Title: Taxogenomics of the order Chlamydiales.

Authors: Pillonel T, Bertelli C, Salamin N, Greub G

Journal: International journal of systematic and evolutionary microbiology

Year: 2015 Apr

Issue: 65

Volume: Pt 4

Pages: 1381-93

DOI: 10.1099/ijs.0.000090

Taxogenomics of the *Chlamydiales*

Trestan Pillonel [1,2], Claire Bertelli [1,2], Nicolas Salamin [2,3] and Gilbert Greub [1,*]

[1] Center for Research on Intracellular Bacteria, Institute of Microbiology, University Hospital Center and University of Lausanne, Lausanne, Switzerland.

[2] SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3] Department of Ecology and Evolution, Biophore, University of Lausanne, Lausanne, Switzerland.

* Corresponding author:

Pr. Gilbert GREUB, Institute of Microbiology, Bugnon 48, CH-1011 Lausanne, Swizerland, Tel : +41213144979, fax : +41213144060, email : gilbert.greub@chuv.ch

Running title : Taxogenomics of the *Chlamydiales*

Keywords : Taxonomy, genomics, core genes, phylogeny, chlamydiae, intracellular bacteria

Word count : 5483

Abstract word count: 202

ABSTRACT

Bacterial classification is a long-standing problem for taxonomists and species definition itself is constantly debated among specialists. The classification of strict intracellular bacteria such as members of the *Chlamydiales* order mainly relies on DNA or protein-based phylogenetic reconstructions because they exhibit few phenotypic differences and are difficult to culture. The availability of full genome sequences allows the comparison of the performance of conserved protein sequences to reconstruct *Chlamydiales* phylogeny. This approach permits the identification of markers that maximize the phylogenetic signal and the robustness of the inferred tree. In this study, a set of 424 core proteins was identified and concatenated to construct a reference species tree. Although individual protein trees present variable topologies, we detected only few cases of incongruence with the reference species tree, which were due to horizontal gene transfers. Detailed analysis of the phylogenetic information of individual protein sequences (i) showed that phylogenies based on single randomly chosen core proteins are not reliable and (ii) led to the identification of twenty taxonomically highly reliable proteins, allowing the construction of a robust tree close to the reference species tree. We recommend to use these protein sequences to precisely classify newly discovered isolates at the family, genus and species levels.

1. INTRODUCTION

Phylogenetic reconstruction based on 16S ribosomal RNA (rRNA) sequences is a widely used approach to infer relationships between bacteria (Fox *et al.*, 1980). Nevertheless, the high conservation of rRNA reduces its discrimination power and makes it insufficient to distinguish closely-related bacterial species (Rosselló-Mora & Amann, 2001). In addition, performance of a single gene in phylogenetic inference can be highly variable for distantly-related species (Aguileta *et al.*, 2008). Indeed, highly conserved sequences with few substitutions are not evolutionary informative whereas sequences evolving very rapidly may have a saturated phylogenetic signal (Goldman, 1998). Horizontal gene transfer (HGT) or recombination events further complicate the reconstruction of species tree because of frequent discrepancies between gene trees. For example, serovars of *Chlamydia trachomatis* were classified based on the major outer membrane protein (*ompA*), but this classification was misleading because of recombination events in *ompA* (Brunelle & Sensabaugh, 2006; Harris *et al.*, 2012).

The *Chlamydiae* phylum was long restricted to one group of closely-related obligate intracellular bacteria classified in a single family, the *Chlamydiaceae*. During the last two decades, new organisms resembling *Chlamydiaceae* were identified in various hosts, such as amoebae, fish and arthropods (Horn, 2008). These so-called "*Chlamydia*-related" bacteria exhibit the same biphasic developmental cycle as *Chlamydiaceae* and all belong to the *Chlamydiales* order. These novel chlamydiae were isolated from different geographical areas, indicating a widespread occurrence in nature. This is also emphasized by the diversity of *Chlamydiales* organisms observed in metagenomics samples (Lagkouvardos *et al.*, 2013).

In 1999, Everett *et al.* proposed to use 16S and 23S rRNA cutoffs of 97, 95 and 90 percent identity to classify members of the *Chlamydiales* order at species, genus and family level, respectively (Everett *et al.*, 1999). Controversies arose because Everett *et al.* proposed to split the *Chlamydiaceae* family into two genera: *Chlamydia* and *Chlamydophila* (Everett *et al.*, 1999). This split was disputed since it was not consistently supported by significant biological differences and 16S rRNA differences were limited (Schachter *et al.*, 2001; Stephens *et al.*, 2009). Thus, the International subcommittee on the taxonomy of the *Chlamydiae* (ISTC) decided to revert to a single genus: *Chlamydia* (Bavoil *et al.*, 2013; Greub, 2010a). However, the rRNA identity cutoffs were accepted by the ISTC but should be used with caution and

flexibility (Greub, 2010b). The ISTC recommends using additional housekeeping genes (Greub, 2013).
Several attempts were made to develop a multilocus approach for the classification of chlamydial species (Klint *et al.*, 2007; Pannekoek *et al.*, 2008). Yet, they concentrated on the *Chlamydiaceae* family and did not consider the maximization of the phylogenetic signal allowing a robust evaluation of the deeper nodes of the *Chlamydiales* phylogeny. The use of multiple and carefully selected loci could both improve the resolution of the current classification and ease the assignment of newly identified species.

Thus, the present work aimed at identifying highly informative protein sequences for phylogenetic inference to allow the reconstruction of robust phylogenetic trees using a limited number of protein sequences. To achieve this goal, we compared currently available genomes from 15 different species belonging to five different families within the *Chlamydiales* order. We first determined the core genes conserved among all 15 species. To exclude potentially horizontally transferred genes, we then tested if the core genes present a congruent phylogenetic signal. Finally, the performance of individual protein sequence to reconstruct the species phylogeny was investigated in order to select sequences that accurately predict the relatedness of chlamydial isolates.

2. METHODS

2.1 *Chlamydiales* genomes

Twenty-one chlamydial genomes, including 15 species from five different families were included in the analysis (Table S1). Predicted protein sequences were retrieved from the NCBI (http://www.ncbi.nlm.nih.gov). Protein sequences from the draft genome of *Protochlamydia naegleriophila* KNic was obtained from the Center for Research on Intracellular Bacteria (CRIB, Lausanne).

2.2 Definition of the core gene set

Orthologs were searched with a reciprocal best BLAST hit (BBH) procedure. It assumes that orthologous sequences are more similar to each other than they are to other sequences. Pairwise BLASTP [version 2.2.24](Altschul *et al.*, 1997) searches were performed between every sequences from all genomes using the BLOSUM62 matrix, 0.1 e-value cut-off and no filter for low complexity regions. When BLASTP resulted in multiple high-scoring segment pairs (HSP), the average identity of the alignment was calculated by weighting the identity of each HSP by its length. Only proteins exhibiting BBH between all pairs of genomes were included in the core gene set.

2.3 Phylogenetic reconstructions

Different genome-scale methods have been developed to construct phylogenetic trees based on features such as gene content or gene order (Snel *et al.*, 2005). However, *Chlamydiales* species exhibit variations in gene content of multiple folds, and there is only poor gene order conservation between different *Chlamydia*-related species (Bertelli *et al.*, 2010; Collingro *et al.*, 2011). Therefore, a reference tree was built based on the sequences of core proteins using three alternative methods: average amino-acid identity, consensus and concatenation of core genes.

Core proteins were aligned using MAFFT 6.850 (Katoh *et al.*, 2002) with default parameters. The quality of the alignment was assessed using GUIDANCE residue pair scores (Penn *et al.*, 2010). The reconstruction of individual core genes was performed with PhyML version 3.0 (Guindon & Gascuel, 2003). According to ProtTest 3 results (Darriba *et al.*, 2011), the LG+Γ+I model of protein evolution was the best suited for 365/424 (86%) proteins (see supplementary table S2). Thus, all analyses were performed using a single model of amino acid replacement, which may have influenced the phylogenetic reconstitution of part of the dataset. A consensus tree derived from the individual core gene trees was constructed using the Extended Majority Rule criterion from the program SumTrees version 3.3.1 from DendropPy library version 3.12.0 (Sukumaran & Holder, 2010).

The reconstruction of a reference species tree was based on the concatenation of the aligned core proteins. Bootstrapped replicates of the concatenated alignment were generated using the SEQBOOT program of the PHYLIP package (J. Felsenstein, University of Washington, Seattle, USA). The trees were reconstructed using PhyML with the LG+Γ+I model. The consensus tree of 100 bootstrap replicates was constructed using SumTrees (Sukumaran & Holder, 2010). Neighbor joining trees were constructed using the bioNJ algorithm with Seaview (Gouy *et al.*, 2010).

2.4 Congruence and strength of the phylogenetic signal

Tree topologies were first compared using the Robinson Fould distance (Robinson & Foulds, 1981) computed using the package Phangorn (Schliep, 2011) in R (R Core Team, 2014). In addition, likelihood-based topological tests were performed to assess the congruence between each individual gene tree , i.e. assess whether individual genes phylogenies agree with one another, using the Shimodaira-Hasegawa test [SH-test] (Shimodaira & Hasegawa, 1999). For a given alignment, this test determines whether the likelihood of a suboptimal tree topology is significantly lower than the likelihood of the most likely tree. The likelihood of each candidate topology was calculated using LG+Γ+I model of substitution. For each core protein alignment, SH-tests were performed with all tree topologies obtained from other core proteins as well as the reference tree topology.

In order to evaluate the strength of the phylogenetic signal of each protein, SH-tests were performed to compare the likelihood of the most likely tree with the likelihood of random and semi-random topologies. Randomizing the topology of subparts of the species tree allowed evaluating the strength of the phylogenetic signal in the different subparts of the tree. Three kinds of semi-random topologies were tested: (i) 100 topologies randomizing the branching between *Chlamydia*-related species only (i.e. all members of the *Chlamydiales* order not belonging to the *Chlamydiaceae* family). (ii) 100 topologies randomizing only the branching between members of the *Chlamydiaceae* family, and (iii) all 15 branching possibilities of the five *Chlamydiales* families. The ability to reject semi-random topologies was evaluated by calculating the mean and standard deviation for the p-values of the three sets of semi-random topologies.

The similarity with the reference tree topology (Robinson-Fould distance), the congruence with this reference topology (SH-test p-value) and the ability to reject semi-random topologies (average and standard deviation of the SH-test p-values) were used to classify the chlamydial core proteins. The classification was done using the VEV clustering model (ellipsoidal, equal shape) implemented in the Mclust package (Fraley & Raftery, 2006). These clusters were used to define a minimal number of core genes to be used to resolve the phylogenetic relationships between members of the *Chlamydiales* order.

2.5 Classification of new chlamydial isolates

Five recently sequenced genomes (Table S1) were classified using the new classification procedure developed in this study. The orthologues of 9 proteins were identified in newly sequenced genomes by retrieving the best BLASTP hits of each 9 proteins from the 21 strains included in this analysis. For each 9 proteins, we confirmed that the best hit of a given protein was 21 times the same hit.

2.6 Pairwise distances

Pairwise identities were calculated based on Needleman-Wunsch global alignments computed using Needle (EMBOSS:6.5.7.0) (Rice *et al.*, 2000). Gaps were not considered in the calculation. Full length ribosomal sequences were extracted using barrnap 0.3 : Bacterial/Archaeal Ribosomal RNA Predictor (Seemann T, 2013; http://www.vicbioinformatics.com/). The average nucleotide identity (ANI) between chlamydial genomes was computed using MUMer (Kurtz *et al.*, 2004), as described by Richter and Rosselló-Móra (Richter & Rosselló-Móra, 2009).

3. RESULTS

3.1 Current criteria do not match the existing *Chlamydiales* classification

16S and 23S rRNA sequences are routinely used for bacterial species identification and classification. For members of the *Chlamydiales* order, cutoffs of 97, 95 and 90 percent identity are generally used to delineate species, genus and family levels (Domman *et al.*, 2014; Everett *et al.*, 1999; Lienard *et al.*, 2011). Nevertheless, the recognized classification frequently does not match these criteria, which are notably not well suited for closely-related strains (Fig. 1). In addition, 23S sequences are generally less conserved than 16S rRNA sequences, which makes the use of identical threshold values for two different genes inadequate. Moreover, rRNA identity does not necessarily reflect whole genome similarity. For example, *Chlamydia abortus* and *Chlamydia caviae* strains share 99.29% 16S rRNA identity (Table S3), 98.09% 23S rRNA identity (Table S4), while their whole genomes exhibit an average nucleotide identity (ANI) of 83.89% (Table S5). Contrary to rRNA, ANI cutoff of 95% reflects the recognized chlamydial species-level classification (Table S5). However, ANI calculation is not possible between distantly-related chlamydial genomes, because genomes cannot be aligned. Protein encoding regions are more appropriate to explore deeper phylogenetic relatedness. Chlamydial strains exhibit important variations in gene content, as *Chlamydia*-related strains present genomes between two and three folds larger than strains from the *Chlamydiaceae* family (Bertelli *et al.*, 2010; Collingro *et al.*, 2011). Nevertheless, members of the *Chlamydiaceae* family, most of which possess less than 1,000 genes, still have a large proportion (57-75%) of their proteome in common with *Chlamydia*-related species (Table S6). *Chlamydia trachomatis* strains share between 94% and 99 % of their predicted proteins. On the other hand, the two strains of *Parachlamydia acanthamoebae* and *Waddlia chondrophila* species share only between 86% and 90% of their predicted proteins. Among the *Chlamydia*-related families, only the genus *Protochlamydia* includes more than one species: *Candidatus* Protochlamydia amoebophila shares 71% of its proteins with the proteome of *Protochlamydia naegleriophila*. Their classification as a single genus is supported by the fact that orthologous proteins exhibit an average identity of 70% (Table S7), a percentage comparable to that observed between species of the *Chlamydia* genus.

The current classification of a given strain at species or family level can hardly be directly linked to the average amino-acid identity of orthologous proteins (Table S7). The *Chlamydiaceae* family and *Chlamydia*-related families are clearly separated, presenting between 44.39 and 45.93 average percent identity. Interestingly, *Simkania negevensis* Z presents a similarly low average amino-acid identity with all other strains (45.68% on average), whereas strains from other *Chlamydia*-related families present average identities higher than 50% between each other (Table S7, Fig. S3). In addition, there are no clear differences between the average identity of species of different genera and species of different families among the *Chlamydia*-related families. Indeed, *Estrella lausannensis* and *Criblamydia sequanensis* (same family) exhibit 52.7% average identity, whereas *W. chondrophila* and *P. acanthamoebae* (different families) exhibit 52.8% average identity. Species from the *Chlamydia* genus exhibit average identities ranging from 62.2% (*C. trachomatis* A-*C. pecorum*) to 94.4% (*C. abortus*-*C. psittaci*). Because of the limited usefulness of average nucleotide and amino-acid identity values, we focused on the identification of an informative restricted set of protein sequences to investigate the relationships between chlamydial strains.

3.2 Core genome and *Chlamydiales* phylogeny

While using a restrictive definition of orthologous proteins as those exhibiting a reciprocal BBH between all 21 genomes, we found a core genome of 424 protein coding genes. The corresponding 424 phylogenetic trees presented 386 different topologies. To reconstruct the *Chlamydiales* species tree, we used three methods: the average amino-acid identity, the consensus of all individual gene trees as well as the Maximum likelihood based on a concatenate of the 424 core genes. All these trees present highly similar topologies (Fig. 2) and reflect the classification recognized by the International Subcommittee for chlamydial taxonomy (Greub, 2010a, b). The former *Chlamydophila* subgroup clearly clusters separately from *C. trachomatis* and *C. muridarum*. Significant variations only occur between members of the former *Chlamydophila* subgroup. These variations involve the closely-related *C. psittaci*, *C. caviae* and *C. abortus* species and the basal branching of *C. pecorum* in the NJ tree based on average protein identities.

The topology of the gene trees frequently varies within the *Chlamydiaceae* family (Fig. 2b). In addition, frequent variations are observed concerning the relationship of the *Parachlamydia* and *Protochlamydia* genera, as well as between the *Waddliaceae* and the *Parachlamydiaceae* families, with two nodes presenting a frequency lower than 50% (Fig. 2b). Similarly, the concatenated tree presents a reduced support for the node connecting the *Parachlamydiaceae* and *Waddliaceae* families (Fig. 2c). The concatenated ML tree was used as a reference tree for all subsequent analyses.

3.3 Individual gene trees differ from the species tree

Each individual gene tree was compared to the reference tree topology (Fig. 2c). Only 7 topologies out of 424 were identical to the reference (without considering *C. trachomatis* strains branching pattern; Fig. 3a). Nevertheless, only 8 individual protein alignments rejected the reference tree topology with an SH-test significance threshold set at 0.2 (Fig. 3b, Table 1). Fig. 3(c) shows one example of strong conflicting phylogenetic signal due to an HGT event. Species of the *Protochlamydia* genus present sequences non-vertically inherited, suggesting the acquisition of a gene by an ancestor of the clade, followed by the loss of the gene copy of chlamydial descent. Other cases rejecting the reference tree generally presented more complex situations where different *Chlamydia*-related species clustered together with different non-chlamydial species (data not shown).

3.4 The phylogenetic signal of individual protein alignments is highly variable

The phylogenetic signal of each protein alignment was investigated using the SH-test in order to identify the most informative protein sequences. For that, we tested whether the likelihoods of semi-random topologies were significantly lower than the likelihood of their most likely tree. As many as 393 alignments rejected random branching within the *Chlamydiaceae* family with an average p-value < 0.001 (Fig. 4 topologies 1-100), while 12 alignments presented an average p-value > 0.05. In contrast, only 42 alignments rejected random branching of the *Chlamydia*-related species with an average p-value < 0.001. Those proteins include proteins widely used for phylogenetic purpose (e.g. *rpoB*, *rpoC*) as well as the six proteins presenting particular evolutionary histories (Table 1). 203 alignments presented an average p-value > 0.05 (Fig. 4 topologies 101-200).

Overall, the less discriminating alignments (with p-value > 0.05) are mostly short (~143 aa) and conserved with an average tree length of 2.15. Ten out of the 12 less discriminating proteins for the randomized *Chlamydiaceae* topologies are ribosomal proteins.

To test the support of the deep branching nodes of the *Chlamydiales* order, the support of all 15 possible branching of the five *Chlamydiales* families was investigated. In this case, p-values are higher than in the case of semi-random *Chlamydiaceae* and *Chlamydia*-related topologies, indicating that individual alignments do not strongly support any branching at the family level. Only 4 alignments present average p-value below 0.05: tgt, hemH, lgB and aroB, and they all reject the reference topology as well (Table 1).

3.5 Selection of optimal markers for the classification of chlamydial isolates

In order to identify the most phylogenetically informative alignments, the alignments were classified in 9 clusters according to two criteria (Table S8). First, the congruence with the reference tree topology was evaluated by the Robinson-Fould distance and the p-value of the SH-test (individual vs reference tree topology). Second, the strength of the phylogenetic signal was estimated by the ability of individual alignments to reject semi-random topologies of the chlamydial tree. The most promising cluster, number two, exhibits high congruence with the reference topology (p-value of SH-test of 0.98 and Robinson-Fould value of 3.7 on average) and low SH-test p-value for the rejection of semi-random topologies (<0.001 for *Chlamydiaceae* and 0.03 for *Chlamydia*-related bacteria, see Table S8).

The optimal number of protein alignments to concatenate and produce a robust phylogeny was estimated by randomly concatenating an increasing number of alignments. Concatenating 5 alignments already resulted in trees with average bootstrap of value 94.7±1.33% (Fig. S2).

Fig. 5 proposes a new classification scheme for the *Chlamydiales* order. Identity cutoffs of 92.5% and 91% for the 16S and 23S rRNA, respectively, are more representative of the recognized classification. Nine additional markers selected among the 20 most informative ones and presenting various degrees of amino acid sequence conservation (Fig. S4) should be used for genus and species delineations.

3.6 Classification of 5 newly sequenced genomes at genus and species level

Five recently-published genomes were used to assess our classification scheme: *Chlamydia avium* 10DC88, *Chlamydia ibidis* 10-1398/6, *Chlamydia suis* MD56, *Chlamydia gallinacea* 08-1274/3, and *Neochlamydia* S13 (See supplementary Tables 9-13). The classification of the first three strains was confirmed as new species of the *Chlamydia* genus without any conflicting result for all 9 proteins. The orthologue of HemL could not be identified in published sequences of *C. gallinacea*, which did not prevent us to confirm the classification of this strain as a new species of the *Chlamydia* genus. Similarly, the orthologue of SucA could not be identified in *Neochlamydia.* Conflicting percentage identity of the 23S rRNA can be observed between Neochlamydia and the two *Parachlamydia-Protochlamydia* genera (Table S13). In addition, FabI presents a percentage identity higher than the cutoff of 78% with the *Parachlamydia* genus, in contrast to DnaA and protein_325. Altogether, these results still suggest that *Neochlamydia* S13 is a new genus of the *Parachlamydiaceae* family, an affiliation which is congruent with current taxonomy.

4. Discussion

To improve phylogeny and classification, sequences used to reconstruct phylogenetic trees must be carefully chosen (i) to maximize the phylogenetic information, and thus the robustness of the tree, and (ii) minimize potential biases due to horizontal gene transfers , to conserved genes or to genes with high mutation rate leading to saturation. Thus, this work focused on the identification of a set of protein sequences presenting a strong phylogenetic signal allowing an accurate classification of new chlamydial isolates. We identified a set of 20 protein sequences that enable to build robust phylogenetic trees congruent (i.e. in agreement) with a tree based on all chlamydial core proteins (Table 2). This protein set should be used to reconstruct the phylogeny of the *Chlamydiales* order and to determine the taxonomic affiliation of a new strain at species, genus and family level.

4.4 Chlamydial classification

Chlamydial phylogeny has been a topic of intense debate during the last decades, focusing mainly on the classification of *Chlamydiaceae* into one or two genera and the use of 16S rRNA for chlamydial classification (Everett *et al.*, 1999; Schachter *et al.*, 2001; Stephens *et al.*, 2009; Voigt *et al.*, 2012). The analysis of 16S rRNA sequences is not sufficient to delineate species and does not always correlate well with whole genome similarity (Chan *et al.*, 2012; Kim *et al.*, 2014). Due to the democratization of bacterial genome sequencing, whole genome analysis is being more and more used for the taxonomy and the systematics of Bacteria (Chun & Rainey, 2014; Ramasamy *et al.*, 2014).

An ANI of 95-96% is one of the metrics proposed to delineate bacterial species (Kim *et al.*, 2014; Richter & Rosselló-Móra, 2009). This criterion effectively reflects the recognized chlamydial taxonomy at species level (Table S5). Nevertheless, this approach is not well suited for higher taxonomic assignation as there are huge variations in ANI values when comparing genomes from the same or different genera (Kim et al., 2014). The average protein identity (API) could be used as an alternative. Chlamydial families exhibit a relatively wide range of protein identities, which question the relevance of the current classification. Indeed, the *Chlamydiales* order present three highly diverging clades (average protein identities < 50%): the *Chlamydiaceae*, the *Simkaniaceae* and the grouping of the *Waddliaceae*, *Parachlamydiaceae* and *Criblamdiaceae* (Fig. 2a, Table S7). In addition, *C. sequanensis*-*E. lausannensis* (same family) exhibit an average identity which is lower than *W. chondrophila*-*P. acanthamoebae* (different families). Nevertheless, as protein sequences saturation can be important with such distantly-related organisms, simple metrics such as the API are probably not the best approach to distinguish intergenus from interfamily relationships.

4.3 A core proteome of 424 proteins

Taking advantage of the availability of an increasing number of complete and draft chlamydial genome sequences, we identified a core set of 424 proteins. Previous studies identified a larger core genome comprising as many as 560 proteins (Collingro *et al.*, 2011), but included no member of the *Criblamydiaceae* family, and only 4 genomes from *Chlamydia*-related species. The present analysis included 9 genomes of *Chlamydia*-related bacteria including two different genera within the *Criblamydiaceae* family. Moreover, the stringent criterion used in the present work to define orthology, as well as the inclusion of 5 draft genomes also explains such a difference.

A reference phylogeny of the *Chlamydiales* order was constructed based on the concatenated core gene set of 424 proteins using three different methods. In each case, the topology obtained was congruent with previous reconstructions of the phylogenetic relationship between a smaller number of chlamydial strains that was based on 37 ribosomal proteins and four additional proteins (Collingro *et al.*, 2011). Our analysis highlighted the fact that due to their small size and high level of conservation, individual ribosomal proteins do not allow to reconstruct robust phylogenies. However, these proteins still reflect the evolutionary history of the species and are useful to construct robust phylogenies when concatenated.

4.3 Different genes trees but few evidences of HGT

Although core genes are expected to share a similar evolutionary history, phylogenetic reconstruction based on individual protein alignments resulted in 356 different tree topologies with most of the variations concentrated on the most basal nodes of the phylogeny. It is possible that some core genes do not share a common evolutionary history, because of errors in inferring orthology or HGT events. However, this is not expected to be frequent here as we only included proteins presenting reciprocal BBH between all pairwise comparisons. Nevertheless, few proteins in the core gene set exhibited evidence for HGT (Table 1, Fig. 3c), which sheds light on the potential limitations of only using BBH for assigning orthology. The alternative is that those trees are only slightly different, these differences resulting from stochastic errors (Jeffroy *et al.*, 2006). Indeed, when the sequences contain only a poor phylogenetic signal, a maximum likelihood tree can be designated optimal by chance (Shimodaira, 2002). For instance, nearly identical sequences among the 21 species do not allow determining the evolutionary relationships of the different sequences with strong confidence. Consequently, different tree topologies can have a highly similar likelihood, and sometime even identical likelihoods, but only one tree is returned. Lack of information can thus result in a range of slightly different trees, despite the fact that all sequences share a similar evolutionary history.

In order to distinguish stochastic errors from conflicting phylogenetic signals, we evaluated the congruence of phylogenetic signals of individual genes with the tree inferred based on the whole dataset. Various methods have been developed to test the congruence of the phylogenetic signal of different genes (Leigh *et al.*, 2011). Those methods have been applied on genomic scale mainly to evaluate phylogenetic congruence of the core genes, as for 13 gammaproteobacteria (Lerat *et al.*, 2003), but the conclusions of such analyses were disputed (Bapteste *et al.*, 2004). It seems not possible to assume that core genes are free of HGT events and effectively share a common evolutionary history because of the difficulty to detect HGT when considering proteins with weak phylogenetic signal (Bapteste *et al.*, 2005; Susko *et al.*, 2006).

4.4 Important variations in the strength of the phylogenetic signal

As we were primarily interested in highly informative proteins, we evaluated the strength of the phylogenetic signal of individual alignments by comparing the likelihood of suboptimal tree topologies with the likelihood of the best tree. This analysis revealed important differences in the amount of phylogenetic signals provided by different protein sequences as well as important differences in the support of different parts of the *Chlamydiales* phylogeny. On the one hand, the classification of the *Chlamydiaceae* family seems highly supported by most of the core genes as almost any random modification of the topology was significantly rejected (Fig. 4). On the other hand, phylogenetic relationships between *Chlamydia*-related species presented reduced support. Moreover, relationships between the 5 different families belonging to the *Chlamydiales* order were not significantly discriminated by any individual gene.

The poor resolution of the basal branches supporting the different chlamydial families probably results from the very ancient divergence of these families, about 0.7 to 1.4 billion years ago (Greub & Raoult, 2003). Multiple amino acid changes probably accumulated at the same sites, rending difficult the reconstruction of the branching of *Chlamydia*-related families. Homoplasy (i.e. convergence) is also known to have a major impact on the lack of phylogenetic resolution (Rokas & Carroll, 2006; Wiens *et al.*, 2003). It can be overcome by increasing the size of the sequence for example by concatenating several gene sequences, as in the present work, or by increasing the number of taxa, in order to detect multiple substitutions (Delsuc *et al.*, 2005; Jeffroy *et al.*, 2006).

4.5 New chlamydial classification procedure

The evaluation of the strength of the phylogenetic signal allowed the selection of 20 highly discriminant and taxonomically informative core proteins that should be used in chlamydial taxonomy. A minimum of 8 of these selected sequences should be used to construct robust trees with an average boostrap above 95% (Fig. S2). In addition to the reconstruction of robust phylogenetic trees, we propose a new classification scheme based on both 16S/23S sequences as well as 9 of these 20 proteins (Fig. 5). Four proteins more conserved than the average (see Supplementary Table 6) were chosen to distinguish different genus, and five highly divergent proteins to distinguish different species. As multiple sequences are proposed to classify new isolates, this approach is robust to a few number of missing genes. In case of conflicting results, a "majority" rule should be first considered, i.e when a single gene provides conflicting results, the majority prevail. When no majority is present, we then propose to adopt a polyphasic taxonomic approach relying on whole genome phylogeny, genetic distances and phenotypic data. We recommend the use of the global pairwise alignment algorithm from Needleman-Wunsch, and to calculate identity values without considering gaps (complete deletion). Indeed, methods used to align sequences and calculate pairwise identity are known to impact the resulting identity score. For instance, multiple sequence alignment, as opposed to pairwise sequence alignment, is known to yield bigger distances, which tend to inflate the number of taxonomic units (Chen *et al.*, 2013; Lagkouvardos *et al.*, 2013; Sun *et al.*, 2012).

The validity of this new approach could be confirmed with the classification of 5 newly sequenced genomes. One case of conflicting data was resolved by using the majority rule. For the two strains missing a gene, the absence of these genes in the full genome cannot be definitely confirmed, since both genomes are incomplete genome assemblies. Indeed, SucA was successfully retrieved in a genome assembly of another *Neochlamydia* strain recently sequenced in Lausanne (unpublished data).

Due to the very divergent sequences of *Chlamydia*-related families, it is impossible to design primers to sequence the proposed genes in any new strain of the *Chlamydiales* order. Thanks to the democratization of new sequencing technologies, we recommend to sequence the whole genome for the taxonomic characterization of available strains and to concatenate the sequences of the 9 genes to derive the taxonomic affiliation of a new strain. Alternatively, when the isolate has not been obtained in culture and the insufficient number of DNA copies present in the sample prevents genome sequencing, it is possible to obtain the sequences of most of the 20 discriminant and taxonomically informative proteins by designing family-level broad-range primers of the corresponding protein-encoding genes.

6 Conclusion

In this study, we explored different approaches to determine the ability of core *Chlamydiales* proteins to produce robust phylogenies. The reconstruction of chlamydial phylogeny based on 424 groups of orthologs belonging to 21 different chlamydial genomes resulted in a wide range of tree topologies, confirming as expected that a single gene sequence is not sufficient to construct robust bacterial phylogeny. Despite the fact that nearly all topologies inferred from individual protein alignments were different, only few strong conflicting phylogenetic signals that led to the rejection of the reference tree were found in the core gene set of the *Chlamydiales*. No straightforward parameter allowed the quantification of phylogenetic information. Consequently, we combined different parameters, such as the rejection of semi-random topologies and the non-rejection of the reference topology to select a small set of protein sequences that optimally reconstruct a highly supported phylogenetic tree of the *Chlamydiales* order and provide a robust classification scheme. At least 9 of these 20 proteins should be used to accurately assign newly discovered chlamydial strains at family, genus and species level within the *Chlamydiales* order.

6 Acknowledgements

The computations were performed at the Vital-IT Center for high-performance computing of the Swiss Institute of Bioinformatics (SIB, Lausanne, http://www.vital-it.ch).

REFERENCES

Aguileta, G., Marthey, S., Chiapello, H., Lebrun, M.-H., Rodolphe, F., Fournier, E., Gendrault-Jacquemard, a & Giraud, T. (2008). Assessing the performance of single-copy genes for recovering robust phylogenies. *Syst Biol* **57**, 613–27.

Altschul, S. F., Madden, T. L., Schäffer, a a, Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–402.

Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R. L. & Doolittle, W. F. (2005). Do orthologous gene phylogenies really support tree-thinking? *BMC Evol Biol* **5**, 33.

Bapteste, E., Boucher, Y., Leigh, J. & Doolittle, W. F. (2004). Phylogenetic reconstruction and lateral gene transfer. *Trends Microbiol* **12**, 406–11.

Bavoil, P., Kaltenboeck, B. & Greub, G. (2013). In Chlamydia veritas. *Pathog Dis* **67**, 89–90.

Bertelli, C., Collyn, F., Croxatto, A., Rückert, C., Polkinghorne, A., Kebbi-Beghdadi, C., Goesmann, A., Vaughan, L. & Greub, G. (2010). The Waddlia genome: a window into chlamydial biology. *PLoS One* **5**, e10890.

Brunelle, B. W. & Sensabaugh, G. F. (2006). The ompA Gene in Chlamydia trachomatis Differs in Phylogeny and Rate of Evolution from Other Regions of the Genome The ompA Gene in Chlamydia trachomatis Differs in Phylogeny and Rate of Evolution from Other Regions of the Genome **74**.

Chan, J. Z.-M., Halachev, M. R., Loman, N. J., Constantinidou, C. & Pallen, M. J. (2012). Defining bacterial species in the genomic era: insights from the genus Acinetobacter. *BMC Microbiol* **12**, 302. BMC Microbiology.

Chen, W., Zhang, C. K., Cheng, Y., Zhang, S. & Zhao, H. (2013). A comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS One* **8**, e70837.

Chun, J. & Rainey, F. a. (2014). Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *Int J Syst Evol Microbiol* **64**, 316–324.

Collingro, A., Tischler, P., Weinmaier, T., Penz, T., Heinz, E., Brunham, R. C., Read, T. D., Bavoil, P. M., Sachse, K. & other authors. **(2011).** Unity in variety--the pan-genome of the Chlamydiae. *Mol Biol Evol* **28**, 3253–70.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–5.

Delsuc, F., Brinkmann, H. & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6**, 361–75.

Domman, D., Collingro, A., Lagkouvardos, I., Gehre, L., Weinmaier, T., Rattei, T., Subtil, A. & Horn, M. (2014). Massive Expansion of Ubiquitination-Related Gene Families within the Chlamydiae. *Mol Biol Evol*.

Everett, K. D., Bush, R. M. & Andersen, a a. **(1999).** Emended description of the order Chlamydiales, proposal of Parachlamydiaceae fam. nov. and Simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family Chlamydiaceae, including a new genus and five new species, and standards. *Int J Syst Bacteriol* **49 Pt 2**, 415–40.

Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S. & other authors. **(1980).** The phylogeny of prokaryotes. *Science* **209**, 457–63.

Fraley, C. & Raftery, A. E. (2006). MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-Based Clustering.

Goldman, N. (1998). Phylogenetic information and experimental design in molecular systematics. *Proc Biol Sci* **265**, 1779–86.

Gouy, M., Guindon, S. & Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**, 221–4.

Greub, G. (2013). International Committee on Systematics of Prokaryotes * Subcommittee on the taxonomy of Chlamydiae: Minutes of the closed meeting, 23 February 2011, Ascona, Switzerland. *Int J Syst Evol Microbiol* **63**, 1934–1935.

Greub, G. (2010a). International Committee on Systematics of Prokaryotes. Subcommittee on the taxonomy of the Chlamydiae: minutes of the inaugural closed meeting, 21 March 2009, Little Rock, AR, USA. *Int J Syst Evol Microbiol* **60**, 2691–3.

Greub, G. (2010b). International Committee on Systematics of Prokaryotes. Subcommittee on the taxonomy of the Chlamydiae: minutes of the closed meeting, 21 June 2010, Hof bei Salzburg, Austria. *Int J Syst Evol Microbiol* **60**, 2694.

Greub, G. & 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 History of the ADP / ATP-Translocase-Encoding Gene , a Parasitism Gene Transferred from a Chlamydiales Ancestor t **69**, 5530–5535.

Guindon, S. & Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst Biol* **52**, 696–704.

Harris, S. R., Clarke, I. N., Seth-Smith, H. M. B., Solomon, A. W., Cutcliffe, L. T., Marsh, P., Skilton, R. J., Holland, M. J., Mabey, D. & other authors. **(2012).** Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. *Nat Genet* **44**, 413–9, S1. Nature Publishing Group.

Horn, M. (2008). Chlamydiae as symbionts in eukaryotes. *Annu Rev Microbiol* **62**, 113–31.

Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends Genet* **22**, 225–31.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–66.

Kim, M., Oh, H.-S., Park, S.-C. & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* **64**, 346–51.

Klint, M., Fuxelius, H.-H., Goldkuhl, R. R., Skarin, H., Rutemark, C., Andersson, S. G. E., Persson, K. & Herrmann, B. (2007). High-resolution genotyping of Chlamydia trachomatis strains by multilocus sequence analysis. *J Clin Microbiol* **45**, 1410–4.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12.

Lagkouvardos, I., Weinmaier, T., Lauro, F. M., Cavicchioli, R., Rattei, T. & Horn, M. (2013). Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. *ISME J*.

Leigh, J. W., Lapointe, F.-J., Lopez, P. & Bapteste, E. (2011). Evaluating phylogenetic congruence in the post-genomic era. *Genome Biol Evol* **3**, 571–87.

Lerat, E., Daubin, V. & Moran, N. a. **(2003).** From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol* **1**, E19.

Lienard, J., Croxatto, A., Prod'hom, G. & Greub, G. (2011). Estrella lausannensis, a new star in the Chlamydiales order. *Microbes Infect* **13**, 1232–41. Elsevier Masson SAS.

Pannekoek, Y., Morelli, G., Kusecek, B., Morré, S. a, Ossewaarde, J. M., Langerak, A. a & van der Ende, A. (2008). Multi locus sequence typing of Chlamydiales: clonal groupings within the obligate intracellular bacteria Chlamydia trachomatis. *BMC Microbiol* **8**, 42.

Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D. & Pupko, T. (2010). GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res* **38**, W23–8.

R Core Team. **(2014).** R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramasamy, D., Mishra, A. K., Lagier, J.-C., Padhmanabhan, R., Rossi, M., Sentausa, E., Raoult, D. & Fournier, P.-E. (2014). A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol* **64**, 384–91.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–7.

Richter, M. & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* **106**, 19126–31.

Robinson, D. F. & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math Biosci* **53**, 131–147.

Rokas, A. & Carroll, S. B. (2006). Bushes in the tree of life. *PLoS Biol* **4**, e352.

Rosselló-Mora, R. & Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiol Rev* **25**, 39–67.

Schachter, J., Stephens, R. S., Timms, P., Kuo, C., Bavoil, P. M., Birkelund, S., Boman, J., Caldwell, H., Campbell, L. a & other authors. **(2001).** Radical changes to chlamydial taxonomy are not necessary just yet. *Int J Syst Evol Microbiol* **51**, 249; author reply 251–3.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–3.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst Biol* **51**, 492–508.

Shimodaira, H. & Hasegawa, M. (1999). Letter to the Editor Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol* 1114–1116.

Snel, B., Huynen, M. a & Dutilh, B. E. (2005). Genome trees and the nature of genome evolution. *Annu Rev Microbiol* **59**, 191–209.

Stephens, R. S., Myers, G., Eppinger, M. & Bavoil, P. M. (2009). Divergence without difference: phylogenetics and taxonomy of Chlamydia resolved. *FEMS Immunol Med Microbiol* **55**, 115–9.

Sukumaran, J. & Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–71.

Sun, Y., Cai, Y., Huse, S. M., Knight, R., Farmerie, W. G., Wang, X. & Mai, V. (2012). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinform* **13**, 107–21.

Susko, E., Leigh, J., Doolittle, W. F. & Bapteste, E. (2006). Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the gamma-proteobacteria. *Mol Biol Evol* **23**, 1019–30.

Voigt, A., Schöfl, G. & Saluz, H. P. (2012). The Chlamydia psittaci genome: a comparative analysis of intracellular pathogens. *PLoS One* **7**, e35097.

Wiens, J. J., Chippindale, P. T. & Hillis, D. M. (2003). When are phylogenetic analyses misled by convergence? A case study in Texas cave salamanders. *Syst Biol* **52**, 501–14.

Table 1 | Protein alignments presenting strong evidence of conflicting phylogenetic signal with the reference tree

gene	Accession *C. trachomatis* D	Orthogroup ID	Tree length	Align. length	RF [*]	SH[†] reference tree	Mean SH[†] random topologies	SD SH[†] random topologies	Mean SH[†] random *Chlam.* Classic	Mean SH[†] random *Chlam.*-like	Mean SH[†] 15 Familiy topo.	Annotation
-	15604821	57	11.08	250	10	0.1363	0.1908	0.2704	0	0.1464±0.0338	0.8352±0.0233	hypothetical protein
tgt	15604913	80	4.28	406	10	0.0000	0.0087	0.0902	0	0.0000	0.0000	queuine tRNA-ribosyltransferase
aroB	15605093	173	9.70	423	8	0.0330	0.0655	0.1815	0	0.0135±0.0190	0.4065±0.0655	3-dehydroquinate synthase
mdhC	15605100	175	4.23	340	14	0.0384	0.1279	0.2000	0	0.1034±0.0180	0.5781±0.0128	malate dehydrogenase
hemH	15605213	228	8.81	375	10	0.0000	0.0165	0.0933	0	0.0000	0.0165±0.0009	ferrochelatase
birA	15605458	360	10.25	263	10	0.0433	0.0631	0.1785	0	0.0026±0.0015	0.5737±0.1646	biotin--protein ligase
nrdB	15605563	408	8.12	368	12	0.1134	0.0930	0.2377	5e-04±0.000283	0.0000	0.7773±0.2868	ribonucleotide-diphosphate reductase subunit beta
glgB	15605602	424	5.45	769	8	0.0000	0.0142	0.1126	0	0.0000	0.0040±0.0057	glycogen branching enzyme

[*] **RF:** Robinson-Fould distance when a given tree topology is compared to the reference tree obtained with the contacteantion of all 424 core protein sequences.

[†] **SH:** p-value of the SH-test.

Table 2 | The 20 most phylogenetically informative proteins of the core genome of the *Chlamydiales*

Gene	Accession *C .trachomatis* D	Orthogroup ID	Tree length	Align. length	RF[*]	SH[†] reference tree	Mean SH[†] random topologies	SD SH[†] random topologies	Mean SH[†] random Chlam. Classic	Mean SH[†] random Chlam.-like	SD SH[†] random Chlam.-like	Mean SH[†] 15 Familiy topolo.	SD SH[2] 15 Familiy topo.	Annotation
sucA	15604773	29	6.36	996	4	0.99	0.08	0.24	0.00	0.00	0.00	0.94	0.05	2-oxoglutarate dehydrogenase subunit E1
tyrS	15604781	32	5.05	446	2	0.95	0.10	0.24	0.00	0.06	0.03	0.95	0.05	tyrosyl-tRNA synthetase
fabI	15604823	59	2.47	325	4	0.97	0.13	0.25	0.00	0.12	0.07	0.94	0.02	enoyl-ACP reductase
pepF	15604831	62	5.32	655	6	0.99	0.10	0.25	0.00	0.04	0.04	0.93	0.02	oligoendopeptidase F
adk	15604847	67	8.08	289	2	1.00	0.09	0.25	0.00	0.01	0.01	0.96	0.01	adenylate kinase
hemL	15604930	83	6.47	496	8	0.90	0.09	0.25	0.00	0.04	0.05	0.92	0.04	glutamate-1-semialdehyde aminotransferase
fabG	15604958	93	4.52	254	2	1.00	0.13	0.28	0.00	0.05	0.02	0.95	0.01	3-ketoacyl-ACP reductase
dnaA	15604971	103	5.45	494	2	0.99	0.09	0.24	0.00	0.05	0.07	0.95	0.00	chromosomal replication initiation protein
clpC	15605007	126	2.31	902	4	1.00	0.08	0.24	0.00	0.01	0.01	0.93	0.05	ClpC protease ATPase
dut	15605013	130	3.99	156	8	0.93	0.15	0.27	0.00	0.11	0.10	0.92	0.02	deoxyuridine 5'-triphosphate nucleotidohydrolase
lpxK	15605127	190	9.66	453	4	1.00	0.13	0.27	0.00	0.06	0.04	0.94	0.03	tetraacyldisaccharide 4'-kinase
argS	15605181	218	4.76	594	6	0.99	0.10	0.25	0.00	0.04	0.05	0.91	0.05	arginyl-tRNA synthetase
gspF	15605299	281	6.57	401	2	1.00	0.09	0.24	0.00	0.01	0.01	0.93	0.08	general secretion pathway protein F
rpoN	15605340	304	8.29	527	4	0.98	0.08	0.24	0.00	0.00	0.00	0.94	0.04	RNA polymerase factor sigma-54
greA	15605367	317	5.46	741	4	0.96	0.07	0.24	0.00	0.00	0.00	0.95	0.04	transcript cleavage factor
topA	15605375	323	3.60	911	0	1.00	0.07	0.24	0.00	4e-04	0.00	0.95	0.06	DNA topoisomerase I/SWI
-	15605380	325	7.07	455	4	0.98	0.09	0.25	0.00	0.00	0.00	0.95	0.02	hypothetical protein
-	15605424	341	4.89	243	4	0.92	0.11	0.25	0.00	0.04	0.02	0.95	0.05	hypothetical protein
ftsK	15605472	364	7.75	958	2	1.00	0.07	0.24	0.00	0.01	0.02	0.94	0.04	cell division protein FtsK
priA	15605511	385	5.18	776	2	1.00	0.09	0.26	0.00	0.00	0.00	0.95	0.05	primosome assembly protein PriA

[*] **RF:** Robinson-Fould distance when a given tree topology is compared to the reference tree obtained with the contacteantion of all 424 core protein sequences.

[†] **SH:** p-value of the SH-test.

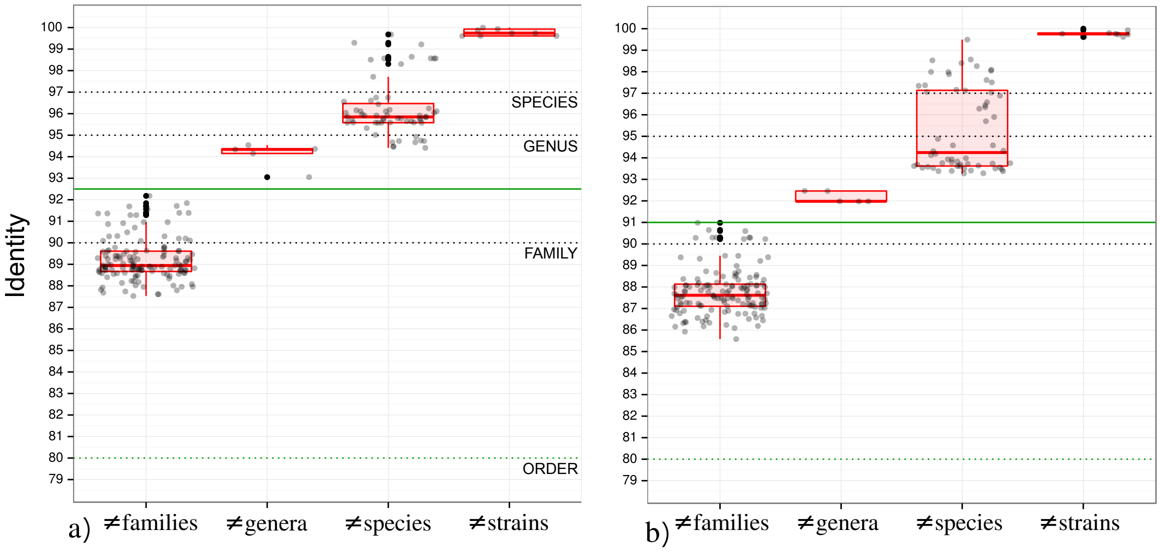


Figure 1: Ribosomal RNA identity based on pairwise global alignments. a) 16S and b) 23S rRNA identity. Dotted lines indicate 16S identity thresholds proposed by Everett in 1999 (Everett et al., 1999). Green lines indicate new proposed thresholds of respectively 92.5 and 91 percent for family

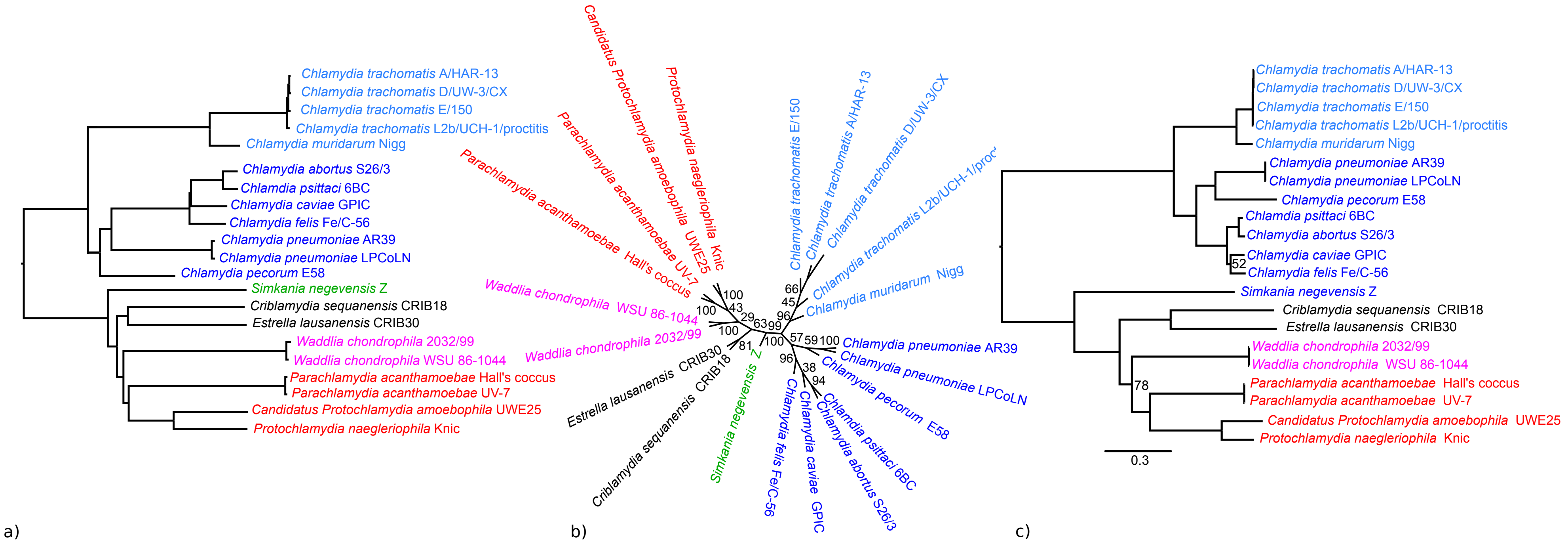

Chlamydia trachomatis A/HAR-13
Chlamydia trachomatis D/UW-3/CX
Chlamydia trachomatis E/150
Chlamydia trachomatis L2b/UCH-1/proctitis
Chlamydia muridarum Nigg
Chlamydia abortus S26/3
Chlamdia psittaci 6BC
Chlamydia caviae GPIC
Chlamydia felis Fe/C-56
Chlamydia pneumoniae AR39
Chlamydia pneumoniae LPCoLN
Chlamydia pecorum E58
Simkania negevensis Z
Criblamydia sequanensis CRIB18
Estrella lausanensis CRIB30
Waddlia chondrophila 2032/99
Waddlia chondrophila WSU 86-1044
Parachlamydia acanthamoebae Hall's coccus
Parachlamydia acanthamoebae UV-7
Candidatus Protochlamydia amoebophila UWE25
Protochlamydia naegleriophila Knic
a)
Candidatus Protochlamydia amoebophila UWE25
Protochlamydia naegleriophila Knic
Parachlamydia acanthamoebae UV-7
Parachlamydia acanthamoebae Hall's coccus
Waddlia chondrophila WSU 86-1044
Waddlia chondrophila 2032/99
Estrella lausanensis CRIB30
Criblamydia sequanensis CRIB18
Simkania negevensis Z
Chlamydia felis Fe/C-56
Chlamydia caviae GPIC
Chlamydia abortus S26/3
Chlamdia psittaci 6BC
Chlamydia pecorum E58
Chlamydia pneumoniae LPCoLN
Chlamydia pneumoniae AR39
Chlamydia muridarum Nigg
Chlamydia trachomatis L2b/UCH-1/proc
Chlamydia trachomatis D/UW-3/CX
Chlamydia trachomatis A/HAR-13
Chlamydia trachomatis E/150
100
43
100
29
63
99
100
81
100
66
45
96
57
59
100
96
38
94
b)
Chlamydia trachomatis A/HAR-13
Chlamydia trachomatis D/UW-3/CX
Chlamydia trachomatis E/150
Chlamydia trachomatis L2b/UCH-1/proctitis
Chlamydia muridarum Nigg
Chlamydia pneumoniae AR39
Chlamydia pneumoniae LPCoLN
Chlamydia pecorum E58
Chlamdia psittaci 6BC
Chlamydia abortus S26/3
Chlamydia caviae GPIC
52
Chlamydia felis Fe/C-56
Simkania negevensis Z
Criblamydia sequanensis CRIB18
Estrella lausanensis CRIB30
Waddlia chondrophila 2032/99
Waddlia chondrophila WSU 86-1044
78
Parachlamydia acanthamoebae Hall's coccus
Parachlamydia acanthamoebae UV-7
Candidatus Protochlamydia amoebophila UWE25
Protochlamydia naegleriophila Knic
0.3
c)

Figure 2: Phylogenetic trees of the Chlamydiales order based on 424 core proteins. a) Midpoint rooted tree constructed by neighbor-joining based on average identity of the genes shared between pairs of genomes (see Suppl. Table S7).Blue: *Chlamydiaceae* family (with the former *Chlamydophila* genus in dark, and *Chlamydia trachomatis* and *Chlamydia muridarum* species in light). Pink: *Simkaniaceae* family. Black: two genus of the *Criblamydiaceae* family. Pink: *Waddliaceae* family. Red: two genus of the *Parachlamydiaceae* family. b) Consensus tree based on the 424 individual core protein phylogenies c) Midpoint rooted ML tree based on concatenation of the 424 core proteins. Bootstrap support values are indicated when inferior to 100.

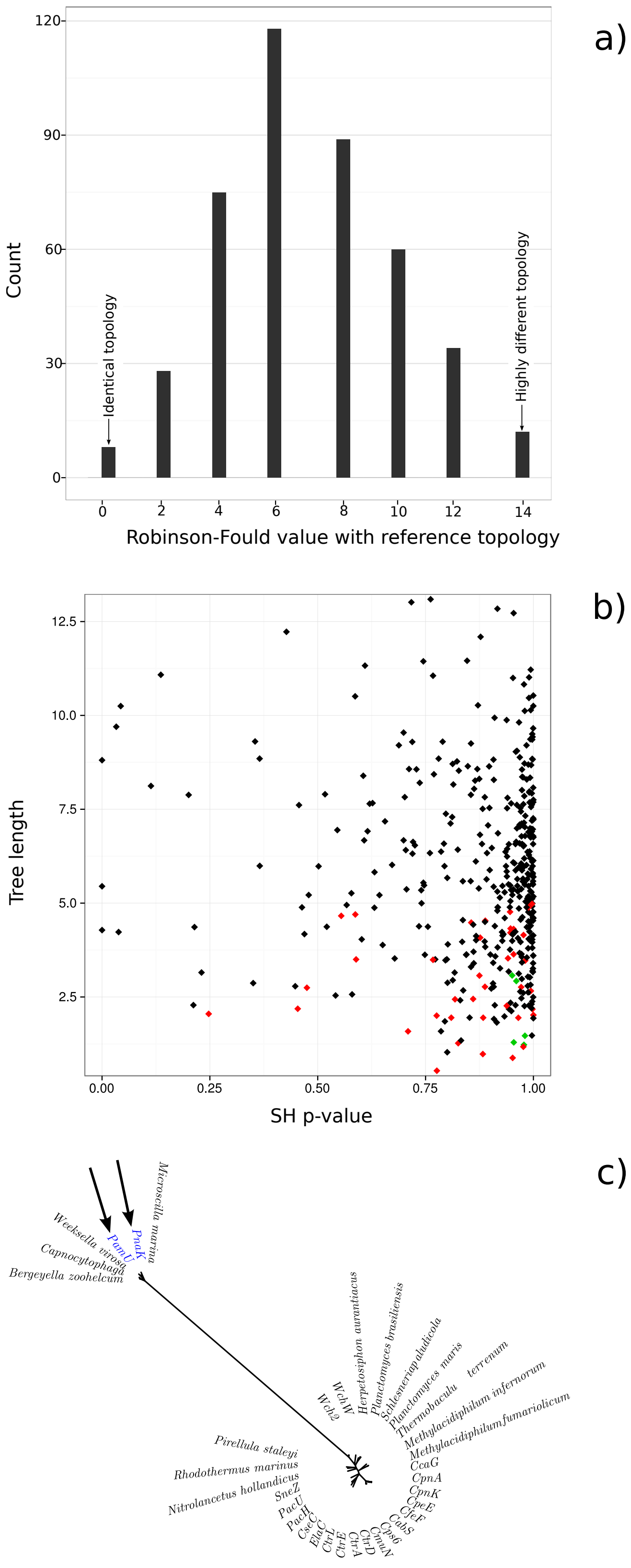


Figure 3: Congruence of *Chlamydiales* phylogeny. a) Robinson-Fould distance of individual gene trees compared to the reference tree topology. A distance of 0 indicates identical topologies. b) SH-test p-value as a function of tree length. The position of the 38 ribosomal proteins is indicated in red. In green is the position of RpoB, RpoC, GyrB, RecA and Ef-Tu, five proteins frequently used for phylogenetic purpose. c) Conflicting phylogeny of ribonucleotide-diphosphate reductase subunit beta (nrdB). The two species of *Protochlamydia* genus (in blue, arrows) cluster with non-chlamydial species. For this analysis, the five best non chlamydial BLAST hits were obtained from the NCBI nr for *Chlamydia trachomatis* D/UW-3/CX, *Simkania negevensis* Z, *Criblamydia sequanensis* CRIB-18, *Estrella lausannensis* CRIB-30, *Waddlia chondrophila* WSU 86-1044, *Protochlamydia naegleriophila* Knic and *Parachlamydia acanthamoebae* Hall's coccus, and redundancy was removed before phylogenetic reconstruction.

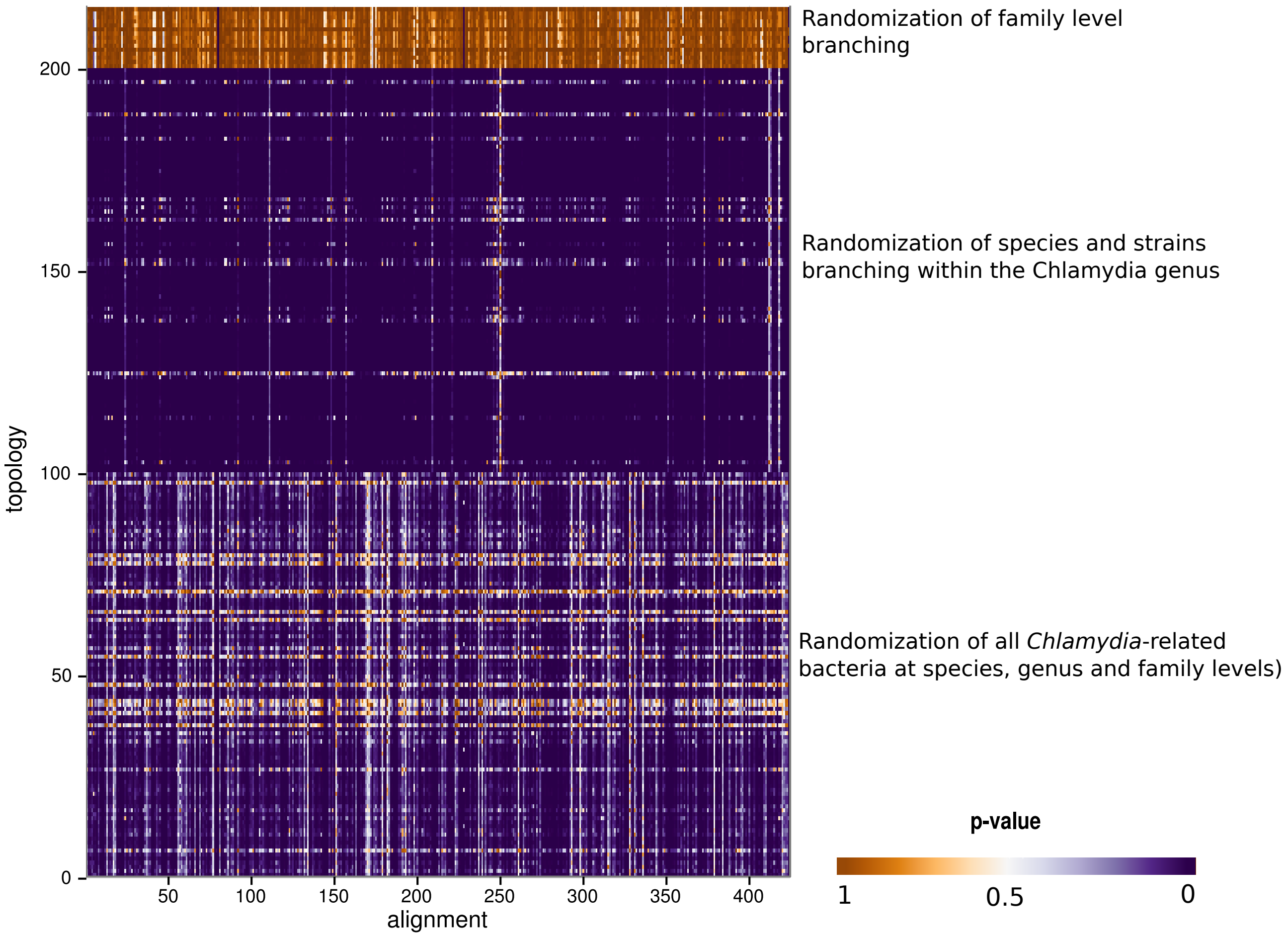


Figure 4: Rejection of random topologies. Heatmap of the SH-test p-value that reflects the statistical power of individual protein alignments to reject semi-random topologies. Topologies 1-100) Fixation of the *Chlamydiaceae* species tree and randomization of the *Chlamydia*-related species position. Topologies 101-200). Fixation of the *Chlamydia*-related species position and randomization of the *Chlamydiaceae* species tree. Topologies 201-215) all 15 possible branching of the 4 *Chlamydia*-related families (intra-family branching was not modified).

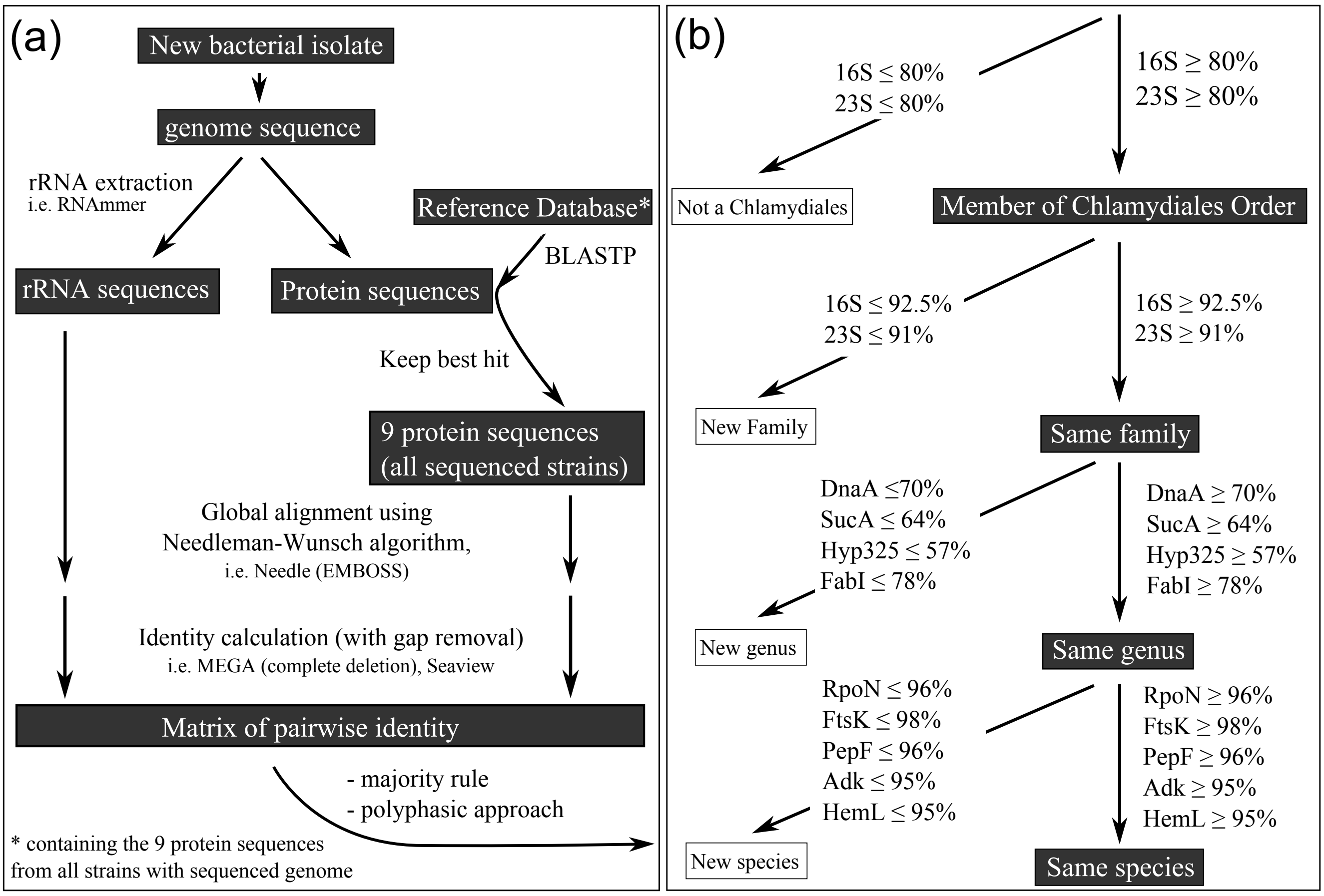


Figure 5: **Classification scheme**. a) Retrieval of 9 conserved taxonomically informative gene products from a newly sequenced strain. b) Classification based on the percentage of sequence identity between 9 protein sequences of the new isolate and all other sequenced members of the *Chlamydiales* order.

International Journal of Systematic and Evolutionary Microbiology

Taxogenomics of the *Chlamydiales*: supplementary material

Trestan Pillonel [1,2], Claire Bertelli [1,2], Nicolas Salamin [2,3] and Gilbert Greub [1,*]

[1] Center for Research on Intracellular Bacteria, Institute of Microbiology, University Hospital Center and University of Lausanne, Lausanne, Switzerland.

[2] SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3] Department of Ecology and Evolution, Biophore, University of Lausanne, Lausanne, Switzerland.

* Corresponding author:

Pr. Gilbert GREUB, Institute of Microbiology, Bugnon 48, CH-1011 Lausanne, Swizerland, Tel : +41213144979, fax : +41213144060, email : gilbert.greub@chuv.ch

Supplementary table 1| Genome informations.

Genome	Family	Abbreviation	Number of proteins	Size (bp)	Accession
Chlamydia trachomatis A/HAR-13	Chlamydiaceae	CtrA	911	1044459	CP000051
Chlamydia trachomatis D/UW-3/CX	Chlamydiaceae	CtrD	895	1042519	AE001273
Chlamydia trachomatis E/150	Chlamydiaceae	CtrE	927	1042996	CP001886
Chlamydia trachomatis L2b/UCH-1/proctitis	Chlamydiaceae	CtrL	873	1038863	AM884177
Chlamdia psittaci 6BC*	Chlamydiaceae	Cps6	967	1171660	CP002549
Chlamydia pneumoniae AR39*	Chlamydiaceae	CpnA	1112	1229853	AE002161
Chlamydia pneumoniae LPCoLN*	Chlamydiaceae	CpnK	1097	1241020	CP001713
Chlamydia muridarum Nigg	Chlamydiaceae	CmuN	903	1072950	AE002160
Chlamydia abortus S26/3*	Chlamydiaceae	CabS	932	1144377	CR848038
Chlamydia pecurum E58*	Chlamydiaceae	CpeE	988	1106197	CP002608
Chlamydia felis Fe/C-56*	Chlamydiaceae	CfeF	1005	1166239	AP006861
Chlamydia caviae GPIC*	Chlamydiaceae	CcaG	998	1173390	AE015926
Parachlamydia acanthamoebae UV-7	Parachlamydiaceae	PacU	2789	3072383	FR872580
Parachlamydia acanthamoebae Hall's coccus	Parachlamydiaceae	PacH	2809	2971261	ACZE00000000
Protochlamydia naegleriophila Knic	Parachlamydiaceae	PnaK	3444	3011277	PRJEB7990
Candidatus *Protochlamydia amoebophila* UWE25	Parachlamydiaceae	PamU	2031	2414465	BX908798
Simkania negevensis Z	Simkaniaceae	SneZ	2381	2496337	FR872582
Waddlia chondrophila WSU 86-1044	Waddliaceae	WchW	1934	2116312	CP001928.1
Waddlia chondrophila 2032/99	Waddliaceae	Wch2	2015	2139757	PRJEA49037
Criblamydia sequanensis CRIB-18	Criblamydiaceae	CseC	2681	3018308	CCJ000000000
Estrella lausannensis CRIB-30	Criblamydiaceae	ElaC	2434	2861702	PRJEB7018
Neochlamydia sp. S13$^+$	Parachlamydiaceae	NeoS	-	-	BASK00000000.1
Chlamydia avium 10DC88$^+$	Chlamydiaceae	Cav1	940	1041170	CP006571.1
Chlamydia gallinacea 08-1274/3$^+$	Chlamydiaceae	Cga0	907	-	NZ_AWUS01000000
Chlamydia ibidis 10-1398/6$^+$	Chlamydiaceae	Cib1	1018	-	APJW01000000
Chlamydia suis MD56$^+$	Chlamydiaceae	CsuM	931	-	AYKJ01000000

*previously named *Chlamydophila*

$^+$newly sequenced strain used to evaluate the classification scheme developed based on 21 chlamydial genomes

Supplementary table 2| Best model of amino acid replacement according to the Bayesian Information Criterion (BIC):

Model	Number of proteins	Number of ribosomal proteins
CpREV+I+G	3	-
Dayhoff	1	1
Dayhoff+I+G	3	1
HIVb+I+G	1	1
JTT+I+G	7	-
JTT+I+G+F	5	-
LG+I+G	365	32
LG+I+G+F	31	1
RtREV+I+G	2	1
RtREV+I+G+F	1	-
VT+I+G	2	-
WAG+I+G	3	2

Supplementary table 3| 16S rRNA pairwise identity.

	CabS	CcaG	CfeF	CmuN	CpeE	CpnA	CpnK	Cps6	CtrA	CtrD	CtrE	CtrL	CseC	ElaC	PacH	PacU	PamU	PnaK	SneZ	Wch2
CcaG	99.29																			
CfeF	98.31	98.51																		
CmuN	96.04	96.11	95.79																	
CpeE	96.56	96.75	96.61	96.16																
CpnA	95.85	95.59	95.33	94.75	95.92															
CpnK	96.11	95.85	95.59	95.01	96.24	99.61														
Cps6	99.68	99.22	98.64	96.11	96.75	96.18	96.44													
CtrA	95.65	95.91	95.58	98.57	95.58	94.47	94.73	95.71												
CtrD	95.65	95.91	95.58	98.57	95.58	94.47	94.73	95.71	100											
CtrE	95.71	95.97	95.77	98.57	95.77	94.67	94.93	95.91	99.74	99.74										
CtrL	95.59	96.04	95.84	98.57	95.58	94.41	94.67	95.78	99.61	99.61	99.87									
CseC	89.47	89.48	88.75	89.58	89.62	88.46	88.72	89.73	88.89	88.83	88.88	88.82								
ElaC	89.62	89.64	89.24	89.4	89.29	88.22	88.72	89.82	89.38	89.26	89.29	89.35	93.05							
PacH	88.52	88.18	88.4	88.72	88.55	89.58	89.6	88.73	89.03	88.9	88.93	89	90.32	91.71						
PacU	88.59	88.55	88.77	88.94	88.61	89.61	89.63	88.79	89.25	89.12	89.15	89.22	90.49	91.84	99.93					
PamU	89.3	89.24	89.13	88.58	88.89	88.63	88.95	89.37	88.73	88.73	88.8	88.87	92.18	91.69	94.15	94.34				
PnaK	88.68	88.93	88.49	88.95	87.95	87.82	88.14	89.05	88.01	88.02	88.08	87.96	90.91	91.56	94.36	94.54	97.71			
SneZ	88.4	88.34	87.53	87.79	87.92	87.96	88.23	88.47	87.74	87.61	87.61	87.68	88.61	88.79	90.71	90.87	89.66	90.11		
Wch2	89.4	89.34	88.98	88.76	89.17	88.58	88.9	89.78	88.87	88.75	88.68	88.75	90.09	90.3	91.29	91.36	90.96	91.39	90.2	
WchW	89.4	89.34	88.98	88.76	89.17	88.58	88.9	89.78	88.87	88.75	88.68	88.75	90.09	90.3	91.29	91.36	90.96	91.39	90.2	100

Supplementary table 4| 23S rRNA pairwise identity.

	CabS	CcaG	CfeF	CmuN	CpeE	CpnA	CpnK	Cps6	CtrA	CtrD	CtrE	CtrL	CseC	ElaC	PacH	PacU	PamU	PnaK	SneZ	Wch2

CcaG	98.09																			
CfeF	98.26	98.53																		
CmuN	94.31	94.33	94.13																	
CpeE	96.28	95.7	95.9	94.18																
CpnA	97.13	96.89	97.06	94.58	96.44															
CpnK	97.34	97.03	97.16	94.88	96.58	99.76														
Cps6	99.49	98.4	98.57	94.58	96.35	97.37	97.51													
CtrA	93.38	93.54	93.27	97.88	93.28	93.55	93.69	93.62												
CtrD	93.49	93.64	93.38	97.99	93.39	93.58	93.79	93.72	99.76											
CtrE	93.69	93.98	93.61	98.05	93.5	93.73	93.93	93.93	99.62	99.73										
CtrL	93.52	93.75	93.41	97.99	93.46	93.62	93.83	93.76	99.62	99.73	99.8									
CseC	86.38	85.92	86.17	86.45	86.19	86.16	86.24	86.32	86.75	86.77	86.65	86.93								
ElaC	86.89	86.96	86.76	87.47	85.58	86.66	86.77	87.23	87.08	87.25	87.24	87.25	91.98							
PacH	88.33	88.37	88.49	88.73	87.98	88	88.06	88.56	88.02	88.13	88.09	88.09	90.29	90.3						
PacU	88.33	88.37	88.49	88.73	87.98	88	88.06	88.56	88.02	88.13	88.09	88.09	90.29	90.3	100					
PamU	87.74	87.89	88.08	87.61	87.23	87.56	87.7	87.96	86.92	87.14	87.14	87.03	90.6	90.65	91.98	91.98				
PnaK	88.21	87.92	88.19	87.47	87.64	88	88.07	88.44	87.03	87.2	87.27	87.27	90.98	90.28	92.46	92.46	97.62			
SneZ	86.51	86.58	86.39	87.13	86.09	86.36	86.56	86.64	86.89	87.06	86.84	87.18	89.31	88.47	90.23	90.23	88.98	89.21		
Wch2	87.78	87.14	87.49	87.83	86.34	87.11	87.21	87.69	87.5	87.54	87.43	87.58	88.65	87.53	89.38	89.38	87.76	87.84	88.14	
WchW	87.71	87.19	87.42	87.8	86.45	87.29	87.38	87.62	87.46	87.51	87.39	87.54	88.64	87.47	89.45	89.45	87.82	87.81	88.07	99.93

Supplementary table 5| Mummer-based average nucleotide identity. Values were only reported if the NUCmer alignment covered a minimum of 50% of the reference genome. Values from strains from the same species are highlighted in black.

	CpnK	Cps6	CabS	CcaG	CmuN	CtrL	CtrA	CtrD	WchW	PacU
CpnA	98.99									
CabS		92.52								
CcaG		84.17	83.89							
CfeF		83.81	83.63	84.63						
CtrL					83.20					
CtrA					83.17	99.07				
CtrD					83.23	99.09	99.62			
CtrE					83.17	99.12	99.28	99.35		
Wch2									99.36	
PacH										99.66

Supplementary table 6 | Number of reciprocal best BLAST hits between the 21 *Chlamydiales* proteomes. Cell colors reflect current classification. Light grey: strains from the same species. Intermediate grey: species from the same genus. Black: species from different genera. White: species from different families.

	CabS	**CcaG**	**CfeF**	**CmuN**	**CpeE**	**CpnA**	**CpnK**	**Cps6**	**CtrA**	**CtrD**	**CtrE**	**CtrL**	**CseC**	**ElaC**	**PacH**	**PacU**	**PamU**	**PnaK**	**SneZ**	**Wch2**
CcaG	901																			
CfeF	905	931																		
CmuN	808	817	818																	
CpeE	845	862	870	800																
CpnA	868	870	882	811	860															
CpnK	863	865	875	803	858	998														
Cps6	915	928	932	809	862	874	871													
CtrA	818	823	830	850	809	815	810	817												
CtrD	817	823	828	848	807	813	808	816	890											
CtrE	811	817	818	842	803	806	803	811	875	872										
CtrL	815	819	823	840	805	813	807	813	868	868	858									
CseC	672	677	676	652	670	671	673	676	648	647	651	647								
ElaC	671	670	681	653	664	676	672	672	653	654	651	647	1571							
PacH	680	682	688	651	667	681	683	687	656	656	655	654	1459	1365						
PacU	682	683	691	654	671	683	686	687	659	657	659	656	1465	1379	2450					
PamU	661	665	678	641	652	662	661	666	647	647	644	644	1151	1145	1213	1224				
PnaK	670	680	689	649	666	677	675	675	652	653	649	650	1384	1330	1405	1408	1433			
SneZ	656	656	664	626	658	657	653	661	640	637	635	634	1035	1013	1016	1029	934	1016		
Wch2	632	634	638	608	626	631	623	632	622	619	617	618	1229	1224	1229	1243	1035	1148	917	
WchW	656	655	661	632	652	650	643	653	645	643	641	643	1259	1253	1252	1260	1059	1171	943	1731

Supplementary table 7| Average identity of orthologous proteins. Cell colors reflect current classification. Light grey: strains from the same species. Intermediate grey: species from the same genus. Black: species from different genera. White: species from different families.

	CabS	CcaG	CfeF	CmuN	CpeE	CpnA	CpnK	Cps6	CtrA	CtrD	CtrE	CtrL	CseC	ElaC	PacH	PacU	PamU	PnaK	SneZ	Wch2
CcaG	84.84																			
CfeF	84.35	85.45																		
CmuN	65.03	65.38	65.25																	
CpeE	67.25	67.12	67.02	62.37																
CpnA	67.94	68.42	67.97	62.82	67.85															
CpnK	68.05	68.42	68.12	62.85	67.96	98.75														
Cps6	94.04	84.9	84.3	65.36	67.12	68.18	68.25													
CtrA	65	65.14	64.89	85.19	62.2	62.71	62.69	65.19												
CtrD	65.06	65.12	65.01	85.25	62.29	62.77	62.78	65.21	99.45											
CtrE	65.16	65.17	65.14	85.41	62.43	62.87	62.84	65.31	99.07	99.23										
CtrL	64.97	65.11	65.07	85.51	62.21	62.66	62.72	65.19	98.77	98.89	98.93									
CseC	44.86	44.58	44.77	44.42	44.61	44.66	44.52	44.77	44.74	44.7	44.53	44.67								
ElaC	44.62	44.62	44.47	44.52	44.66	44.43	44.39	44.7	44.59	44.55	44.5	44.7	52.86							
PacH	45.51	45.25	45.38	45.15	45.55	45.36	45.15	45.3	45.16	45.14	45.16	45.2	51.2	50.7						
PacU	45.55	45.27	45.39	45.18	45.48	45.28	45.15	45.41	45.19	45.21	45.16	45.25	51.38	50.63	99.17					
PamU	45.58	45.55	45.22	45.43	45.54	45.51	45.44	45.49	45.22	45.21	45.24	45.31	50.95	50.54	54.93	54.96				
PnaK	45.91	45.42	45.56	45.76	45.93	45.83	45.76	45.85	45.63	45.58	45.54	45.69	51.42	50.34	55.35	55.34	70.28			
SneZ	44.52	44.88	44.61	44.76	44.59	44.56	44.59	44.55	44.61	44.61	44.59	44.69	46.02	45.87	47.25	47.21	48.07	47.98		
Wch2	45.17	45.15	45.11	45.04	45.02	45.1	45.07	45.04	44.84	44.88	44.81	44.91	50.04	50.05	53.02	52.94	52.9	52.92	47.72	
WchW	45.53	45.56	45.54	45.4	45.32	45.61	45.59	45.5	45.24	45.24	45.13	45.26	50.18	50.29	53.09	53.11	53.28	53.19	47.99	99.2

Supplementary table 8| Mean and standard deviation of the parameters used to assess the phylogenetic information of each protein sequence for each of the 9 clusters

cluster	tree length	Align. length	RF[1]	SH[2] Reference tree	Mean SH[2] random topologies	SD SH[2] random topologies	Mean SH[2] random *Chlam.* Classic	SD SH[2] random Chlam. Classic	Mean SH[2] random *Chlam*.-like	SD SH[2] random *Chlam*.-like	Mean SH[2] 15 Familiy topo.	SD SH[2] 15 Familiy topo.
Mean value												
1	4.73	257.87	8.89	0.62	0.12	0.23	0.00	0.00	0.11	0.07	0.78	0.08
2	5.66	553.60	3.70	0.98	0.10	0.25	0.00	0.00	0.03	0.03	0.94	0.04
3	6.79	299.04	8.22	0.74	0.19	0.29	4. e-06	6.78e-06	0.22	0.09	0.96	0.02
4	2.15	143.05	9.60	0.80	0.19	0.26	0.12	0.03	0.10	0.08	0.90	0.10
5	4.59	205.33	8.22	0.84	0.12	0.22	0.00	0.00	0.08	0.04	0.83	0.23
6	5.68	544.65	6.61	0.91	0.10	0.26	2.10e-06	0.00	0.02	0.02	0.97	0.02
7	6.33	376.79	5.90	0.99	0.13	0.27	0.00	0.00	0.09	0.07	0.98	0.02
8	4.57	654.56	5.81	0.94	0.07	0.22	1.92e-06	0.00	0.00	0.00	0.88	0.13
9	6.28	335.78	6.89	0.92	0.15	0.27	0.00	0.00	0.19	0.13	0.95	0.05
SD value												
1	2.46	161.93	2.74	0.35	0.07	0.06	0.00	0.00	0.14	0.09	0.27	0.09
2	1.91	254.82	2.08	0.03	0.02	0.01	0.00	0.00	0.04	0.03	0.01	0.02
3	2.36	139.69	3.28	0.21	0.07	0.03	1.48e-05	2.09e-05	0.17	0.08	0.03	0.02
4	1.13	72.93	2.72	0.22	0.08	0.04	0.17	0.04	0.14	0.12	0.08	0.11
5	2.58	139.63	3.23	0.16	0.05	0.05	0.01	0.00	0.09	0.05	0.15	0.20
6	2.24	274.56	2.64	0.10	0.02	0.02	1.44e-05	0.00	0.01	0.02	0.01	0.02
7	2.18	155.21	3.08	0.02	0.03	0.02	0.00	0.00	0.04	0.04	0.01	0.02
8	2.38	322.02	2.03	0.06	0.02	0.02	1.38e-05	0.00	0.01	0.00	0.06	0.09
9	2.40	118.05	2.63	0.06	0.04	0.03	0.00	0.00	0.10	0.11	0.05	0.07

[1] **RF:** Robinson-Fould distance when a given tree topology is compared to the reference tree obtained with the concatenation of all 424 core protein sequences.
[2] **SH:** p-value of the SH-test.

Supplementary table 9| Pairwise protein sequence identity between *C. gallinacea* and 21 chlamydial species. Cells colored in blue present identity values higher than the defined threshold values (column T, see Fig. 5). Darker colors indicate higher identity. *C. gallinacea* is part of the *Chlamydia* genus, but clearly belongs to a new species, as reflected by the low identities of the RpoN, PepF, Adk and FtsK protein sequences with other *Chlamydia* species. HemL orthologue could not be found in the published draft sequences.

	gene	T	Accession	CtrA	CtrD	CtrE	CtrL	CpeE	CabS	Cps6	CcaG	CfeF	CmuN	CpnA	CpnK	ElaC	CseC	PacH	PacU	PamU	PnaK	SneZ	Wch2	WchW
fam.	16S	92.50%		95.31	95.31	95.44	95.44	95.58	98.18	98.38	98.05	97.53	96.36	95	95.26	90.08	89.66	88.47	88.83	89.42	88.38	89.35	89.47	89.47
	23S	91%		93.63	93.73	93.88	93.77	94.09	96.79	96.89	96.11	96.31	95.79	96.2	96.48	86.89	86.45	88.17	88.17	87.26	87.62	86.58	87.35	87.28
genus	DnaA	70%	WP_021828315.1	74.23	74.23	74.23	74.01	72.77	81.74	81.74	80.43	80.43	73.57	73.48	73.48	43.86	41.59	45.18	45.18	46.65	43.98	40.91	45.18	45.18
	FabI	78%	WP_021828235.1	84.56	84.56	84.56	85.23	83.56	87.63	84.62	87.96	87.29	84.90	83.61	83.61	64.78	58.53	61.13	60.80	62.67	64.33	67.11	66.78	66.78
	protein_325	57%	WP_021828758.1	65.48	65.48	65.24	65.24	65.72	79.48	79.76	79.29	79.53	65.48	71.53	71.29	42.65	41.75	40.00	40.24	43.65	40.52	39.71	42.11	42.11
	SucA	64%	WP_021828263.1	67.19	67.33	67.41	67.30	68.88	78.59	79.47	78.70	79.54	67.88	71.76	72.09	46.36	44.78	45.46	45.30	45.74	46.70	44.44	46.47	46.47
species	RpoN	96%	WP_021828751.1	47.73	47.73	47.49	47.26	56.09	66.75	68.41	67.46	67.62	51.32	57.04	56.93	34.39	33.90	33.01	32.76	31.65	32.05	30.52	30.94	30.94
	PepF	96%	WP_021828433.1	65.67	65.50	65.67	65.67	65.62	75.58	76.74	76.08	76.91	65.33	67.11	67.77	43.57	43.48	42.66	42.57	43.55	45.82	39.53	43.26	43.26
	Adk	95%	WP_021828371.1	51.43	51.43	51.43	51.43	54.81	63.46	64.53	64.42	65.38	51.90	56.25	56.73	42.11	41.55	43.54	43.54	40.78	42.08	37.98	41.06	41.06
	FtsK	98%	WP_021828651.1	72.90	72.81	72.77	72.77	73.12	81.19	82.35	82.43	82.43	73.14	71.85	72.11	50.66	50.98	52.18	52.18	52.65	50.91	50.61	48.03	50.07
	HemL	95%	WP_021828504.1																					

[1] **T :** Threshold values defined for species and genus delineation (see Fig. 5)

Supplementary table 10| Pairwise protein sequence identity between *C. avium* and 21 chlamydial species. Cells colored in blue present identity values higher than the defined threshold values (column T, see Fig. 5). Darker colors indicate higher identity. *C. avium* is part of the *Chlamydia* genus, but clearly belongs to a new species, as reflected by the low identities of the RpoN, PepF, Adk, FtsK and HemL protein sequences with other *Chlamydia* species.

	gene	T	Accession	CtrA	CtrD	CtrE	CtrL	CmuN	CabS	Cps6	CcaG	CfeF	CpeE	CpnA	CpnK	ElaC	CseC	PacH	PacU	PamU	PnaK	SneZ	Wch2	WchW
fam.	**16S**	92.50%		95.69	95.69	95.82	95.89	95.65	97.99	98.18	98.05	97.34	96.55	94.94	95.19	89.83	89.45	88.03	88.4	89.21	88.57	88.62	88.74	88.74
	23S	91%		93.7	93.81	93.95	93.84	94.07	97.06	97.23	96.35	96.55	96.07	96.54	96.72	86.27	86.55	88.13	88.13	87.35	87.13	86.96	87.47	87.4
genus	**DnaA**	70%	AHK63150.1	75.77	75.77	75.77	75.55	75.33	85.22	85.65	85.00	85.00	76.03	77.17	77.17	44.32	42.95	46.00	46.00	45.12	42.77	43.96	47.25	47.25
	FabI	78%	AHK63066.1	83.89	83.89	83.89	84.56	84.23	89.63	86.29	89.63	89.30	83.56	82.94	82.94	65.66	59.20	61.79	61.46	62.67	65.42	67.91	65.44	65.44
	protein_325	57%	AHK63629.1	63.44	63.44	63.44	63.21	64.62	78.59	79.76	79.20	78.82	66.90	69.34	69.10	42.58	41.67	40.81	41.05	44.60	41.87	40.10	39.05	39.05
	SucA	64%	AHK63096.1	67.41	68.19	67.29	67.07	68.63	78.96	79.73	79.51	78.74	68.58	70.31	70.76	44.97	44.59	44.21	44.26	45.88	45.77	44.19	44.85	44.85
species	**RpoN**	96%	AHK63622.1	47.74	47.74	47.51	47.51	50.24	68.35	69.78	68.57	68.03	54.65	56.87	56.70	35.04	34.62	34.05	34.05	33.01	32.13	31.34	31.73	31.73
	PepF	96%	AHK63281.1	66.01	65.68	65.84	65.84	66.83	77.14	78.13	77.30	78.45	67.27	67.93	68.42	44.71	43.87	42.60	43.00	43.52	45.45	41.48	42.07	42.07
	Adk	95%	AHK63212.1	49.77	49.77	49.77	49.77	50.70	61.79	63.05	66.98	63.68	55.19	54.29	54.29	39.38	39.19	41.41	41.41	37.44	40.47	37.84	42.38	42.38
	FtsK	98%	AHK63517.1	73.60	73.48	73.48	73.48	72.84	81.49	82.87	82.72	83.61	75.22	74.51	74.77	52.17	53.00	52.18	52.18	51.84	52.77	51.84	49.41	49.41
	HemL	95%	AHK63273.1	58.61	58.61	58.61	58.61	59.81	65.44	65.90	64.52	66.82	59.91	60.28	60.51	41.81	39.62	43.84	44.08	43.30	41.15	39.86	40.28	40.28

[1] **T :** Threshold values defined for species and genus delineation (see Fig. 5)

Supplementary table 11| Pairwise protein sequence identity between *C. ibidis* and 21 chlamydial species. Cells colored in blue present identity values higher than the defined threshold values (column T, see Fig. 5). Darker colors indicate higher identity. *C. ibidis* is part of the *Chlamydia* genus, but clearly belongs to a new species, as reflected by the low identities of the RpoN, PepF, Adk, FtsK and HemL protein sequences with other *Chlamydia* species.

	gene	T	accession	CtrA	CtrD	CtrE	CtrL	CmuN	CabS	Cps6	CcaG	CfeF	CpeE	CpnA	CpnK	ElaC	CseC	PacH	PacU	PamU	PnaK	SneZ	Wch2	WchW
fam.	**16S**	92.50%		95.52	95.52	95.59	95.4	95.98	97.08	97.27	97.08	96.56	96.82	96.63	96.89	89	89.96	89.15	89.18	88.62	88.81	88.02	89.12	89.12
	23S	91%		93.78	93.89	93.95	93.92	94.57	96.82	96.99	96.38	96.85	96.03	96.57	96.64	86.7	87.21	88.48	88.48	88.1	88.51	86.59	87.64	87.6
genus	**DnaA**	70%	WP_020370094.1	73.79	73.79	74.01	73.79	73.57	82.83	83.26	83.26	83.04	75.6	76.96	76.96	41.76	42.27	45.48	45.48	46.01	43.5	38.79	46.12	46.12
	FabI	78%	WP_020370277.1	84.56	84.56	84.56	85.23	86.58	82.94	81.27	86.62	86.29	84.56	85.62	85.62	67.46	60.07	66.22	65.89	68.26	65.89	69.02	67.45	67.45
	protein_325	57%	WP_020370681.1	68.11	68.11	67.63	67.87	66.43	75.89	76.36	76.42	75.83	65.01	70.59	70.59	43.23	42.76	45.61	45.37	42.96	42.38	36.6	41.15	41.15
	SucA	64%	WP_020370037.1	66.74	67.07	66.96	67.07	68.11	76.6	77.81	76.38	75.61	67.96	70.65	70.7	45.6	45.5	43.63	43.63	44.42	45.31	42.92	44.36	44.36
species	**RpoN**	96%	WP_020370688.1	46.73	45.67	45.95	46.12	48.4	60.24	60.95	59.05	59.95	50.6	55.61	55.4	34.15	37.97	35.38	35.14	33.09	32.49	31.23	31.8	31.8
	PepF	96%	WP_020370240.1	64.19	64.19	64.03	64.03	65.68	68.75	69.41	69.74	68.75	64.31	66.45	66.28	43.07	43.6	42.83	43.17	45.21	47.73	40.34	43.46	43.46
	Adk	95%	WP_020370158.1	55.14	55.14	55.14	55.14	54.67	50.94	50.98	52.61	53.77	50	52.58	53.52	39.91	37.74	37.91	37.91	36.15	38.03	38.5	38.5	38.5
	FtsK	98%	WP_020370586.1	72.14	71.84	72.14	72.14	72.26	78.23	78.9	78.11	78.36	73.88	72.08	72.34	50.66	52.47	49.93	49.93	50.53	51.92	50.68	49.67	49.8
	HemL	95%	WP_020370234.1	55.02	55.02	55.02	55.4	56.63	59.72	59.72	60.19	60.88	53.54	55.12	55.35	41.47	40.91	42.93	43.17	42.14	43.65	39.71	39.52	39.52

[1]**T :** Threshold values defined for species and genus delineation (see Fig. 5)

Supplementary table 12| Pairwise protein sequence identity between *C. suis* and 21 chlamydial species. Cells colored in blue present identity values higher than the defined threshold values (column T, see Fig. 5). Darker colors indicate higher identity. *C. suis* is part of the *Chlamydia* genus, but clearly belongs to a new species, as reflected by the low identities of the RpoN, PepF, Adk, FtsK and HemL protein sequences with other *Chlamydia* species.

	gene	T	accession	CtrA	CtrD	CtrE	CtrL	CmuN	CabS	Cps6	CcaG	CfeF	CpeE	CpnA	CpnK	ElaC	CseC	PacH	PacU	PamU	PnaK	SneZ	Wch2	WchW
fam.	**16S**	92.50%		97.98	97.98	98.05	98.18	98.31	95.3	95.5	95.56	95.87	95.34	94.24	94.5	90.09	89.62	88.64	88.87	88.52	89.22	88.05	88.45	88.45
	23S	91%		98.1	98.21	98.41	98.21	97.93	93.58	93.96	93.99	93.61	93.62	93.85	94.06	87	86.53	88.5	88.5	87.17	87.54	87.29	88.06	88.15
genus	**DnaA**	70%	ESN89490.1	93.64	93.64	94.08	93.64	92.76	79.96	80.62	79.3	78.63	72.85	74.67	74.67	43.94	43.15	46.79	46.79	44.7	44.28	43.45	44.47	44.47
	FabI	78%	ESN89713.1	93.6	93.6	93.6	94.28	96.97	83.84	80.81	85.19	84.51	84.51	85.52	85.52	65.66	60.14	63.64	63.3	65.66	64.98	66.67	65.99	65.66
	protein_325	57%	ESN89143.1	83.49	83.49	83.02	83.25	85.38	64.85	65.32	65.32	63.29	61.7	64.68	64.68	43.37	41.93	42.45	42.45	41.73	40.86	38.35	39.57	39.57
	SucA	64%	ESN89761.1	86.49	86.52	86.6	86.49	88.18	68.07	68.51	67.62	68.4	66.33	65.12	65.45	45.52	45.38	43.99	43.88	43.79	46.57	43.5	44.6	44.6
species	**RpoN**	96%	ESN89190.1	75.41	75.64	75.88	75.41	80.05	51.54	51.78	51.67	51.55	50.6	46.35	46.78	31.73	32.69	32.29	32.45	30.94	31.84	30.64	31.88	31.88
	PepF	96%	ESN89716.1	86.84	86.68	86.68	86.84	89.64	65.79	65.89	67.38	68.6	65.35	63.7	63.86	44.48	45.71	44.92	45.36	43.98	47.27	42.09	45.39	45.39
	Adk	95%	ESN89683.1	81.22	81.22	81.22	81.22	82.61	46.48	48.53	49.3	50.7	45.54	46.48	46.48	40.47	39.44	42.58	42.58	40.19	42.72	39.91	40.1	40.1
	FtsK	98%	ESN89046.1	92.23	92.23	92.36	92.48	92.24	73.72	75.78	75.03	74.23	69.27	70.09	69.83	50.68	53.58	50.94	50.94	50.13	50.53	49.86	49.8	49.4
	HemL	95%	ESN89503.1	76.78	76.78	76.78	76.3	80.09	56.53	56.53	57.24	58.19	56.53	56.46	56.46	38.7	41.77	43.03	42.79	45.83	45.41	40.05	41.02	41.02

[1]**T :** Threshold values defined for species and genus delineation (see Fig. 5)

Supplementary table 13| Pairwise protein sequence identity between *Neochlamydia S13* and 21 chlamydial species. Cells colored in blue present identity values higher than the defined threshold values (column T, see Fig. 5). Darker colors indicate higher identity. *Neochlamydia* presents conflicting 23S identities with members of the *Parachlamydiaceae* family. In such case, the majority prevails. This strain forms a new *Parachlamydiaceae* genus, as reflected by the low identities of the DnaA and protein_325. SucA orthologue could not be found in the published draft sequences.

	gene	T	CtrA	CtrD	CtrE	CtrL	CmuN	CabS	CcaG	CfeF	CpeE	CpnA	CpnK	Cps6	ElaC	CseC	PacH	PacU	PamU	PnaK	SneZ	Wch2	WchW
fam.	16S	92.50%	88.68	88.56	88.46	88.52	89.21	89.85	89.91	89.96	89.19	89.64	89.74	90.14	91.06	90.76	93.21	93.42	93.21	92.56	89.99	90.8	90.8
	23S	91%	88.11	88.34	88.09	88.12	87.97	86.64	86.87	86.82	86.28	86.71	86.84	87.06	89.08	90.64	91.4	91.4	90.6	90.85	89.11	87.68	87.62
genus	DnaA	70%	44.29	44.29	44.29	44.29	43.61	44.59	44.39	43.95	45.52	45.62	45.84	43.95	50.9	49.2	57.96	57.96	60.27	59.76	47.89	55.38	55.38
	FabI	78%	65.53	65.53	65.53	65.53	64.09	65.1	64.21	63.88	68.47	69.05	69.05	64.09	68.47	63.88	79.33	79.67	74.92	75.67	70.33	73.49	73.49
	protein_325	57%	44.34	44.34	44.1	44.58	43.99	44.05	44.47	44.93	43.24	44.79	44.55	44.26	45.41	44.34	51.9	51.9	47.74	49.05	40.48	47.75	47.75
	SucA	64%																					
species	RpoN	96%	37.84	37.47	37.8	37.47	33.82	35.99	36.23	35.27	34.38	34.37	34.37	35.61	46.58	41.36	56.24	56.03	49.17	50.52	38.76	51.57	51.57
	PepF	96%	44.34	44.34	44.1	44.58	43.99	44.05	44.47	44.93	43.24	44.79	44.55	44.26	45.41	44.34	51.9	51.9	47.74	49.05	40.48	47.75	47.75
	Adk	95%	38.5	38.5	38.5	38.5	40.85	43.9	45.24	41.04	44.93	40.19	40.67	44.9	47.25	53.95	55.45	55.45	50.46	52.75	41.01	54.93	56.37
	FtsK	98%	65.53	65.53	65.53	65.53	64.09	65.1	64.21	63.88	68.47	69.05	69.05	64.09	68.47	63.88	79.33	79.67	74.92	75.67	70.33	73.49	73.49
	HemL	95%	46.06	45.73	45.9	45.73	47.07	45.15	46.66	46.49	47.42	45.92	45.76	45.82	56.37	57.36	58.97	61.4	58.61	65.03	53.43	62.4	62.4

[1]**T :** Threshold values defined for species and genus delineation (see Fig. 5)

Supplementary figure 1 | Visualization of protein clusters by principal component analysis

Principal component analysis (PCA) of the criteria used to evaluate the phylogenetic information and congruence of individual protein sequences. The 9 clusters are highlighted in different colors. Most proteins share similar information and therefore cluster together. Cluster 2 (in blue) present the best overall characteristics and was selected for subsequent analysis. Most proteins diverging from the core are those rejecting the reference topology (Table 1) and presenting HGT events (arrows). One outlier, the 50S ribosomal protein L16, was removed from the PCA visualization. It was the most uninformative protein alignment as evaluated by SH-tests with semi-random topologies.

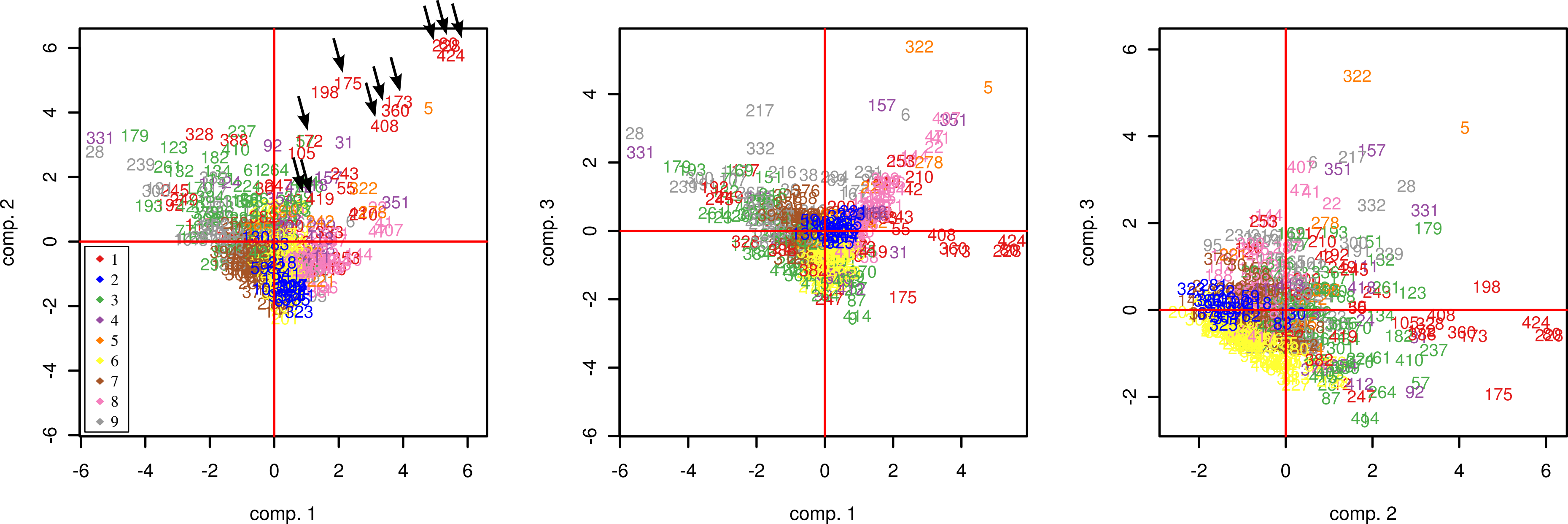
Supplementary figure 2 | Evaluation of the optimal number of concatenated genes needed to reconstruct a robust phylogeny of the *Chlamydiales* order.

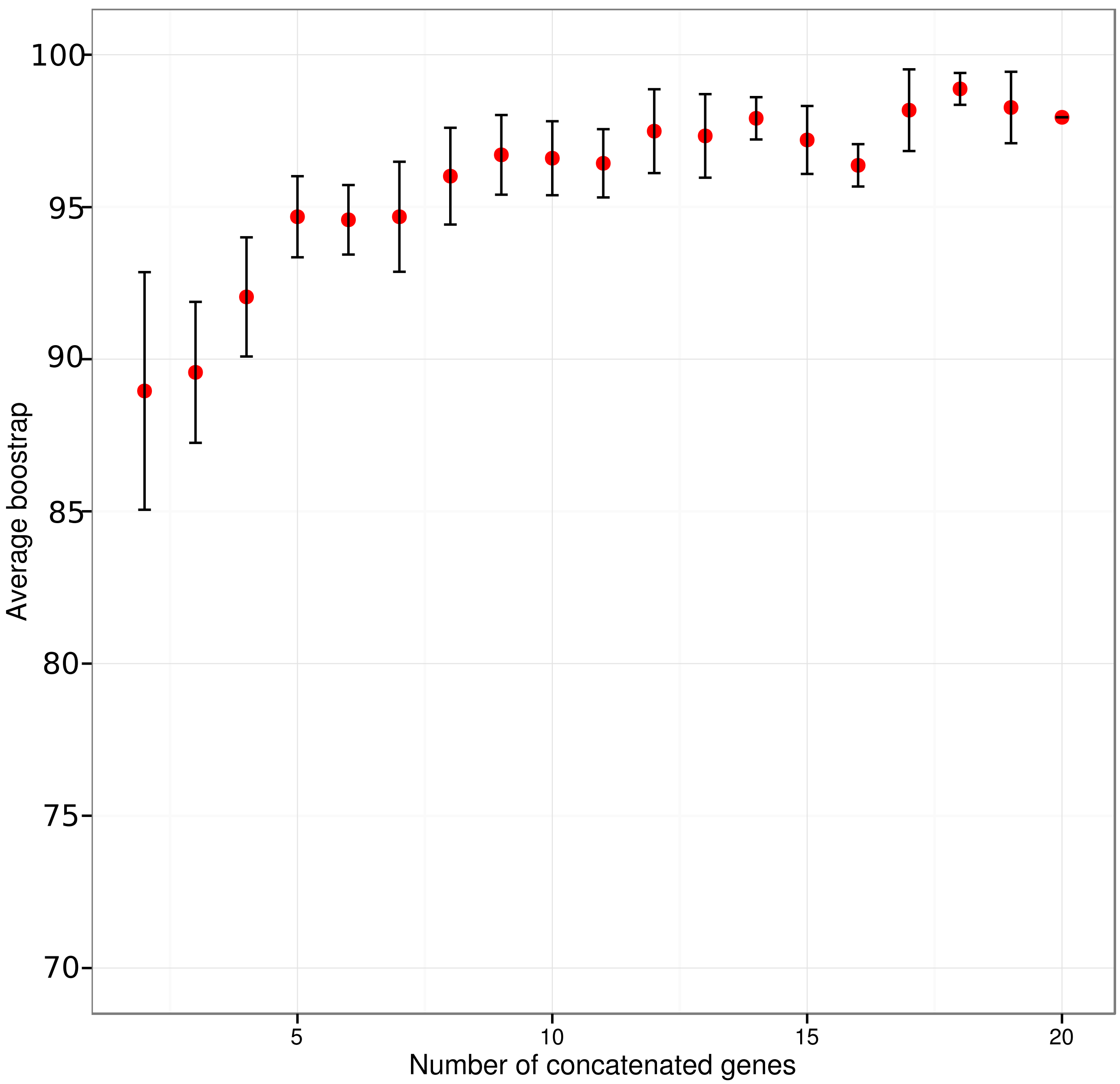
Alignments were randomly sampled 5 times with replacement among the best 20 markers. Errors bars reflect the variation between the 5 samples, but concatenations of increasingly higher number of alignments tend to include the same alignments. The 20th is a concatenation of all 20 alignments. Please note that using ≥ 8 protein sequences provide an average boostrap value > 95%.

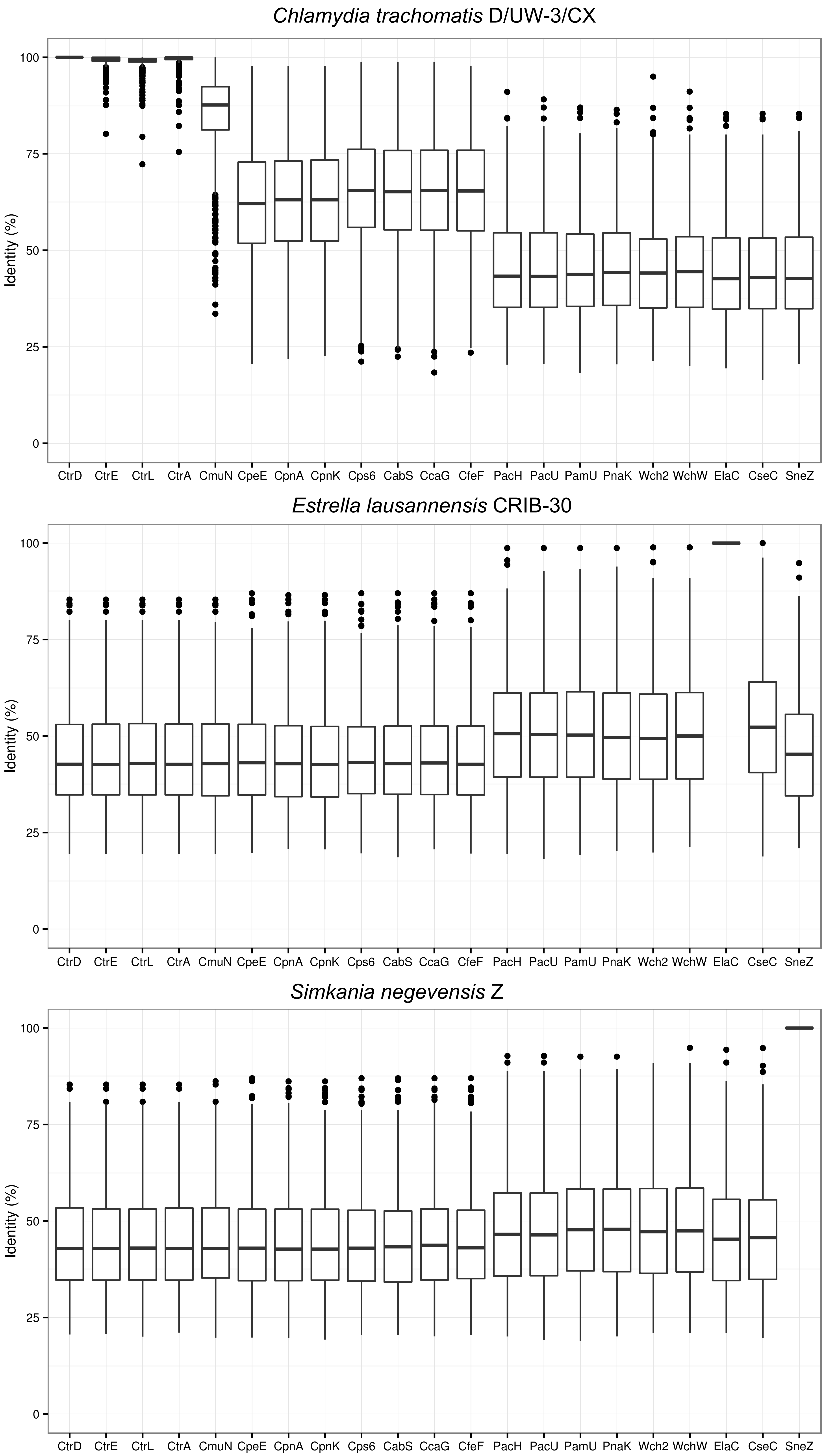
Supplementary figure 3 | Boxplot of identity of reciprocal BLASTP

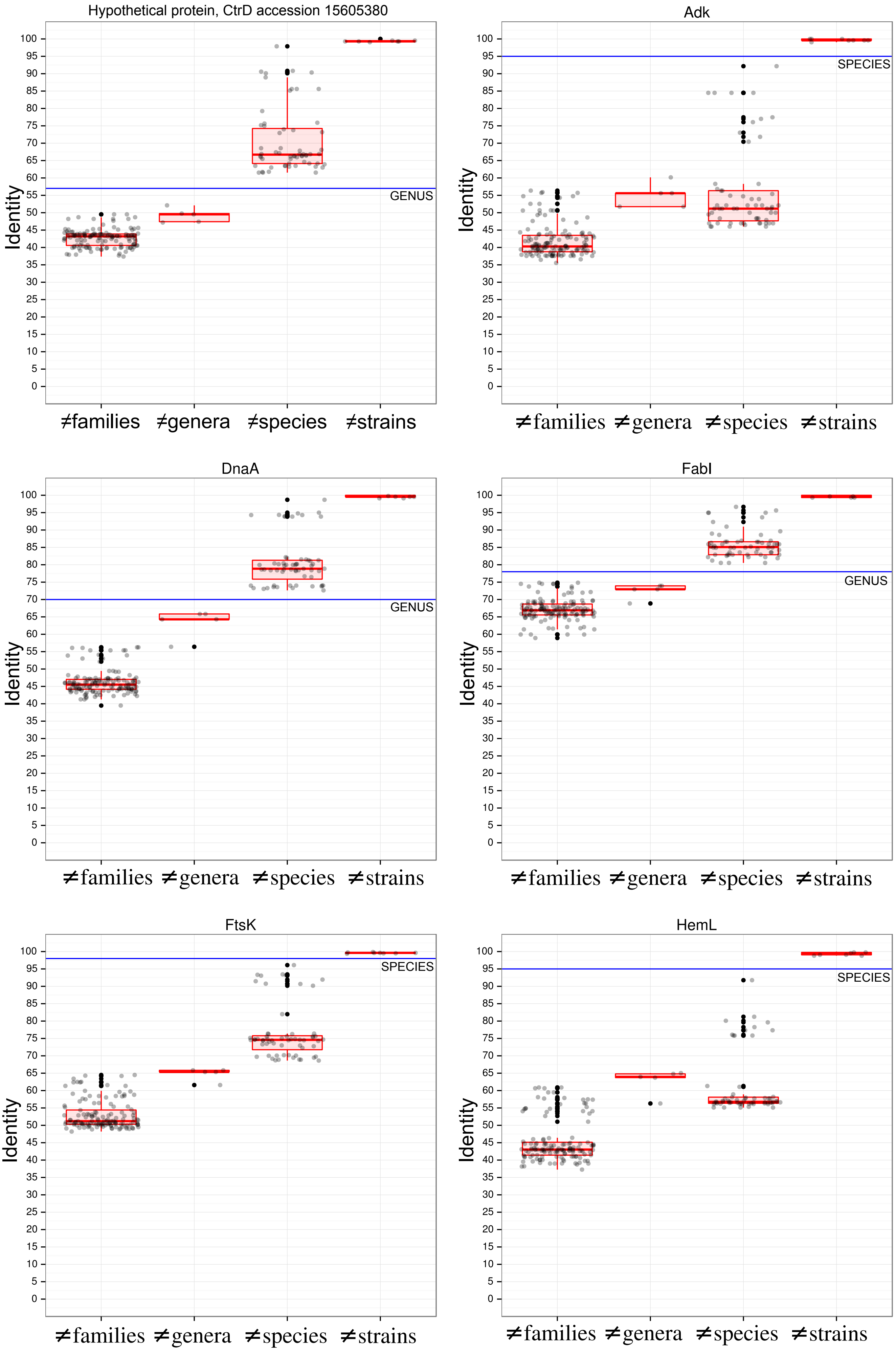
Supplementary figure 4 | Pairwise identity of selected markers

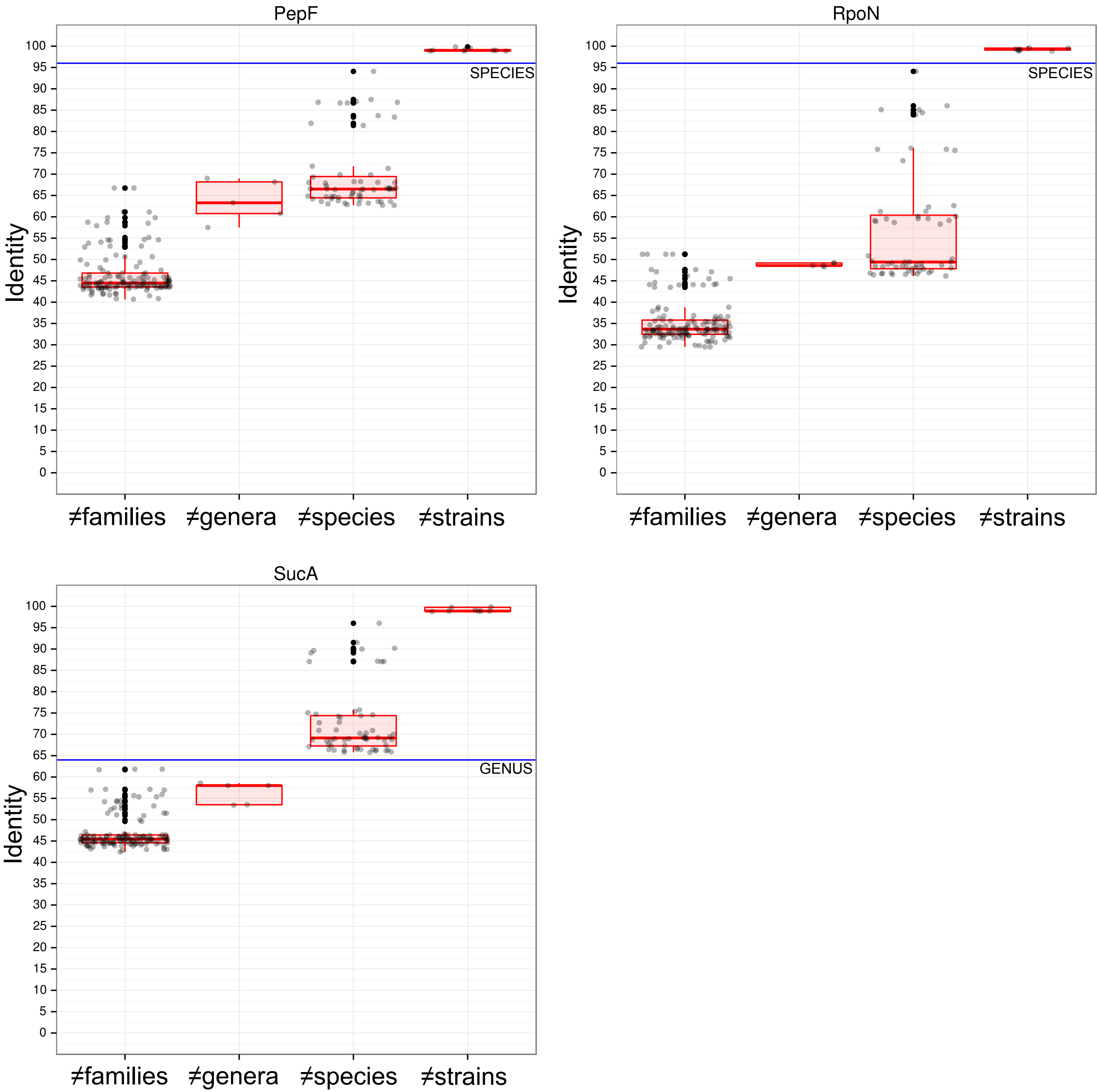
The conservation of proteins selected for the classification of new chlamydial isolates is represented here as a boxplot of pairwise identity between strains belonging to different taxonomical level. Blue lines indicate the classification cutoff value selected for each protein to classify *Chlamydiales* at the species and genus levels (Figure 5).


comp. 1
comp. 2
comp. 1
comp. 3
comp. 2
comp. 3
-6
-4
-2
0
2
4
6
1
2
3
4
5
6
7
8
9
280
424
175
198
173
360
5
408
331
179
28
328
388
237
123
182
410
92
172
105
31
243
55
322
351
407
323
217
332
157
6
253
278
210
42
65
175
414
87
407
351
47
22
144
198
123
261
237
410
57
92
264
412
247


100
95
90
85
80
75
70
Average boostrap
5
10
15
20
Number of concatenated genes


Chlamydia trachomatis D/UW-3/CX
Estrella lausannensis CRIB-30
Simkania negevensis Z
Identity (%)
0
25
50
75
100
CtrD
CtrE
CtrL
CtrA
CmuN
CpeE
CpnA
CpnK
Cps6
CabS
CcaG
CfeF
PacH
PacU
PamU
PnaK
Wch2
WchW
ElaC
CseC
SneZ


Hypothetical protein, CtrD accession 15605380
Adk
DnaA
FabI
FtsK
HemL
Identity
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
GENUS
SPECIES
≠families
≠genera
≠species
≠strains


PepF
RpoN
SucA
Identity
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
SPECIES
GENUS
≠families
≠genera
≠species
≠strains