D526–D534 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 30 October 2019*
doi: 10.1093/nar/gkz924

ChlamDB: a comparative genomics database of the phylum *Chlamydiae* and other members of the *Planctomycetes-Verrucomicrobiae-Chlamydiae* superphylum

Trestan Pillonel*, Florian Tagini, Claire Bertelli and Gilbert Greub

Institute of Microbiology, Lausanne University Hospital and University of Lausanne, Bugnon 48, 1011 Lausanne, Switzerland

Received August 15, 2019; Revised October 04, 2019; Editorial Decision October 04, 2019; Accepted October 08, 2019

ABSTRACT

ChlamDB is a comparative genomics database containing 277 genomes covering the entire *Chlamydiae* phylum as well as their closest relatives belonging to the *Planctomycetes-Verrucomicrobiae-Chlamydiae* (PVC) superphylum. Genomes can be compared, analyzed and retrieved using accessions numbers of the most widely used databases including COG, KEGG ortholog, KEGG pathway, KEGG module, Pfam and InterPro. Gene annotations from multiple databases including UniProt (curated and automated protein annotations), KEGG (annotation of pathways), COG (orthology), TCDB (transporters), STRING (protein–protein interactions) and InterPro (domains and signatures) can be accessed in a comprehensive overview page. Candidate effectors of the Type III secretion system (T3SS) were identified using four *in silico* methods. The identification of orthologs among all PVC genomes allows users to perform large-scale comparative analyses and to identify orthologs of any protein in all genomes integrated in the database. Phylogenetic relationships of PVC proteins and their closest homologs in RefSeq, comparison of transmembrane domains and Pfam domains, conservation of gene neighborhood and taxonomic profiles can be visualized using dynamically generated graphs, available for download. As a central resource for researchers working on chlamydia, chlamydia-related bacteria, verrucomicrobia and planctomyces, ChlamDB facilitates the access to comprehensive annotations, integrates multiple tools for comparative genomic analyses and is freely available at https://chlamdb.ch/. Database URL: https://chlamdb.ch/

INTRODUCTION

All known members of the phylum *Chlamydiae* are obligate intracellular bacteria exhibiting a unique life cycle. Described chlamydial species cause a broad range of diseases in various species of birds, fishes, reptiles, amphibians, marsupials and mammals (1), and include major human pathogens such as *Chlamydia trachomatis*—a leading cause of blindness and infertility (1,2). *Chlamydiae* are difficult to cultivate and genetic manipulations are only available for a few species, which drastically slows down the understanding of their fascinating biology. Other members of the *Planctomycetes-Verrucomicrobiae-Chlamydiae* (PVC) superphylum include the closest relatives of the *Chlamydiae*: The *Planctomycetes* are extremely attractive for the field of evolutionary cell biology given their peculiar intracellular compartments (3). Like *Chlamydiae*, they replicate using an FtsZ-independent mechanism but contrarily to the *Chlamydiae*, *Planctomycetales* were shown to have a complete peptidoglycan cell wall (4–7). There is currently no database allowing an easy access and comparison of comprehensive genomics data for members of the PVC superphylum. A database focusing on the curation of chlamydial genome annotation was recently published (8), but it is limited to three species of the genus *Chlamydia*. A phylum-scale perspective including comparative data with the closest free-living relatives of the *Chlamydiae* would provide significant added value for the research community given the conserved intracellular lifestyle of these bacteria that were estimated to diverge over 700 million years ago (9). The PVCbase (10) provides updated automated protein annotations of forty-two PVC genomes, but only offers limited browsing capabilities and no comparative data. ChlamDB offers a centralized resource for genomic data and annotations of the entire PVC-superphylum. Its simple search engine allows browsing protein annotations, identifying orthologs in PVC genomes and performing a variety of comparative analyses.

*To whom correspondence should be addressed. Tel: +41 21 314 49 79; Fax: + 41 21 314 40 60; Email: trestan.pillonel@chuv.ch

Table 1. Overview of ChlamDB content

Phylum	# genomes	# species
Chlamydiae	221	48
Planctomycetes	20	20
Verrucomicrobia	34	12
Lentisphaerae	1	1
Kiritimatiellaeota	1	1
TOTAL	**277**	**82**

Genomic data and search

ChlamDB release 2.0 integrates data from 277 PVC genomes of 82 different species (Table 1), retrieved from GenBank (11) or RefSeq (12) (when GenBank records were not annotated). It includes all complete PVC genomes as well as draft genomes of the *Chlamydiae* phylum to increase the diversity of genera and species represented in the database. Draft genomes of the most studied *Chlamydia* species were discarded to reduce unnecessary redundancy in the database. Most genomes ($n = 221$) belong to the *Chlamydiae* phylum, including 86 *C. trachomatis*, 20 *Chlamydia muridarum*, 20 *Chlamydia psittaci* and 12 *Chlamydophila pneumoniae* genomes, thus allowing intra-species comparison for these important human pathogens. Species-level diversity was shown to determine *C. trachomatis* tissue tropism, hence showing the interest of such comparisons to elucidate novel aspects of chlamydial lifestyle and pathogenesis. To allow for broader comparisons, this database also contains the genomes of 34 *Verrucomicrobia,* 20 *Planctomycetes*, 1 *Lentisphaerae* and 1 *Kiritimatiellaeota*. Among the 34 *Verrucomicrobia*, there are 23 *Akkermansia muciniphila*, a bacterium commonly found in the human gut (13).

The database provides various tools for comparing, analyzing and retrieving genomic data. A simple Boolean search interface allows querying the database for specific entries using NCBI protein accessions and locus tags or UniProt accessions. Accessions numbers of widely-used databases such as COG (14), KEGG ortholog (KO) (15), KEGG pathway (16), KEGG module, Pfam (17) and InterPro (18) are also recognized and can be used to search for proteins with specific annotations. The annotation of individual genomes can be browsed in tables of genes that are accessible directly from the front web page. In addition, sequence homology searches can be performed through a BLAST interface integrating the different blast flavours (BLASTp, BLASTn, tBLASTn and BLASTx) (19).

Individual protein annotation view

Searching for a protein allows to access a 'locus' page, designed to summarize automated and imported functional annotations, and provides comprehensive comparative data to facilitate the interpretation of annotations (Figure 1). It integrates annotations from multiple databases including UniProt (curated and automated protein annotations) (20), KEGG (annotation of pathways), COG (orthology), TCDB (transporters) (21), STRING (protein-protein interactions) (22) and InterPro (domains and signatures). The different tabs at the top of the page link to additional data such as the list of orthologs in other PVC genomes (Figure 1C), identified using OrthoFinder (23). Orthologs are listed in a table containing the locus tag, the gene name, the name of the organism, the product, the percentage of amino acid identity as compared to the reference locus and the UniProt annotation score. Orthologs that were reviewed on SwissProt are flagged to quickly identify orthologs with manually curated annotations. Additional tabs link to (i) a precomputed phylogeny of the orthologous group, (ii) a second phylogeny that includes the closest non-PVC RefSeq hits of each sequence of the orthogroup, allowing to investigate the phylogenetic relationship of PVC proteins and their closest homologs available in public databases (Figures 1J and 2J), precomputed homology searches with (iii) RefSeq and (iv) SwissProt databases (200 top hits), (v) links to published literature based on text-mining from the STRING database (24) and PaperBLAST hits (25) and (vi) candidate functional interactors. Putative interactors were predicted in-house from genomic data alone using phylogenetic profiling and investigation of conserved gene neighborhood (see online methods) (Figure 1G). See (26) and (27) for the rationale justifying use of those two approaches.

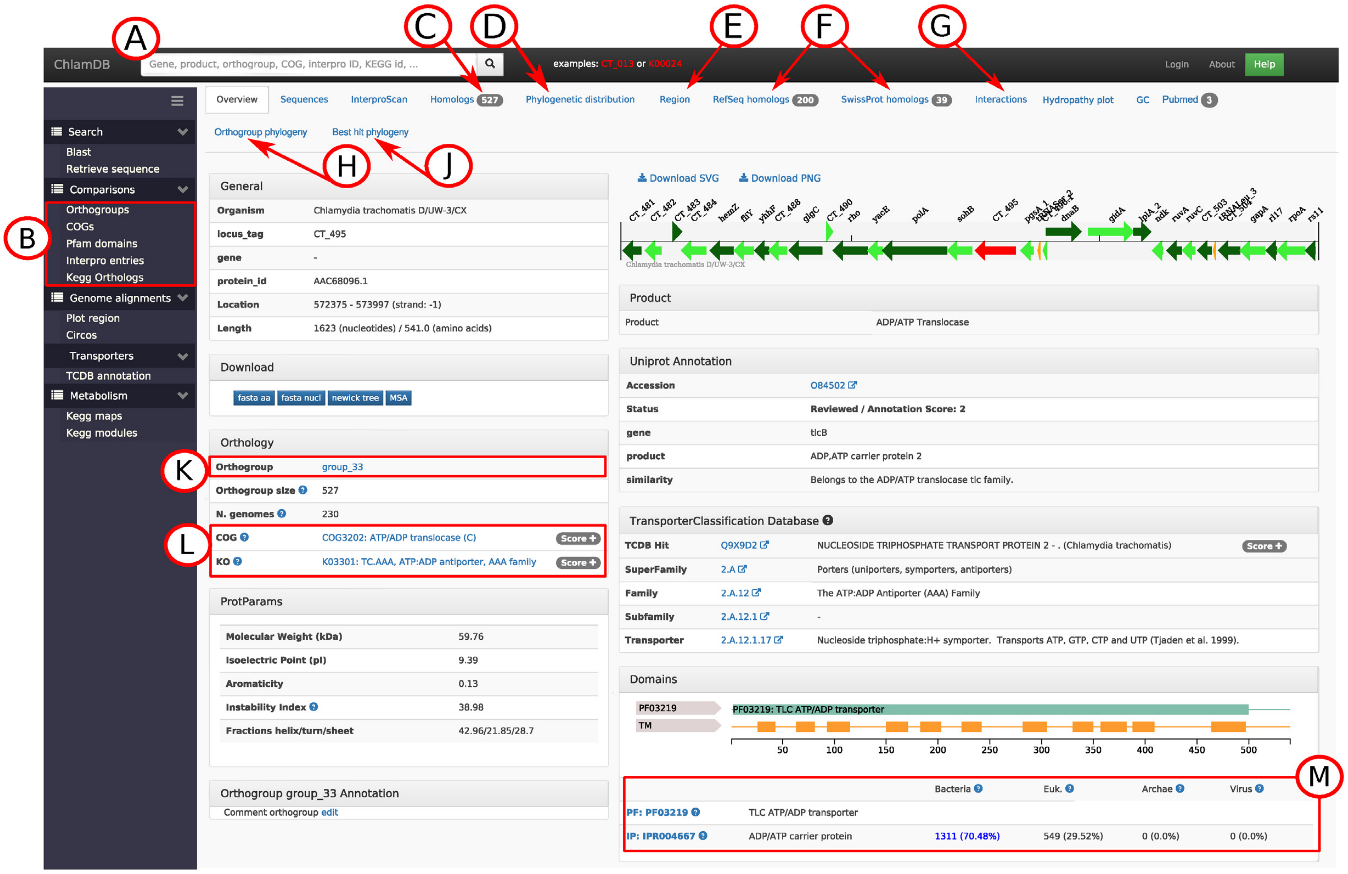
We put a strong emphasis on the visual representation of the data (Figure 2). The pattern of presence/absence of orthologous groups within the PVC superphylum can be visualized with help of an annotated reference phylogeny (Figures 1D and 2D). The reference phylogeny was reconstructed with FastTree (28) (default parameters, JTT+CAT model) based on the concatenated alignment of 32 single copy orthologs conserved in at least 266 out of the 277 genomes.

The organization of transmembrane and Pfam domains in orthologs can be easily compared along the phylogeny of the orthologous group (Figures 1H and 2H). The conservation of proteins encoded in the direct neighborhood (23 kb upstream and downstream) of the protein of interest can also be visualized (Figures 1E and 2E).

The 'orthogroup' link (Figure 1K) provides an overview of the annotation of orthologs including gene name, product, COG annotation, KEGG annotation, InterPro annotations, number of transmembrane domains and sequence length. It allows verifying the consistency of annotations among putative orthologs and identifying wrongly grouped proteins (e.g. non-orthologous proteins sharing a domain).

Annotation of candidate type III secretion system effectors

Chlamydiae use a type III secretion system (T3SS) to deliver effector proteins that will allow the bacterium to overcome eukaryotic host defenses and to manipulate host cells. Effectors are difficult to identify because they evolve quickly and are much less conserved than proteins encoding components of the T3SS apparatus (29,30). Between 5 and 8% of *Chlamydia* spp. coding sequences (CDS) are estimated to be effectors (31). Candidate T3SS effectors were identified using four different machine-learning classifiers that were trained with known effector sequences: BPBAac (32), effectiveT3 (33), DeepT3 (34) and T3_MM (35). In addition, we tagged proteins harboring eukaryotic domains rarely found in bacterial genomes. Such domains are known to



General

Organism	Chlamydia trachomatis D/UW-3/CX
locus_tag	CT_495
gene	-
protein_id	AAC68096.1
Location	572375 - 573997 (strand: -1)
Length	1623 (nucleotides) / 541.0 (amino acids)

Download

Orthology

Orthogroup	group_33
Orthogroup size	527
N. genomes	230
COG	COG3202: ATP/ADP translocase (C)
KO	K03301: TC.AAA, ATP:ADP antiporter, AAA family

ProtParams

Molecular Weight (kDa)	59.76
Isoelectric Point (pI)	9.39
Aromaticity	0.13
Instability Index	38.98
Fractions helix/turn/sheet	42.96/21.85/28.7

Orthogroup group_33 Annotation

Comment orthogroup edit

Product

Product	ADP/ATP Translocase

Uniprot Annotation

Accession	O84502
Status	Reviewed / Annotation Score: 2
gene	tlcB
product	ADP,ATP carrier protein 2
similarity	Belongs to the ADP/ATP translocase tlc family.

TransporterClassification Database

TCDB Hit	Q9X9D2	NUCLEOSIDE TRIPHOSPHATE TRANSPORT PROTEIN 2 - . (Chlamydia trachomatis)
SuperFamily	2.A	Porters (uniporters, symporters, antiporters)
Family	2.A.12	The ATP:ADP Antiporter (AAA) Family
Subfamily	2.A.12.1	-
Transporter	2.A.12.1.17	Nucleoside triphosphate:H+ symporter. Transports ATP, GTP, CTP and UTP (Tjaden et al. 1999).

Domains

		Bacteria	Euk.	Archae	Virus
PF: PF03219	TLC ATP/ADP transporter				
IP: IPR004667	ADP/ATP carrier protein	1311 (70.48%)	549 (29.52%)	0 (0.0%)	0 (0.0%)

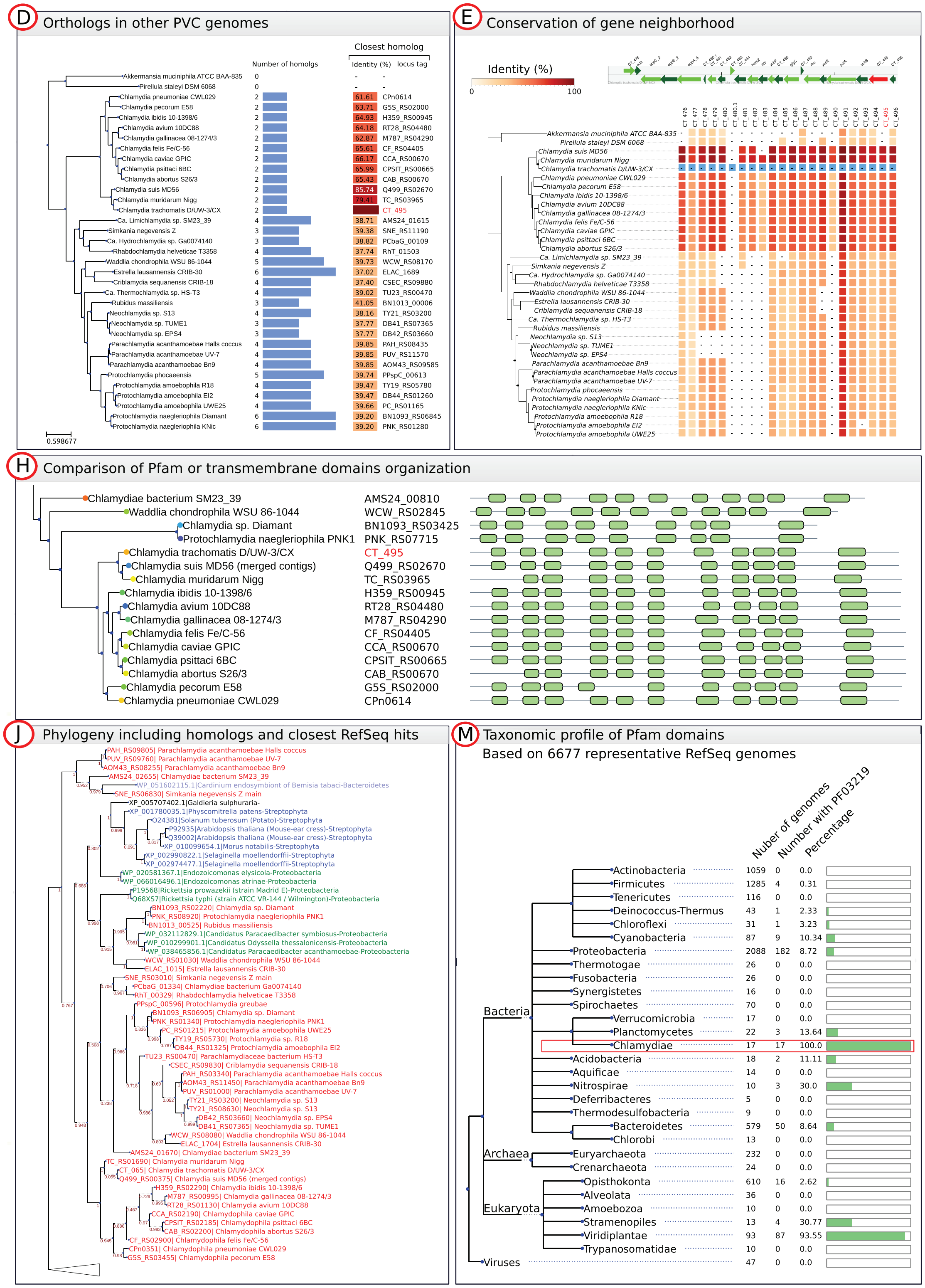
Figure 1. Protein annotation page of CT_495, an ADP/ATP transporter. (**A**) Main search bar. (**B**) Menu to access comparative analyses tools for the comparison of genome content based on the clustering of proteins into orthologous groups and for comparison of COG, KEGG, Pfam and InterPro annotations. (**C**) 'Homologs' tab with the list of the 527 orthologs of CT_495 in other PVC genomes. (**D**) Tab with the reference species phylogeny and the pattern of presence/absence of orthologs in each genome of the database as well as the locus tag of the closest ortholog in each genome. (**E**) 'Region' tab showing the conservation of proteins encoded in the direct neighborhood of the target protein. (**F**) Best hits in RefSeq and SwissProt databases. (**G**) Predicted protein interactors based on phylogenetic profiling and conservation of gene neighborhood. (**H**) Phylogenetic trees of the orthogroup and associated Pfam and transmembrane (TM) domain organization of each protein. (**J**) Phylogenetic tree including the best RefSeq hits of each protein of the orthologous group. (**K**) Name of the orthologous group with link to an overview of the annotation of the considered orthogroup (here 527 orthologs for group_33). (**L**) COG and KEGG annotations with link to the detailed list of proteins annotated with the same COG/KO in other genomes of the database; (**M**) Pfam and InterPro annotations with basic taxonomic information from the InterPro website: the numbers and percentages of proteins harboring this domain that are classified as Bacteria, Eukaryote, Archaea and Virus (data retrieved from InterPro version 60). Clicking on the Pfam accession numbers links to more detailed taxonomic information and a detailed list of proteins harboring the same domain in 6677 representative RefSeq genomes.

be frequently involved in bacteria–host interactions (36,37). The ADP/ATP transporter domain (InterPro accession IPR004667) is for instance frequently found in both bacteria (70.48%) and eukaryotes (29.52%) (Figure 1L). A dedicated page allows visualizing the taxonomic distribution of each COG and Pfam domains across respectively 2,031 (for COG) and 6,677 (for Pfam) representative Archaea, Bacteria, Eukaryotes and Viruses genomes (Figure 1M and 2M). The detailed list of identified homologs can (for instance) be used to quickly determine whether a candidate effector protein harbors a domain predominantly identified in the genome of eukaryotes and other intracellular bacterial parasites such as *Rickettsia* or *Legionella*.

Comparative genomics and data mining tools

Since *C. trachomatis* genome became one of the first sequenced genomes (38), hundreds of *Chlamydiae* genomes have been sequenced. Comparisons of complete genomes of different strains and species can help identify genetic variations that can be involved in defining tissue tropism or host specificity (39), or identify genes essential to the unique intracellular lifestyle of *Chlamydiae*. ChlamDB allows users to perform various comparative analyses based on orthologous proteins to identify highly conserved and genome-specific or clade-specific orthologous groups (Figure 3.1 and 3.2). Whole genome comparisons can be visualized using interactive circular genome maps, Venn diagrams or heat maps (Figure 3.3, 3.4 and 3.5). In addition, ChlamDB enables the alignment of local genomic regions in two or more genomes (Figure 3.6).

Pfam domains, KEGG orthologs and InterPro entries can also be compared to identify clade-specific or highly conserved protein features (Figure 3.7). A simple form enables the user to compare the size of gene families or the frequency of domains/KEGG annotations in each genome, allowing the identification of large protein families or fre-



	Nuber of genomes	Number with PF03219	Percentage
Actinobacteria	1059	0	0.0
Firmicutes	1285	4	0.31
Tenericutes	116	0	0.0
Deinococcus-Thermus	43	1	2.33
Chloroflexi	31	1	3.23
Cyanobacteria	87	9	10.34
Proteobacteria	2088	182	8.72
Thermotogae	26	0	0.0
Fusobacteria	26	0	0.0
Synergistetes	16	0	0.0
Spirochaetes	70	0	0.0
Verrucomicrobia	17	0	0.0
Planctomycetes	22	3	13.64
Chlamydiae	17	17	100.0
Acidobacteria	18	2	11.11
Aquificae	14	0	0.0
Nitrospirae	10	3	30.0
Deferribacteres	5	0	0.0
Thermodesulfobacteria	9	0	0.0
Bacteroidetes	579	50	8.64
Chlorobi	13	0	0.0
Euryarchaeota	232	0	0.0
Crenarchaeota	24	0	0.0
Opisthokonta	610	16	2.62
Alveolata	36	0	0.0
Amoebozoa	10	0	0.0
Stramenopiles	13	4	30.77
Viridiplantae	93	87	93.55
Trypanosomatidae	10	0	0.0
Viruses	47	0	0.0

Figure 2. Selected examples of comparative data that can be retrieved from protein annotation pages. Panels are named according to the links shown in Figure 1. (**D**) Profile of presence/absence of orthologs in other PVC genomes with the identity of the closest ortholog in each genome. (**E**) Visualization of the conservation of proteins encoded in the neighborhood of CT_495. (**H**) Comparative view of transmembrane domains organization of CT_495 (TlcB) orthologs. (**J**) Phylogeny of CT_495 orthologous group including the closest identified RefSeq homologs. Red labels are proteins from the ChlamDB database whereas blue and green labels indicate non-PVC proteins. In this example, a sequence of *Cardinium*, a *Bacteroidetes* endosymbiont of the whitefly *Bemisia tabaci*, is clustering with *Chlamydiae* spp. and Proteobacteria symbionts such as *Paracaedibacter* and *Rickettsia* spp. are clustering with other *Chlamydia* spp., suggesting multiple events of horizontal gene transfer. Panels present only a subset of the 277 genomes currently present in the database to fit on a single page. Complete figures can be retrieved from the ChlamDB website. (**M**) Overview of the taxonomic profile of the Pfam domain PF03219 in 6677 representative RefSeq genomes. We can observe that this domain can be found in 100% of *Chlamydiae* genomes and in most *Viridiplantae*, but also in some genomes of other bacterial phyla. The detailed list of hits can also be browsed and downloaded.

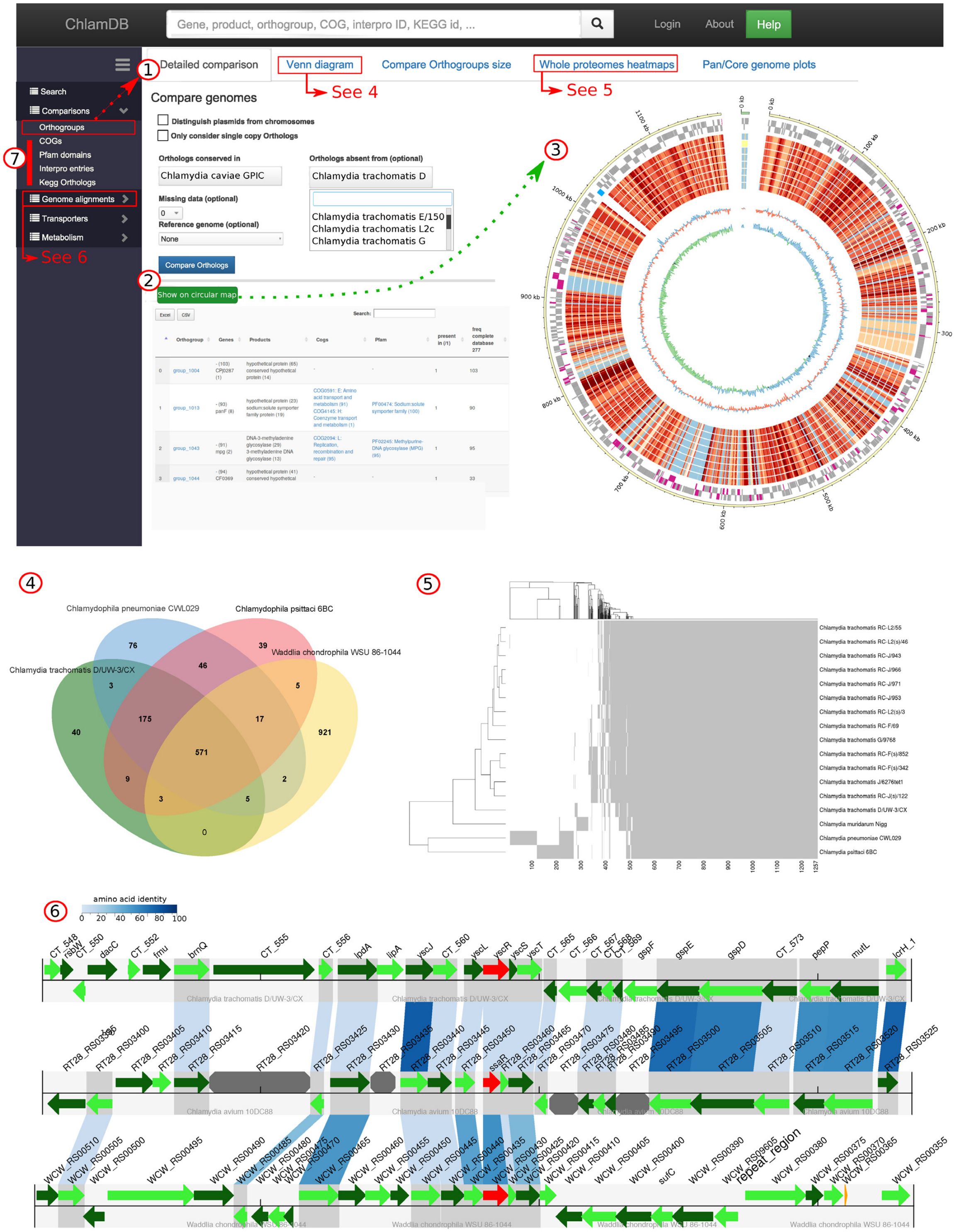
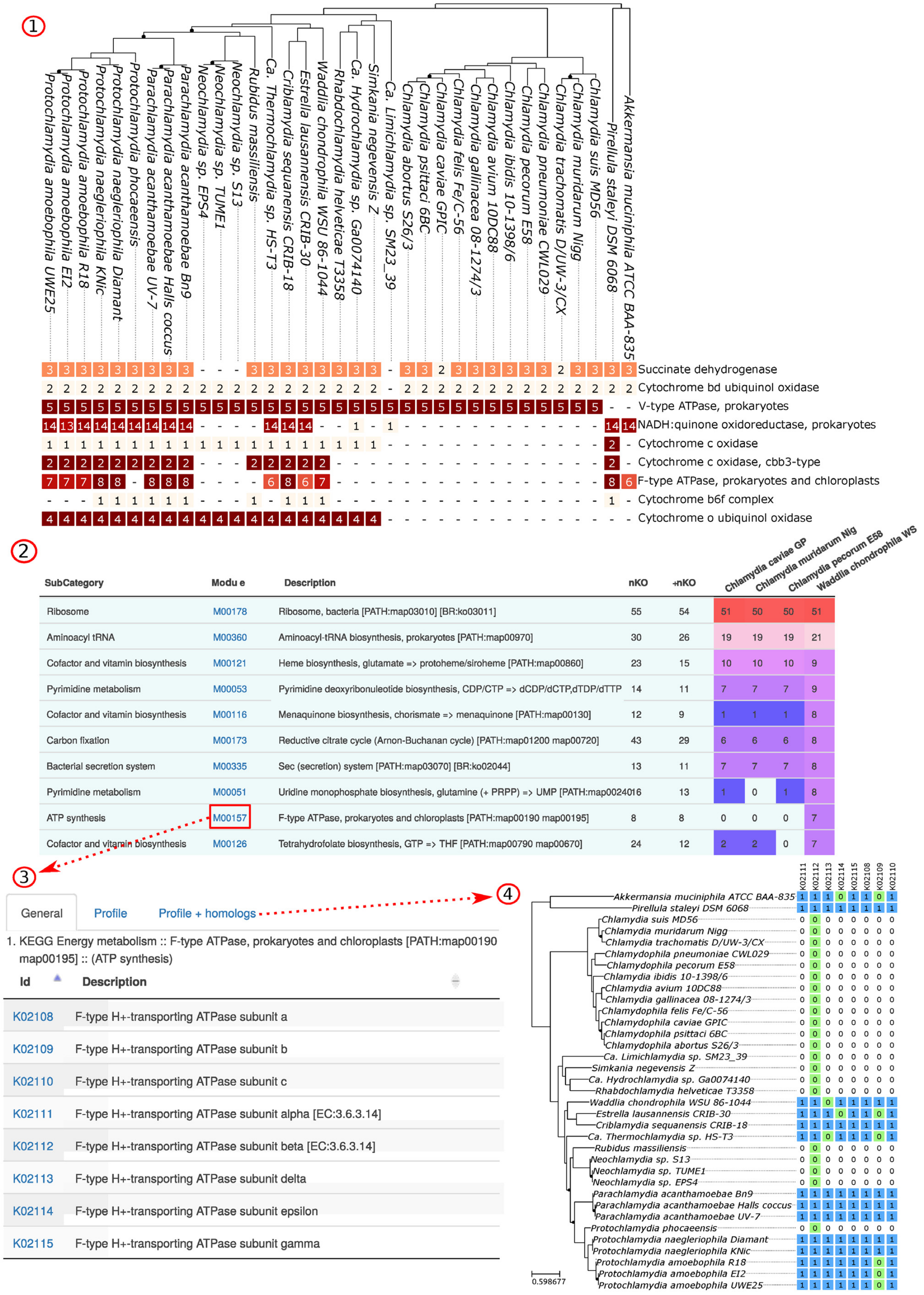

ChlamDB
Gene, product, orthogroup, COG, interpro ID, KEGG id, ...
Login
About
Help
Search
Comparisons
Orthogroups
COGs
Pfam domains
Interpro entries
Kegg Orthologs
Genome alignments
Transporters
Metabolism
See 6
1
Detailed comparison
Venn diagram
See 4
Compare Orthogroups size
Whole proteomes heatmaps
See 5
Pan/Core genome plots
Compare genomes
Distinguish plasmids from chromosomes
Only consider single copy Orthologs
Orthologs conserved in
Chlamydia caviae GPIC
Orthologs absent from (optional)
Chlamydia trachomatis D
Chlamydia trachomatis E/150
Chlamydia trachomatis L2c
Chlamydia trachomatis G
Missing data (optional)
Reference genome (optional)
None
Compare Orthologs
2
Show on circular map
3
7
4
Chlamydophila pneumoniae CWL029
Chlamydophila psittaci 6BC
Waddlia chondrophila WSU 86-1044
Chlamydia trachomatis D/UW-3/CX
76
39
46
5
3
175
17
40
921
571
9
2
3
5
0
5
Chlamydia trachomatis RC-L2/55
Chlamydia trachomatis RC-L2(s)/46
Chlamydia trachomatis RC-J/943
Chlamydia trachomatis RC-J/966
Chlamydia trachomatis RC-J/971
Chlamydia trachomatis RC-J/953
Chlamydia trachomatis RC-L2(s)/3
Chlamydia trachomatis RC-F/69
Chlamydia trachomatis G/9768
Chlamydia trachomatis RC-F(s)/852
Chlamydia trachomatis RC-F(s)/342
Chlamydia trachomatis J/6276tet1
Chlamydia trachomatis RC-J(s)/122
Chlamydia trachomatis D/UW-3/CX
Chlamydia muridarum Nigg
Chlamydia pneumoniae CWL029
Chlamydia psittaci 6BC
6
amino acid identity
0 20 40 60 80 100
Chlamydia trachomatis D/UW-3/CX
Chlamydia avium 10DC88
Waddlia chondrophila WSU 86-1044

Figure 3. Comparative analyses based on orthologous groups. (**1**) The orthogroup comparison allows retrieving (**2**) a list of orthologous groups that are conserved in a group of genomes and absent from another group of genomes, that can be visualized on a circular map of one reference genome (**3**). The outer gray circles indicate CDS encoded on the leading and lagging strand of the genome. Proteins encoded in the genome of *Chlamydia caviae* and absent from the genomes of eight *Chlamydia trachomatis* strains are highlighted in pink. The inner red/blue circle indicates the conservation of each protein in the genomes included in the comparison (the red scale reflects protein identity, blue indicate absence of identified homolog). Regions of interest can be clicked, redirecting the user to the corresponding protein annotation page. (**4**) Up to six genomes can also be compared using interactive Venn diagrams. The complete list of shared or specific orthologous groups can be retrieved by clicking on the numbers. (**5**) Conservation of orthologous groups can be visualized as heat maps of the presence (gray) and absence (white) of each orthologous group in selected genomes. (**6**) Local alignments between distantly related genomes can be done using one locus (in red) as anchor. Green features are open reading frames (ORFs). Octagonal gray features are pseudogenes. Blue boxes linking ORFs from different genomes reflect the conservation of orthologs (the blue color scale reflects protein sequences identity). (**7**) Similar comparative analyses as in panels 1–5 can be done based on COG, Pfam, InterPro and KEGG orthologs annotations.



SubCategory	Modu e	Description	nKO	+nKO	Chlamydia caviae GP	Chlamydia muridarum Nig	Chlamydia pecorum E58	Waddlia chondrophila WS
Ribosome	M00178	Ribosome, bacteria [PATH:map03010] [BR:ko03011]	55	54	51	50	50	51
Aminoacyl tRNA	M00360	Aminoacyl-tRNA biosynthesis, prokaryotes [PATH:map00970]	30	26	19	19	19	21
Cofactor and vitamin biosynthesis	M00121	Heme biosynthesis, glutamate => protoheme/siroheme [PATH:map00860]	23	15	10	10	10	9
Pyrimidine metabolism	M00053	Pyrimidine deoxyribonuleotide biosynthesis, CDP/CTP => dCDP/dCTP,dTDP/dTTP	14	11	7	7	7	9
Cofactor and vitamin biosynthesis	M00116	Menaquinone biosynthesis, chorismate => menaquinone [PATH:map00130]	12	9	1	1	1	8
Carbon fixation	M00173	Reductive citrate cycle (Arnon-Buchanan cycle) [PATH:map01200 map00720]	43	29	6	6	6	8
Bacterial secretion system	M00335	Sec (secretion) system [PATH:map03070] [BR:ko02044]	13	11	7	7	7	8
Pyrimidine metabolism	M00051	Uridine monophosphate biosynthesis, glutamine (+ PRPP) => UMP [PATH:map0024016	13		1	0	1	8
ATP synthesis	M00157	F-type ATPase, prokaryotes and chloroplasts [PATH:map00190 map00195]	8	8	0	0	0	7
Cofactor and vitamin biosynthesis	M00126	Tetrahydrofolate biosynthesis, GTP => THF [PATH:map00790 map00670]	24	12	2	2	0	7

1. KEGG Energy metabolism :: F-type ATPase, prokaryotes and chloroplasts [PATH:map00190 map00195] :: (ATP synthesis)

Id	Description
K02108	F-type H+-transporting ATPase subunit a
K02109	F-type H+-transporting ATPase subunit b
K02110	F-type H+-transporting ATPase subunit c
K02111	F-type H+-transporting ATPase subunit alpha [EC:3.6.3.14]
K02112	F-type H+-transporting ATPase subunit beta [EC:3.6.3.14]
K02113	F-type H+-transporting ATPase subunit delta
K02114	F-type H+-transporting ATPase subunit epsilon
K02115	F-type H+-transporting ATPase subunit gamma

Figure 4. Comparative analysis of KEGG Pathways and Modules. (**1**) Comparison of KEGG modules of the 'ATP synthesis' category. Numbers indicate the amount of annotated Kegg Orthologs (KO) for each module. We can observe that F-type ATP-synthase subunits were identified in only some species of the phylum and were probably lost independently by *Neochlamydia/rubidus* strains and *Protochlamydia phocaensis*. (**2**) Module data can be browsed as tables. (**3**) Detailed lists of Kegg Orthologs (KO) can be retrieved from linked pages. (**4**) The pattern of presence or absence of each KO in each genome of the phylum can also be investigated visually (blue cells) for any module and pathway. Green cells indicate genomes for which no protein was annotated with the corresponding KO but an ortholog was identified based on OrthoFinder data. They could indicate either wrongly annotated proteins or non-orthologous proteins wrongly clustered in the same orthologous group. Panels 1 and 4 present only a subset of the 277 genomes currently present in the database to fit on a single page. Complete figures can be retrieved from the ChlamDB website.

quent domains. For instance, the polymorphic membrane protein family (Pmp), a family of proteins involved in adhesion identified in all sequenced *Chlamydiaceae* genomes (40), is present in up to 28 copies in *C. psittaci* CP3 genome. Interestingly, the Pfam domain **PF05150** ('*Legionella pneumophila* major outer membrane protein domain'), a domain extremely rarely identified outside of the *Legionella* genus (see https://chlamdb.ch/pfam_profile/PF05150/phylum) is present in 219 copies within the PVC superphylum (https://chlamdb.ch/fam/PF05150/pfam). This domain is also the most frequent domain identified in the genome of *Simkania negevensis* (36 occurrences). Proteins harboring this domain were probably acquired by horizontal gene transfer by *Chlamydiae*, *Legionella* or both and might share similar functions.

Annotations from the KEGG database were used to classify proteins into metabolic pathways and modules (16). Data for individual pathways and modules can be retrieved by searching KEGG accessions in the main search bar. In addition, KEGG annotations in various genomes can be compared as annotated phylogenies (Figure 4.1) and interactive bar charts or accessed from summary tables available for each genome (Figure 4.2). Modules and pathways pages detail KEGG orthologs associated to a given entry (Figure 4.3) and report the list of orthologs identified in each PVC genome (Figure 4.4).

Implementation, methods and updates

The interface was developed using the Django framework (https://www.djangoproject.com/). Data are stored on a MySQL server and visualized with existing JavaScript libraries allowing to draw interactive plots and tables such as jvenn.js (41), datatables.js (https://datatables.net), cytoscape.js (42) and feature−viewer.js (https://github.com/calipho-sib/feature-viewer) (43). The python module GenomeDiagram is used to draw genome schematics, including alignments of multiple genomic locations (44). Circular representations of genomes and plasmids are made with Circos (45). The Ete3 Python module is used to draw phylogenetic trees with associated metadata (46). Some plots are also made using R (47), ggplot2 (48) and plotly (https://plot.ly). Annotations, phylogenetic trees and multiple sequence alignments can be downloaded from the website. A detailed description of the methods used to precompute functional and comparative analyses and setup the database is available online (https://www.chlamdb.ch/docs/index.html). The code source of the website is freely available on Github and issues can be reported online (https://github.com/metagenlab/chlamdb). This database has been developed at the Centre for Research on Intracellular Bacteria (CRIB) in Lausanne and will be maintained and updated at least once a year.

CONCLUSION AND FUTURE DIRECTIONS

As the number of genome sequences quickly increases, there is a need for a centralized genomics resource providing updated annotations and extensive comparative genomics capabilities for the PVC superphylum. A superphylum-specific database has a significant added value with respect to large-scale genomic databases such as PATRIC (49) or Microscope (50): ChlamDB greatly facilitates access to comprehensive annotations and comparative data meaningful to the *Chlamydia* and *PVC* research community, with an intuitive interface and a special focus on visual representations of comparative data. Easy access to precomputed homology searches and phylogenetic reconstructions will help researchers to investigate the function and evolutionary history of proteins encoded in *PVC* genomes. Annotations of proteins specific for intracellular life such as predictions of type III secretion system effectors and identification of eukaryote-like domains will also facilitate the identification of uncharacterized proteins that might be involved in chlamydia-host interactions.

Since the annotation of PVC genomes stored in Genbank is generally not up-to-date with the most recent research, the existing ChlamDB could be extended to allow manual curation of the annotation and tracking of protein annotation history. Indeed, successful examples of community-curated databases exist for major pathogens, such as the Pseudomonas Database (www.pseudomonas.com) (51). The inference of orthologous relationships could be used to propagate the annotation of characterized proteins to less studied members of the phylum.

ACKNOWLEDGEMENTS

We are grateful to Roland Sahli for his support and for providing the computational resources necessary for the development and setup of this database. We would like to thank Valentin Scherz for valuable discussions as well as Carole Kebbi-Beghdadi, Ludovic Pilloux, Silvia Ardissone, Sébastien Aeby, Marie de Barsy, Nicolas Jacquier, Firuza Bayramova, Aurélie Scherler, Alyce Taylor-Brown and all other members of the Center for Research on Intracellular Bacteria (CRIB) for their support and feedbacks during the development of ChlamDB.

FUNDING

This work results from the close interaction between the Laboratory of Genomics and Metagenomics and the Center for Research on Intracellular Bacteria, both led by Prof. Greub. Funding of research in genomics and on intracellular bacteria by Greub's research team is supported by various grants, including grants from the Swiss National Science Foundation (Sinergia grant n° CRSII3-141837; SNSF n°3200BO-116445; SNSF n°310030-162603; SNSF n°10531C-170280), as well as by grants from Foundations (Jürg Tschopp award and Leenaards award).

Conflict of interest statement. None declared.

REFERENCES

1. Bachmann,N.L., Polkinghorne,A. and Timms,P. (2014) Chlamydia genomics: providing novel insights into chlamydial biology. *Trends Microbiol.*, **22**, 464–472.
2. Leonard,C.A. and Borel,N. (2014) Chronic chlamydial diseases: from atherosclerosis to urogenital infections. *Curr. Clin. Microbiol. Rep.*, **1**, 61–72.
3. Rivas-Marín,E. and Devos,D.P. (2018) The Paradigms They Are a-Changin': past, present and future of PVC bacteria research. *Antonie Van Leeuwenhoek*, **111**, 785–799.

4. Jacquier,N., Viollier,P.H. and Greub,G. (2015) The role of peptidoglycan in chlamydial cell division: towards resolving the chlamydial anomaly. *FEMS Microbiol. Rev.*, **39**, 262–275.
5. Jacquier,N., Frandi,A., Pillonel,T., Viollier,P. and Greub,G. (2014) Cell wall precursors are required to organize the chlamydial division septum. *Nat. Commun.*, **5**, 3578 .
6. Jeske,O., Schüler,M., Schumann,P., Schneider,A., Boedeker,C., Jogler,M., Bollschweiler,D., Rohde,M., Mayer,C., Engelhardt,H. *et al.* (2015) Planctomycetes do possess a peptidoglycan cell wall. *Nat. Commun.*, **6**, 1–7.
7. van Teeseling,M.C.F., Mesman,R.J., Kuru,E., Espaillat,A., Cava,F., Brun,Y.V., VanNieuwenhze,M.S., Kartal,B. and van Niftrik,L. (2015) Anammox Planctomycetes have a peptidoglycan cell wall. *Nat. Commun.*, **6**, 1–6.
8. Putman,T., Hybiske,K., Jow,D., Afrasiabi,C., Lelong,S., Cano,M.A., Stupp,G.S., Waagmeester,A., Good,B.M., Wu,C. *et al.* (2019) ChlamBase: a curated model organism database for the Chlamydia research community. *Database*, **2019**, 1–9.
9. Greub,G. and Raoult,D. (2003) History of the ADP/ATP-translocase-encoding gene, a parasitism gene transferred from a Chlamydiales ancestor to plants 1 billion years ago. *Appl. Environ. Microbiol.*, **69**, 5530–5535.
10. Bordin,N., González-Sánchez,J.C. and Devos,D.P. (2018) PVCbase: an integrated web resource for the PVC bacterial proteomes. *Database*, **2018**,1–10.
11. Sayers,E.W., Cavanaugh,M., Clark,K., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2019) GenBank. *Nucleic Acids Res.*, **47**, D94–D99.
12. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R. *et al.* (2018) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.
13. Derrien,M., Collado,M.C., Ben-Amor,K., Salminen,S. and de Vos,W.M. (2008) The Mucin degrader Akkermansia muciniphila is an abundant resident of the human intestinal tract. *Appl. Environ. Microbiol.*, **74**, 1646–1648.
14. Galperin,M.Y., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
15. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
16. Kanehisa,M. and Sato,Y (2019) KEGG Mapper for inferring cellular functions from protein sequences. *Protein Sci.*, 1–8.
17. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
18. Mitchell,A.L., Attwood,T.K., Babbitt,P.C., Blum,M., Bork,P., Bridge,A., Brown,S.D., Chang,H.-Y., El-Gebali,S., Fraser,M.I. *et al.* (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.*, **47**, D351–D360.
19. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 1–9.
20. UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
21. Saier,M.H., Reddy,V.S., Tsu,B.V., Ahmed,M.S., Li,C. and Moreno-Hagelsieb,G. (2016) The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Res.*, **44**, D372–D379.
22. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
23. Emms,D.M. and Kelly,S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, **16**, 1–14.
24. Szklarczyk,D., Morris,J.H., Cook,H., Kuhn,M., Wyder,S., Simonovic,M., Santos,A., Doncheva,N.T., Roth,A., Bork,P. *et al.* (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.*, **45**, D362–D368.
25. Price,M.N. and Arkin,A.P. (2017) PaperBLAST: Text Mining Papers for Information about Homologs . *mSystems*, **2**, e00039-17.
26. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
27. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 4285–4288.
28. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
29. Nogueira,T., Touchon,M. and Rocha,E.P.C. (2012) Rapid evolution of the sequences and gene repertoires of secreted proteins in bacteria. *PLoS One*, **7**, e49403.
30. Dehoux,P., Flores,R., Dauga,C., Zhong,G. and Subtil,A. (2011) Multi-genome identification and characterization of chlamydiae-specific type III secretion substrates: the Inc proteins. *BMC Genomics*, **12**, 1–20.
31. Valdivia,R.H. (2008) Chlamydia effector proteins and new insights into chlamydial cellular microbiology. *Curr. Opin. Microbiol.*, **11**, 53–59.
32. Wang,Y., Zhang,Q., Sun,M. and Guo,D. (2011) High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics*, **27**, 777–784.
33. Jehl,M.-A., Arnold,R. and Rattei,T. (2011) Effective—a database of predicted secreted bacterial proteins. *Nucleic Acids Res.*, **39**, D591–D595.
34. Xue,L., Tang,B., Chen,W. and Luo,J. (2019) DeepT3: deep convolutional neural networks accurately identify Gram-negative bacterial type III secreted effectors using the N-terminal sequence. *Bioinformatics*, **35**, 2051–2057.
35. Wang,Y., Sun,M., Bao,H. and White,A.P. (2013) T3_MM: a Markov model effectively classifies bacterial type III secretion signals. *PLoS One*, **8**, e58173.
36. Gimenez,G., Bertelli,C., Moliner,C., Robert,C., Raoult,D., Fournier,P.-E. and Greub,G. (2011) Insight into cross-talk between intra-amoebal pathogens. *BMC Genomics*, **12**, 1–14.
37. Ponting,C.P., Aravind,L., Schultz,J., Bork,P. and Koonin,E.V. (1999) Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J. Mol. Biol.*, **289**, 729–745.
38. Stephens,R.S., Kalman,S., Lammel,C., Fan,J., Marathe,R., Aravind,L., Mitchell,W., Olinger,L., Tatusov,R.L., Zhao,Q. *et al.* (1998) Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. *Science*, **282**, 754–759.
39. Read,T.D., Brunham,R.C., Shen,C., Gill,S.R., Heidelberg,J.F., White,O., Hickey,E.K., Peterson,J., Utterback,T., Berry,K. *et al.* (2000) Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. *Nucleic Acids Res.*, **28**, 1397–1406.
40. Becker,E. and Hegemann,J.H. (2014) All subtypes of the Pmp adhesin family are implicated in chlamydial virulence and show species-specific function. *Microbiologyopen*, **3**, 544–556.
41. Bardou,P., Mariette,J., Escudié,F., Djemiel,C. and Klopp,C. (2014) jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, **15**, 1–7.
42. Franz,M., Lopes,C.T., Huck,G., Dong,Y., Sumer,O. and Bader,G.D. (2016) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, **32**, 309–311.
43. Gaudet,P., Michel,P.-A., Zahn-Zabal,M., Britan,A., Cusin,I., Domagalski,M., Duek,P.D., Gateau,A., Gleizes,A., Hinard,V. *et al.* (2017) The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.*, **45**, D177–D182.
44. Pritchard,L., White,J.A., Birch,P.R.J. and Toth,I.K. (2006) GenomeDiagram: a python package for the visualization of large-scale genomic data. *Bioinformatics*, **22**, 616–617.
45. Krzywinski,M., Schein,J., Birol,İ., Connors,J., Gascoyne,R., Horsman,D., Jones,S.J. and Marra,M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
46. Huerta-Cepas,J., Serra,F. and Bork,P. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, **33**, 1635–1638.

47. R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.
48. Wickham,H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer, NY.
49. Wattam,A.R., Davis,J.J., Assaf,R., Boisvert,S., Brettin,T., Bun,C., Conrad,N., Dietrich,E.M., Disz,T., Gabbard,J.L. *et al.* (2017) Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.*, **45**, D535–D542.
50. Vallenet,D., Calteau,A., Cruveiller,S., Gachet,M., Lajus,A., Josso,A., Mercier,J., Renaux,A., Rollin,J., Rouy,Z. *et al.* (2017) MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. *Nucleic Acids Res.*, **45**, D517–D528.
51. Winsor,G.L., Griffiths,E.J., Lo,R., Dhillon,B.K., Shay,J.A. and Brinkman,F.S.L. (2016) Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. *Nucleic Acids Res.*, **44**, D646–D653.