ Top

Database ID	92
UniprotID	Q7CQD4
Name	Guanine nucleotide exchange factor sopE2
Type	Type III secretion system effector (T3SE)
Gene Name	sopE2
Organism	*Salmonella typhimurium*
Gene Ontology Terms	GO:0005576;GO:0005096;GO:0005085;GO:0031532;GO:0090630;GO:0009405
Function	Activator for CDC42 by directly engaging this Rho GTPase and acting as potent guanine nucleotide exchange factor (GEF). This activation results in actin cytoskeleton rearrangements and stimulates membrane ruffling, promoting bacterial entry into non- phagocytic cells. Also activates NF-kB, p38 and ERK kinases, which are known to be involved in the induction of IL-8 expression. Chaperone InvB is required for secretion, translocation and stabilization of intracellular levels of sopE2.
Sequence	MTNITLSTQHYRIHRSDVEPVKEKTTEKDIFAKSITAVRNSFISLSTSLSDRFSLHQQTDIPTTHFHRGNASEGRAVLTSKTVKDFMLQKLNSLDIKGNASKDPAYARQTCEAILSAVYSNNKDQCCKLLISKGVSITPFLKEIGEAAQNAGLPGEIKNGVFTPGGAGANPFVVPLIASASIKYPHMFINHNQQVSFKAYAEKIVMKEVTPLFNKGTMPTPQQFQLTIENIANKYLQNAS
Length	240 amino acids
IBS	

Pfam Domains (Experimental/Computational Results)) ↑ Top

id	type	start	end
SopE_GEF	Pfam	76	240
SecIII_SopE_N	Pfam	2	75

Mutation (Experimental Results) ↑ Top

POSITION	WILD-TYPE AA	MUTANT	DISEASE/NOTES	REFERENCE
125	Q	H	-	unknown

Pathway (Experimental Results) ↑ Top

Pathway Accession Number	Description
stm05100	Bacterial invasion of epithelial cells
stm05132	Salmonella infection

Multiple Sequence Alignments (Computational Results) ↑ Top

```
Q7CQD4    1 MTNITLSTQHYRIHRSDVEPVKEKTTEKDIFAKSITAVRNSFISLSTSLSDRFSLHQQTD
Q5PHN0    1 MANITLSTQHYRIHRSDVEPVKEKTTDKDVFAKSITAVRNSFISLSTSLSDRFSLHLQTD

Q7CQD4   61 IPTTHFHRGNASEGRAVLTSKTVKDFMLQKLNSLDIKGNASKDPAYARQTCEAILSAVYS
Q5PHN0   61 IPTTHFHRGSASEGRAVLTSKTVKDFMLQKLNSLDIKGNASKDPAYARQTCEAMLSAVYS

Q7CQD4  121 NNKDQCCKLLISKGVSITPFLKEIGEAAQNAGLPGEIKNGVFTPGGAGANPFVVPLIASA
Q5PHN0  121 NNKDHCCKLLISKGVSITPFLKEIGEAAQNAGLPGEIKNGVFTPGGAGANPFVVPLIAAA

Q7CQD4  181 SIKYPHMFINHNQQVSFKAYAEKIVMKEVTPLFNKGTMPTPQQFQLTIENIANKYLQNAS
Q5PHN0  181 SIKYPHMFINHNQQVSFKAYAEKIVMKEVTPLFNKGTMPTPQQFQLTIENIANKYLQNAS
```

Protein Alignment (Computational Results) ↑ Top

- Download alignment file
- Download orthologs information

Figure 6. Output of the sample search against SecretEPDB using UniProt ID "Q7CQD4" as the query. The results are displayed and organized by different annotation categories, including protein detailed information, sequence alignment, protein structure, multiple sequence alignments, Pfam domain, disorder region prediction, disorder picture, protein mutation and metabolic/signaling pathway.

activate host cell regulatory proteins to initiate actin cytoskeleton rearrangements. For SopE2, first discovered in more than ten years ago[62–64], this occurs because it mimics host cell guanine nucleotide exchange factors, or GEFs. Users with an interest in this biological phenomena who are investigating the UniProt ID "Q7CQD4", corresponding to SopE2will access results that are displayed and organized according to the major annotation categories (Fig. 6).

In order to simplify entry browsing, each entry can be displayed in accordance with their type (i.e. T3SE, T4SE or T6SE) at the browse webpage. Users can download the entire database of SecretEPDB in the SQL format. Alternatively, protein structures and MSAs of the entries are available for download. To collect the up-to-date experimental findings of secretion effectors, we provide an option for researchers to submit their recent results to SecretEPDB via an online entry submission webpage (available at http://secretepdb.erc.monash.edu.au/submission.jsp). At the "Submission" webpage, two submission modules (quick submission and formal submission) are available for users to submit their recently discovered effectors to SecretEPDB. When using the 'quick submission' module, users can simply submit a new effector protein by providing brief information, such as subject (describing the protein name or identity) and description (providing the UniProt accession/link, PubMed ID/link or the title of the literature paper if possible). After successfully receiving the request, the database administrator will then carefully review the submission and accordingly update the database after verification. When using the 'formal submission' module, users are required to provide more detailed information necessary for annotating the entries that they would like to submit. Such information includes contact information, and protein general information, including protein name, species, gene name, molecular weight, effector type, protein sequence, etc. Users are also encouraged to provide additional (optional) information such as Uniprot ID, protein structural and functional annotations. Each submission will be subject to further scrutiny prior to being included in the database and made publicly available. Furthermore, our database team will regularly maintain and update the database by means of searching recently published literature papers and keeping track of the updates of UniProt and PubMed.

We have also made available a 'Timeline' module (http://secretepdb.erc.monash.edu.au/timeline.action), through which users can readily view the information of each major recent update. This enables users to rapidly track recent update history, time and entries included in all recent major updates.

Conclusion

In this work, we develop a new web-based knowledgebase, termed SecretEPDB, which provides comprehensive annotations of effector proteins of three major bacterial secretion systems T3SS, T4SS and T6SS. The annotations provided by SecretEPDB include protein functional annotation, protein 3D structure, Pfam domains, metabolic pathways, and protein evolutionary information. All entries documented in SecretEPDB have been manually annotated and experimentally verified. We anticipate that SecretEPDB will be a useful resource for generating novel hypothesis of the translocation mechanisms and function of secretion effectors and contribute to a better understanding of the functional characterization of these proteins and their corresponding secretion systems.

References

1. Russell, A. B., Peterson, S. B. & Mougous, J. D. Type VI secretion system effectors: poisons with a purpose. *Nature reviews. Microbiology* **12,** 137–148, doi: 10.1038/nrmicro3185 (2014).
2. Costa, T. R. *et al.* Secretion systems in Gram-negative bacteria: structural and mechanistic insights. *Nature reviews. Microbiology* **13,** 343–359, doi: 10.1038/nrmicro3456 (2015).
3. Chang, J. H., Desveaux, D. & Creason, A. L. The ABCs and 123s of bacterial secretion systems in plant pathogenesis. *Annual review of phytopathology* **52,** 317–345, doi: 10.1146/annurev-phyto-011014-015624 (2014).
4. Durand, E., Cambillau, C., Cascales, E. & Journet, L. VgrG, Tae, Tle, and beyond: the versatile arsenal of Type VI secretion effectors. *Trends in microbiology* **22,** 498–507, doi: 10.1016/j.tim.2014.06.004 (2014).
5. Economou, A. *et al.* Secretion by numbers: Protein traffic in prokaryotes. *Molecular microbiology* **62,** 308–319, doi: 10.1111/j.1365-2958.2006.05377.x (2006).
6. Galan, J. E., Lara-Tejero, M., Marlovits, T. C. & Wagner, S. Bacterial type III secretion systems: specialized nanomachines for protein delivery into target cells. *Annual review of microbiology* **68,** 415–438, doi: 10.1146/annurev-micro-092412-155725 (2014).
7. Pearson, J. S., Zhang, Y., Newton, H. J. & Hartland, E. L. Post-modern pathogens: surprising activities of translocated effectors from E. coli and Legionella. *Current opinion in microbiology* **23,** 73–79, doi: 10.1016/j.mib.2014.11.005 (2015).
8. Martinez-Garcia, P. M., Ramos, C. & Rodriguez-Palenzuela, P. T346Hunter: a novel web-based tool for the prediction of type III, type IV and type VI secretion systems in bacterial genomes. *PloS one* 10, e0119317, doi: 10.1371/journal.pone.0119317 (2015).
9. McGuckin, M. A., Linden, S. K., Sutton, P. & Florin, T. H. Mucin dynamics and enteric pathogens. *Nature reviews. Microbiology* **9,** 265–278, doi: 10.1038/nrmicro2538 (2011).
10. Wandersman, C. Concluding remarks on the special issue dedicated to bacterial secretion systems: function and structural biology. *Research in microbiology* **164,** 683–687, doi: 10.1016/j.resmic.2013.03.008 (2013).
11. Block, A. & Alfano, J. R. Plant targets for Pseudomonas syringae type III effectors: virulence targets or guarded decoys? *Current opinion in microbiology* **14,** 39–46, doi: 10.1016/j.mib.2010.12.011 (2011).
12. Zechner, E. L., Lang, S. & Schildbach, J. F. Assembly and mechanisms of bacterial type IV secretion machines. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **367,** 1073–1087, doi: 10.1098/rstb.2011.0207 (2012).
13. Basler, M. Type VI secretion system: secretion by a contractile nanomachine. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **370,** doi: 10.1098/rstb.2015.0021 (2015).
14. Yang, X., Guo, Y., Luo, J., Pu, X. & Li, M. Effective identification of Gram-negative bacterial type III secreted effectors using position-specific residue conservation profiles. *PloS one* **8,** e84439, doi: 10.1371/journal.pone.0084439 (2013).
15. Cascales, E. & Christie, P. J. The versatile bacterial type IV secretion systems. *Nature Reviews Microbiology* **1,** 137–149 (2003).
16. Souza, R. C. *et al.* AtlasT4SS: a curated database for type IV secretion systems. *BMC microbiology* **12,** 172, doi: 10.1186/1471-2180-12-172 (2012).
17. Ilangovan, A., Connery, S. & Waksman, G. Structural biology of the Gram-negative bacterial conjugation systems. *Trends in microbiology* **23,** 301–310, doi: 10.1016/j.tim.2015.02.012 (2015).
18. Tseng, T.-T., Tyler, B. M. & Setubal, J. C. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC microbiology* **9,** S2 (2009).
19. Altindis, E., Dong, T., Catalano, C. & Mekalanos, J. Secretome analysis of Vibrio cholerae type VI secretion system reveals a new effector-immunity pair. *mBio* **6,** e00075, doi: 10.1128/mBio.00075-15 (2015).

20. Arnold, R. *et al.* Sequence-based prediction of type III secreted proteins. *PLoS pathogens* **5,** e1000376, doi: 10.1371/journal.ppat.1000376 (2009).
21. Dong, X., Zhang, Y. J. & Zhang, Z. Using weakly conserved motifs hidden in secretion signals to identify type-III effectors from bacterial pathogen genomes. *PloS one* **8,** e56632, doi: 10.1371/journal.pone.0056632 (2013).
22. McDermott, J. E. *et al.* Computational prediction of type III and IV secreted effectors in gram-negative bacteria. *Infection and immunity* **79,** 23–32, doi: 10.1128/IAI.00537-10 (2011).
23. Pukatzki, S., McAuley, S. B. & Miyata, S. T. The type VI secretion system: translocation of effectors and effector-domains. *Current opinion in microbiology* **12,** 11–17, doi: 10.1016/j.mib.2008.11.010 (2009).
24. Salomon, D. *et al.* Marker for type VI secretion system effectors. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 9271–9276, doi: 10.1073/pnas.1406110111 (2014).
25. Shrivastava, S. & Mande, S. S. Identification and functional characterization of gene components of Type VI Secretion system in bacterial genomes. *PloS one* **3,** e2955, doi: 10.1371/journal.pone.0002955 (2008).
26. Voth, D. E., Broederdorf, L. J. & Graham, J. G. Bacterial Type IV secretion systems: versatile virulence machines. *Future microbiology* **7,** 241–257, doi: 10.2217/fmb.11.150 (2012).
27. Wang, Y., Sun, M., Bao, H. & White, A. P. T3_MM: a Markov model effectively classifies bacterial type III secretion signals. *PloS one* **8,** e58173, doi: 10.1371/journal.pone.0058173 (2013).
28. Wang, Y., Zhang, Q., Sun, M. A. & Guo, D. High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics* **27,** 777–784, doi: 10.1093/bioinformatics/btr021 (2011).
29. Zou, L., Nan, C. & Hu, F. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics* **29,** 3135–3142, doi: 10.1093/bioinformatics/btt554 (2013).
30. An, Y. *et al.* Comprehensive assessment and performance improvement of predictors for effector proteins of bacterial secretion systems III, IV, and VI. *Briefings in Bioinformatics* in press (2016).
31. Bi, D. *et al.* SecReT4: a web-based bacterial type IV secretion system resource. *Nucleic acids research* **41,** D660–665, doi: 10.1093/nar/gks1248 (2013).
32. Dong, X., Lu, X. & Zhang, Z. BEAN 2.0: an integrated web resource for the identification and functional analysis of type III secreted effectors. *Database: the journal of biological databases and curatio*n, bav064, doi: 10.1093/database/bav064 (2015).
33. Li, J. *et al.* SecReT6: a web-based resource for type VI secretion systems found in bacteria. *Environmental microbiology* **17,** 2196–2202, doi: 10.1111/1462-2920.12794 (2015).
34. Wang, Y., Huang, H., Sun, M. a., Zhang, Q. & Guo, D. T3DB: an integrated database for bacterial type III secretion system. *BMC bioinformatics* **13,** 66 (2012).
35. Huang, Y. H., Rose, P. W. & Hsu, C. N. Citing a Data Repository: A Case Study of the Protein Data Bank. *PloS one* **10,** e0136631, doi: 10.1371/journal.pone.0136631 (2015).
36. UniProt, C. UniProt: a hub for protein information. *Nucleic acids research* **43,** D204–212, doi: 10.1093/nar/gku989 (2015).
37. Liu, W. *et al.* IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* **31,** 3359–3361, doi: 10.1093/bioinformatics/btv362 (2015).
38. Gille, C., Birgit, W. & Gille, A. Sequence alignment visualization in HTML5 without Java. *Bioinformatics* **30,** 121–122, doi: 10.1093/bioinformatics/btt614 (2014).
39. Kanehisa, M. *et al.* KEGG_Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research* **42,** D199–205, doi: 10.1093/nar/gkt1076 (2014).
40. Xue, Y. *et al.* GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Molecular & cellular proteomics* **7,** 1598–1608 (2008).
41. Tay, D. M. *et al.* T3SEdb: data warehousing of virulence effectors secreted by the bacterial Type III Secretion System. *BMC bioinformatics* **11** Suppl 7, S4, doi: 10.1186/1471-2105-11-S7-S4 (2010).
42. Wang, Y., Wei, X., Bao, H. & Liu, S.-L. Prediction of bacterial type IV secreted effectors by C-terminal features. *BMC genomics* **15,** 1 (2014).
43. Lifshitz, Z. *et al.* Computational modeling and experimental validation of the Legionella and Coxiella virulence-related type-IVB secretion signal. *Proceedings of the National Academy of Sciences* **110,** E707–E715 (2013).
44. Russell, A. B. *et al.* A widespread bacterial type VI secretion effector superfamily identified using a heuristic approach. *Cell host & microbe* **11,** 538–549 (2012).
45. Russell, A. B. *et al.* Diverse type VI secretion phospholipases are functionally plastic antibacterial effectors. *Nature* **496,** 508–512 (2013).
46. Raymond, B. *et al.* Subversion of trafficking, apoptosis, and innate immunity by type III secretion system effectors. *Trends in microbiology* **21,** 430–441 (2013).
47. Dolezal, P. *et al.* Legionella pneumophila secretes a mitochondrial carrier protein during infection. *PLoS pathogens* **8,** e1002459, doi: 10.1371/journal.ppat.1002459 (2012).
48. Zhu, W. *et al.* Comprehensive identification of protein substrates of the Dot/Icm type IV transporter of Legionella pneumophila. *PloS one* **6,** e17638, doi: 10.1371/journal.pone.0017638 (2011).
49. Ensminger, A. W. Legionella pneumophila, armed to the hilt: justifying the largest arsenal of effectors in the bacterial world. *Current opinion in microbiology* **29,** 74–80 (2016).
50. O'Shea, J. P. *et al.* pLogo: a probabilistic approach to visualizing sequence motifs. *Nature methods* **10,** 1211–1212, doi: 10.1038/nmeth.2646 (2013).
51. Burstein, D. *et al.* Genome-scale identification of Legionella pneumophila effectors using a machine learning approach. *PLoS pathogens* **5,** e1000508 (2009).
52. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic acids research* **42,** D222–230, doi: 10.1093/nar/gkt1223 (2014).
53. Rose, P. W. *et al.* PDB_The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic acids research* **39,** D392–401, doi: 10.1093/nar/gkq1021 (2011).
54. Cheng, J., Sweredoski, M. J. & Baldi, P. Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Mining and Knowledge Discovery* **11,** 213–222, doi: 10.1007/s10618-005-0001-y (2005).
55. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nature reviews Molecular cell biology* **6,** 197–208 (2005).
56. Jones, D. T. & Ward, J. J. Prediction of disordered regions in proteins from position specific score matrices. *Proteins* **53** Suppl 6, 573–578, doi: 10.1002/prot.10528 (2003).
57. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. VSL2B_Length-dependent prediction of protein intrinsic disorder. *BMC bioinformatics* **7,** 208, doi: 10.1186/1471-2105-7-208 (2006).
58. Oates, M. E. *et al.* D2P2: database of disordered protein predictions. *Nucleic acids research*, gks1226 (2012).
59. Gille, C., Fahling, M., Weyand, B., Wieland, T. & Gille, A. Alignment-Annotator web server: rendering and annotating sequence alignments. *Nucleic acids research* **42,** W3–6, doi: 10.1093/nar/gku400 (2014).
60. Magnan, C. N. & Baldi, P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* **30,** 2592–2597, doi: 10.1093/bioinformatics/btu352 (2014).
61. Sievers, F. *et al.* Clustal Omega_Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7,** 539, doi: 10.1038/msb.2011.75 (2011).

62. Bakshi, C. *et al.* Identification of SopE2, a Salmonellasecreted protein which is highly homologous to SopE and involved in bacterial invasion of epithelial cells. *Journal of bacteriology* **182,** 2341–2344 (2000).
63. Cherayil, B. J., McCormick, B. A. & Bosley, J. Salmonella enterica serovar typhimurium-dependent regulation of inducible nitric oxide synthase expression in macrophages by invasins SipB, SipC, and SipD and effector SopE2. *Infection and immunity* **68,** 5567–5574 (2000).
64. Stender, S. *et al.* Identification of SopE2 from Salmonella typhimurium, a conserved guanine nucleotide exchange factor for Cdc42 of the host cell. *Molecular microbiology* **36,** 1206–1221 (2000).

Acknowledgements

We thank Terry Kwok and John Boyce for expert feedback on the database. G.I.W. is a recipient of the Australian Research Council (ARC) Discovery Outstanding Researcher Award (DORA). T.L. is an ARC Australian Laureate Fellow. This work was supported by the NHMRC Program in Cellular Microbiology (Grant 1092262).

Author Contributions

Y.A., J.W. and J.S. conducted the research and experimentation, prepared the draft figures, performed data collection, computational analyses and implemented the web server. C.L., J.R., Y.Z., T.N., M.H., T.A., G.I.W., J.S. provided expertise for the analysis of all data and reviewed the web server. Y.Z., G.I.W., J.S. and T.L. conceived and designed the project, managed communication between co-authors, checked all data and wrote the paper.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: An, Y. *et al.* SecretEPDB: a comprehensive web-based resource for secreted effector proteins of the bacterial types III, IV and VI secretion systems. *Sci. Rep.* **7**, 41031; doi: 10.1038/srep41031 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.