Taxogenomics of the order *Chlamydiales*

Trestan Pillonel,1,2 Claire Bertelli,1,2 Nicolas Salamin2,3 and Gilbert Greub1

1Center for Research on Intracellular Bacteria, Institute of Microbiology, University Hospital Center and University of Lausanne, Lausanne, Switzerland
2SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland
3Department of Ecology and Evolution, Biophore, University of Lausanne, Lausanne, Switzerland

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 order *Chlamydiales* mainly relies on DNA- or protein-based phylogenetic reconstructions because these organisms 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 reconstruct 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 reconstruction of a robust tree close to the reference species tree. We recommend using these protein sequences to precisely classify newly discovered isolates at the family, genus and species levels.

**INTRODUCTION**

Phylogenetic reconstruction based on 16S rRNA gene sequences is a widely used approach to infer relationships between bacteria (Fox et al., 1980). Nevertheless, the high conservation of rRNA reduces its discriminatory power and makes it insufficient to distinguish closely related bacterial species (Rossello-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 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 trees because of frequent discrepancies between gene trees. For example, serovars of *Chlamydia trachomatis* were classified on the basis of 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 phylum Chlamydiae was long restricted to one group of closely related obligate intracellular bacteria classified in a single family, *Chlamydiaceae*. During the last two decades, new organisms resembling members of the family *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 members of the family *Chlamydiaceae* and all belong to the order *Chlamydiales*. These novel chlamydiae were isolated from different geographical areas, indicating a widespread occurrence in nature. This is also emphasized by the diversity of organisms representing the order *Chlamydiales* observed in metagenomics samples (Lagkouvardos et al., 2014).

In 1999, Everett et al. proposed to use 16S and 23S rRNA gene sequence cut-offs of 97, 95 and 90% identity to classify members of the order *Chlamydiales* at the species, genus and family levels, respectively (Everett et al., 1999). Controversies arose because Everett et al. (1999) proposed to split the family *Chlamydiaceae* into two genera: *Chlamydia* and *Chlamydophila*. This split was disputed since it was not consistently supported by significant biological differences and 16S rRNA gene sequence differences were limited.

**Abbreviations:** ANI, average nucleotide identity; BBH, best BLAST hit; HGT, horizontal gene transfer; ISTC, International Subcommittee on the Taxonomy of the *Chlamydiae*; SH-test, Shimodaira-Hasegawa test.

Thirteen supplementary tables and four supplementary figures are available with the online Supplementary Material.
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 (Pen 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 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 reconstructed using the Extended Majority Rule criterion from the program SumTrees version 3.3.1 from Dendropy 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 reconstructed using SumTrees (Sukumaran & Holder, 2010). Neighbour-joining trees were reconstructed using the bioNJ algorithm with Seaview (Gouy et al., 2010).

**Convergence and strength of the phylogenetic signal.** Tree topologies were first compared using the Robinson–Foulds distance (Robinson & Foulds, 1981) computed using the package phangorn with the LG model (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 gene 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 the 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 evaluation of 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 order Chlamydiales not belonging to the family Chlamydiaceae), (ii) 100 topologies randomizing only the branching between members of the family Chlamydiaceae, and (iii) all 15 branching possibilities of the five families of the order Chlamydiales. 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–Foulds distance), the congruence with this reference topology (SH-test P value) and the ability to reject semi-random topologies (mean 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 order Chlamydiales. **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 nine proteins were identified in newly sequenced genomes by retrieving the best
BLASTP hits of each of the nine proteins from the 21 strains included in this analysis. For each of the nine proteins, we confirmed that the best hit of a given protein was 21 times the same hit.

**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 rRNA Predictor (http://www.vicbioinformatics.com/). The average nucleotide identity (ANI) between chlamydial genomes was computed using MUMmer (Kurtz et al., 2004), as described by Richter & Rossello-Mora (2009).

**RESULTS**

Current criteria do not match the existing classification of the order *Chlamydiales*

16S and 23S rRNA gene sequences are routinely used for bacterial species identification and classification. For members of the order *Chlamydiales*, cut-offs of 97, 95 and 90 % 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 gene 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) and 98.09 % 23S rRNA identity (Table S4), while their whole genomes exhibit an ANI of 83.89 % (Table S5).

Contrary to rRNA analysis, an ANI cut-off 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 threefold larger than strains from the family *Chlamydiaceae* (Bertelli et al., 2010; Collingro et al., 2011). Nevertheless, members of the family *Chlamydiaceae*, most of which possess less than 1000 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 the species *Parachlamydia acanthamoebae* and *Waddlia chondrophila* share only between 86 and 90 % of their predicted proteins. Among the *Chlamydia*-related families, only the genus ‘*Candidatus 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 genus *Chlamydia*.

The current classification of a given strain at the species or family level can hardly be directly linked to the average amino acid identity of orthologous proteins (Table S7). The family *Chlamydiaceae* and *Chlamydia*-related families are clearly separated, presenting between 44.39 and 45.93 average percentage identity. Interestingly, *Simkania negevensis* Z^T^ presents a similarly low average amino acid identity with

---

**Fig. 1.** rRNA identity based on pairwise global alignments. (a) 16S and (b) 23S rRNA identity. Dotted lines indicate 16S identity thresholds proposed by Everett et al. (1999). Solid lines indicate new proposed thresholds of, respectively, 92.5 and 91 % for family.
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. S1). 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 Parachlamydia acanthamoebae (different families) exhibit 52.8 % average identity. Species from the genus Chlamydia exhibit average identities ranging from 62.2 % (Chlamydia trachomatis – Chlamydia pecorum) to 94.4 % (‘Chlamydia abortus’–Chlamydia 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.

**Core genome and phylogeny of the order Chlamydiales**

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 species tree for the order Chlamydiales, 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 ISTC (Greub, 2010a, b). The former Chlamydophila subgroup clearly clusters separately from Chlamydia trachomatis and Chlamydia muridarum. Significant variations only occur between members of the former Chlamydophila subgroup. These variations involve the closely related species Chlamydia psittaci, ‘Chlamydia caviae’ and ‘Chlamydia abortus’ and the basal branching of Chlamydia pecorum in the neighbour-joining tree based on average protein identities.

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

**Individual gene trees differ from the species tree**

Each individual gene tree was compared with the reference tree topology (Fig. 2c). Only seven topologies out of 424 were identical to the reference (without considering the branching pattern of Chlamydia trachomatis strains; Fig. 3a). Nevertheless, only eight 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 genus ‘Ca. Protochlamydia’ 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).

**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 family Chlamydiaceae 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 41 alignments rejected random branching of the Chlamydia-related species with an average P value <0.001. Those proteins included proteins widely used for phylogenetic purposes (e.g. RpoB, RpoC) as well as some of the eight proteins presenting particular evolutionary histories (Table 1). A total of 203 alignments presented an average P value >0.05 (Fig. 4 topologies 101–200).

Overall, the less discriminating alignments (with P value >0.05) were 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 order Chlamydiales, the support of all 15 possible branchings of the five families of the order Chlamydiales 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 four alignments present average P values below 0.05: Tgt, HemH, GlgB and ArOβ, and they all reject the reference topology as well (Table 1).

**Selection of optimal markers for the classification of chlamydial isolates**

In order to identify the most phylogenetically informative alignments, the alignments were classified in nine clusters according to two criteria (Table S8, Fig. S2). First, the congruence with the reference tree topology was evaluated by the Robinson–Foulds distance and the P value of the
**Fig. 2.** Phylogenetic trees of the order *Chlamydiales* based on 424 core proteins. (a) Midpoint-rooted tree reconstructed by neighbour-joining based on average identity of the genes shared between pairs of genomes (see Table S7). (b) Consensus tree based on the 424 individual core protein phylogenies. (c) The bar represent the mean number of amino acid substitution per site. Midpoint-rooted maximum-likelihood tree based on concatenation of the 424 core proteins. Bootstrap support values are indicated when inferior to 100. 1, family *Chlamydiaceae*; 2, family *Simkaniaceae*, 3, two genera of the family ‘Criblamydiaceae’; 4, family *Waddliaceae*; 5, two genera of the family *Parachlamydiaceae.*
Fig. 3. Congruence of phylogeny of the order Chlamydiales. (a) Robinson–Foulds distance of individual gene trees compared with the reference tree topology. A distance of 0 indicates identical topologies. (b) SH-test P value as a function of tree length. The positions of the 38 ribosomal proteins are indicated as stars. Black dots indicate the positions of RpoB, RpoC, GyrB, RecA and El-Tu, five proteins frequently used for phylogenetic purposes. (c) Conflicting phylogeny of ribonucleotide-diphosphate reductase subunit beta (NrdB). The two species of the genus ‘Ca. Protochlamydia’ (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, S. negevensis ZT, ‘Criblamydia sequanensis’ CRIB-18, ‘E. lausannensis’ CRIB-30, W. chondrophila WSU 86-1044T, ‘Protochlamydia naegleriophila’ KNic and Parachlamydia acanthamoebae Hall’s coccus, and redundancy was removed before phylogenetic reconstruction.
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–Foulds value of 3.7 on average) and low SH-test $P$ value for the rejection of semi-random topologies ($<0.001$ for Chlamydiales 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 five alignments already resulted in trees with average bootstrap of value 94.7 ± 1.33 % (Fig. S3).

Fig. 5 proposes a new classification scheme for the order Chlamydiales. Identity cut-offs 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.

### Classification of five 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 gallinarum’ 08-1274/3 and Neochlamydia sp. S13 (see Tables S9–S13). The classification of the first three strains as novel species of the genus Chlamydia was confirmed without any conflicting results for all nine proteins. The orthologue of HemL could not be identified in published sequences of ‘Chlamydia gallinarum’, which did not prevent us from confirming the classification of this strain as a novel species of the genus Chlamydia. Similarly, the orthologue of SucA could not be identified in Neochlamydia sp. S13. Conflicting percentage identity of the 23S rRNA gene can be observed between the genus Neochlamydia and the two genera Parachlamydia and ‘Ca. Protochlamydia’ (Table S13). In addition, FabI presents a percentage identity higher than the cut-off of 78 % with the genus Parachlamydia, in contrast to DnaA and protein_325. Altogether, these results still suggest that Neochlamydia sp. S13 represents a new genus of the family Parachlamydiaceae, an affiliation which is congruent with current taxonomy.

### 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) to minimize potential biases due to HGTs, conserved genes or genes with high mutation rate leading to saturation. Thus, this work focused on the identification of a set of protein sequences presenting a

---

**Table 1.** Table alignments presenting strong evidence of conflicting phylogenetic signal with the reference tree

<table>
<thead>
<tr>
<th>Gene Accession</th>
<th>Orthogroup ID</th>
<th>Chlamydia traditionis D</th>
<th>Tree length</th>
<th>Align. length</th>
<th>RF* SH</th>
<th>Mean SH random topologies</th>
<th>Mean SH reference tree</th>
<th>SH identification of a set of protein sequences presenting...</th>
<th>SH, $P$ value of the SH-test.</th>
</tr>
</thead>
<tbody>
<tr>
<td>15604421</td>
<td>87</td>
<td>11.08</td>
<td>2.50</td>
<td>10</td>
<td>0.1863</td>
<td>0.0008</td>
<td>0.2304</td>
<td>0.0987</td>
<td>0.0000</td>
</tr>
<tr>
<td>15604473</td>
<td>80</td>
<td>12.28</td>
<td>4.06</td>
<td>10</td>
<td>0.0120</td>
<td>0.0182</td>
<td>0.0440</td>
<td>0.0002</td>
<td>0.0000</td>
</tr>
<tr>
<td>15605009</td>
<td>173</td>
<td>9.70</td>
<td>4.23</td>
<td>10</td>
<td>0.0120</td>
<td>0.0182</td>
<td>0.0440</td>
<td>0.0002</td>
<td>0.0000</td>
</tr>
<tr>
<td>15605100</td>
<td>175</td>
<td>12.28</td>
<td>4.06</td>
<td>10</td>
<td>0.0120</td>
<td>0.0182</td>
<td>0.0440</td>
<td>0.0002</td>
<td>0.0000</td>
</tr>
<tr>
<td>15605538</td>
<td>108</td>
<td>5.45</td>
<td>5.76</td>
<td>8</td>
<td>0.0000</td>
<td>0.0126</td>
<td>0.0440</td>
<td>0.0002</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

*RF, Robinson–Foulds distance when a given tree topology is compared with the reference tree obtained with the concatenation of all 424 core protein sequences.

**SH, value of the SH-test.$^*$**
strong phylogenetic signal allowing an accurate classification of new chlamydial isolates. We identified a set of 20 protein sequences that enabled us 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 order Chlamydiales and to determine the taxonomic affiliation of a new strain at the species, genus and family level.

Chlamydial classification

Chlamydial phylogeny has been a topic of intense debate during the last decades, focusing mainly on the classification of the family Chlamydiaceae into one or two genera and the use of 16S rRNA gene sequences for chlamydial classification (Everett et al., 1999; Schachter et al., 2001; Stephens et al., 2009; Voigt et al., 2012). The analysis of 16S rRNA gene 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 used more and more for studying 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 & Rossello-Móra, 2009). This criterion effectively reflects the recognized chlamydial taxonomy at the 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 questions the relevance of the current classification. Indeed, the order Chlamydiaceae presents three highly diverging clades (average protein identities <50%): the family Chlamydiaceae, the family Simkaniaceae and the grouping of the families Waddliaceae, Parachlamydiaceae and ‘Criblamydiaceae’ (Fig. 2a, Table S7). In addition, ‘Criblamydiaceae’
sequanensis’ and ‘E. lausannensis’ (same family) exhibit an average identity that is lower than that of W. chondrophila and Parachlamydia acanthamoebae (different families). Nevertheless, as signal saturation of protein sequences can be important with such distantly related organisms, simple metrics such as the API are probably not the best approach to distinguish inter-genus from inter-family relationships.

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 family ‘Criblamydiaceae’, and only four genomes from Chlamydia-related species. The present analysis included nine genomes of Chlamydia-related bacteria including two different genera within the family ‘Criblamydiaceae’. Moreover, the stringent criterion used in the present work to define orthology as well as the inclusion of five draft genomes also explains such a difference.

A reference phylogeny of the order Chlamydiales was reconstructed 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 reconstruction of robust phylogenies. However, these proteins still reflect the evolutionary history of the species and are useful for reconstructing robust phylogenies when concatenated.

Different gene trees but little evidence 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,
Table 2. The 20 most phylogenetically informative proteins of the core genome of the order Chlamydiales

<table>
<thead>
<tr>
<th>Gene</th>
<th>Accession</th>
<th>Orthogroup ID</th>
<th>Tree length</th>
<th>Align. length</th>
<th>RF*</th>
<th>SH† reference tree</th>
<th>Mean SH† random topologies</th>
<th>SD SH† random topologies</th>
<th>Mean SH† Chlamydia classic</th>
<th>Mean SH† Chlamydia-like</th>
<th>Mean SH† Chlamydia-like</th>
<th>Mean SH† 15 family topologies</th>
<th>SD SH† 15 family topologies</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>sucA</td>
<td>15604773</td>
<td>29</td>
<td>6.36</td>
<td>996</td>
<td>4</td>
<td>0.99</td>
<td>0.08</td>
<td>0.24</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.94</td>
<td>0.05</td>
<td>2-Oxoglutarate dehydrogenase subunit E1</td>
</tr>
<tr>
<td>tyrS</td>
<td>15604781</td>
<td>32</td>
<td>5.05</td>
<td>446</td>
<td>2</td>
<td>0.95</td>
<td>0.10</td>
<td>0.24</td>
<td>0.00</td>
<td>0.06</td>
<td>0.03</td>
<td>0.95</td>
<td>0.05</td>
<td>Tyrosyl-tRNA synthase</td>
</tr>
<tr>
<td>fabI</td>
<td>15604823</td>
<td>59</td>
<td>2.47</td>
<td>325</td>
<td>4</td>
<td>0.97</td>
<td>0.13</td>
<td>0.25</td>
<td>0.00</td>
<td>0.12</td>
<td>0.07</td>
<td>0.94</td>
<td>0.02</td>
<td>Enolyl-ACP reductase</td>
</tr>
<tr>
<td>pepF</td>
<td>15604831</td>
<td>62</td>
<td>5.32</td>
<td>655</td>
<td>6</td>
<td>0.99</td>
<td>0.10</td>
<td>0.25</td>
<td>0.00</td>
<td>0.04</td>
<td>0.04</td>
<td>0.93</td>
<td>0.02</td>
<td>Oligoendopeptidase F</td>
</tr>
<tr>
<td>adk</td>
<td>15604847</td>
<td>67</td>
<td>8.08</td>
<td>289</td>
<td>2</td>
<td>1.00</td>
<td>0.09</td>
<td>0.25</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.96</td>
<td>0.01</td>
<td>Adenylate kinase</td>
</tr>
<tr>
<td>hemL</td>
<td>15604930</td>
<td>83</td>
<td>6.47</td>
<td>496</td>
<td>8</td>
<td>0.90</td>
<td>0.09</td>
<td>0.25</td>
<td>0.00</td>
<td>0.04</td>
<td>0.05</td>
<td>0.92</td>
<td>0.04</td>
<td>Glutamate-1-semialdehyde aminotransferase</td>
</tr>
<tr>
<td>fabG</td>
<td>15604958</td>
<td>93</td>
<td>4.52</td>
<td>254</td>
<td>2</td>
<td>1.00</td>
<td>0.13</td>
<td>0.28</td>
<td>0.00</td>
<td>0.05</td>
<td>0.02</td>
<td>0.95</td>
<td>0.01</td>
<td>3-Ketoacyl-ACP reductase</td>
</tr>
<tr>
<td>dnaA</td>
<td>15604971</td>
<td>103</td>
<td>5.45</td>
<td>494</td>
<td>2</td>
<td>0.99</td>
<td>0.09</td>
<td>0.24</td>
<td>0.00</td>
<td>0.05</td>
<td>0.07</td>
<td>0.95</td>
<td>0.00</td>
<td>Chromosomal replication initiation protein</td>
</tr>
<tr>
<td>clpC</td>
<td>15605007</td>
<td>126</td>
<td>2.31</td>
<td>902</td>
<td>4</td>
<td>1.00</td>
<td>0.08</td>
<td>0.24</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.93</td>
<td>0.05</td>
<td>ClpC protease ATPase</td>
</tr>
<tr>
<td>dut</td>
<td>15605013</td>
<td>130</td>
<td>3.99</td>
<td>156</td>
<td>8</td>
<td>0.93</td>
<td>0.15</td>
<td>0.27</td>
<td>0.00</td>
<td>0.11</td>
<td>0.10</td>
<td>0.92</td>
<td>0.02</td>
<td>Deoxyuridine 5′-triphosphate nucleotidohydrolase</td>
</tr>
<tr>
<td>lpxK</td>
<td>15605127</td>
<td>190</td>
<td>9.66</td>
<td>453</td>
<td>4</td>
<td>1.00</td>
<td>0.13</td>
<td>0.27</td>
<td>0.00</td>
<td>0.06</td>
<td>0.04</td>
<td>0.94</td>
<td>0.03</td>
<td>Tetraacyldisaccharide 4′-kinase</td>
</tr>
<tr>
<td>argS</td>
<td>15605181</td>
<td>218</td>
<td>4.76</td>
<td>594</td>
<td>6</td>
<td>0.99</td>
<td>0.10</td>
<td>0.25</td>
<td>0.00</td>
<td>0.04</td>
<td>0.05</td>
<td>0.91</td>
<td>0.05</td>
<td>Arginyl-tRNA synthase</td>
</tr>
<tr>
<td>gspF</td>
<td>15605299</td>
<td>281</td>
<td>6.57</td>
<td>401</td>
<td>2</td>
<td>1.00</td>
<td>0.09</td>
<td>0.24</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.93</td>
<td>0.08</td>
<td>General secretion pathway protein F</td>
</tr>
<tr>
<td>rpoN</td>
<td>15605340</td>
<td>304</td>
<td>8.29</td>
<td>527</td>
<td>4</td>
<td>0.98</td>
<td>0.08</td>
<td>0.24</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.94</td>
<td>0.04</td>
<td>RNA polymerase factor sigma-54</td>
</tr>
<tr>
<td>greA</td>
<td>15605367</td>
<td>317</td>
<td>5.46</td>
<td>741</td>
<td>4</td>
<td>0.96</td>
<td>0.07</td>
<td>0.24</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.95</td>
<td>0.04</td>
<td>Transcript cleavage factor</td>
</tr>
<tr>
<td>topA</td>
<td>15605375</td>
<td>323</td>
<td>3.60</td>
<td>911</td>
<td>0</td>
<td>1.00</td>
<td>0.07</td>
<td>0.24</td>
<td>0.00</td>
<td>4 × 10⁻⁴</td>
<td>0.00</td>
<td>0.95</td>
<td>0.06</td>
<td>DNA topoisomerase I/ SWI</td>
</tr>
<tr>
<td>–</td>
<td>15605380</td>
<td>325</td>
<td>7.07</td>
<td>455</td>
<td>4</td>
<td>0.98</td>
<td>0.09</td>
<td>0.25</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.95</td>
<td>0.02</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>–</td>
<td>15605424</td>
<td>341</td>
<td>4.89</td>
<td>243</td>
<td>4</td>
<td>0.92</td>
<td>0.11</td>
<td>0.25</td>
<td>0.00</td>
<td>0.04</td>
<td>0.02</td>
<td>0.95</td>
<td>0.05</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>ftsK</td>
<td>15605472</td>
<td>364</td>
<td>7.75</td>
<td>958</td>
<td>2</td>
<td>1.00</td>
<td>0.07</td>
<td>0.24</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.94</td>
<td>0.04</td>
<td>Cell division protein FtsK</td>
</tr>
<tr>
<td>priA</td>
<td>15605511</td>
<td>385</td>
<td>5.18</td>
<td>776</td>
<td>2</td>
<td>1.00</td>
<td>0.09</td>
<td>0.26</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.95</td>
<td>0.05</td>
<td>Primosome assembly protein PriA</td>
</tr>
</tbody>
</table>

*RF, Robinson–Foulds distance when a given tree topology is compared with the reference tree obtained with the concatenation of all 424 core protein sequences.
†SH, P value of the SH-test.
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 determination of 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 the 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 have been disputed (Bapteste et al., 2004). It does not seem possible to assume that core genes are free of HGT events and effectively share a common evolutionary history because of the difficulty of detecting HGT when considering proteins with weak phylogenetic signal (Bapteste et al., 2005; Susko et al., 2006).

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 phylogeny of the order Chlamydiales. On the one hand, the classification of the family Chlamydiaceae 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 five different families belonging to the order Chlamydiales 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, rendering 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).

New chlamydiial 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 eight of these selected sequences should be used to reconstruct robust trees with an average bootstrap above 95% (Fig. S3). In addition to the reconstruction of robust phylogenetic trees, we propose a new classification scheme based on both 16S–23S rRNA gene sequences as well as nine of these 20 proteins (Fig. 5). Four proteins more conserved than the average (see Table S7) were chosen to distinguish different genera, and five highly divergent proteins to distinguish different species. As multiple sequences are proposed for classification of new isolates, this approach is robust to a small number of missing genes. In the 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 to use 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., 2014; Sun et al., 2012).

The validity of this new approach could be confirmed with the classification of five 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 strain representing the genus Neochlamydia 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 order Chlamydiales. Thanks to the democratization of new sequencing technologies, we recommend sequencing the whole genome for the taxonomic characterization of available strains and concatenating the sequences of the nine genes to derive the taxonomic affiliation of a novel strain. Alternatively, when the isolate has not been obtained in culture and an 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.

CONCLUSION

In this study, we explored different approaches to determine the ability of core proteins of members of the order Chlamydiales to produce robust phylogenies. The reconstruction of chlamydial phylogeny based on 424 groups of orthologues 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 reconstruct robust bacterial phylogeny. Despite the fact that nearly all topologies inferred from individual protein alignments were different, only a few strongly conflicting phylogenetic signals leading to the rejection of the reference tree were found in the core gene set of the order 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 order Chlamydiales and provide a robust classification scheme. At least nine of these 20 proteins should be used to accurately assign newly discovered chlamydial strains at the family, genus and species level within the order Chlamydiales.

ACKNOWLEDGEMENTS

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

REFERENCES


