Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Pan-genome analysis of whole genome of Chlamydia trachomatis uncovers a highly conserved core and structured accessory variations

Leonardo Miranda Miranda Santos
Universidade Federal do Pará
João Victor Souza Rodrigues
joaovsrodrigues14@gmail.com
Universidade Federal do Pará
Gabriel Albuquerque Albuquerque Xavier
Universidade Federal do Pará
Felipe Bonfim Freitas
Evandro Chagas Institute
Rodrigo Vellasco Duarte Silvestre
Evandro Chagas Institute
Luiz Fernando Almeida Machado
Universidade Federal do Pará
Ricardo Ishak
Universidade Federal do Pará
Edivaldo Costa Sousa Júnior
Evandro Chagas Institute
Maisa Silva de Sousa
Universidade Federal do Pará

Research Article

Keywords: Chlamydia trachomatis, pangenome, comparative genomics, core genome, accessory genome, gene presence/absence

Posted Date: October 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11198035/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Objective

To characterize the composition, saturation, and structure of the pangenome from 70 complete chromosomal genomes of *Chlamydia trachomatis*.

Methods

The complete genomes and their annotations were obtained from NCBI. Gene families were reconstructed using Roary v3.13.0, employing a 95% protein identity threshold, and classified into core, soft-core, shell, and cloud components. Ten permutations of sample ordering were used to construct accumulation curves. The accessory component was evaluated using Jaccard distance, hierarchical clustering, a gene content tree, and the distribution of protein sequence identity.

Results

A total of 979 genes families were identified: 772 core (78.9%), 65 soft-core (6.6%), 96 shell (9.8%), and 46 cloud (4.7%). The accessory component comprised 207 families, including 22 singletons. The gene repertoire per genome ranged from 872 to 903 families. With 50 genomes, 99.5% of the observed pangenome had already been recovered, and the inclusion of the 70th genome added an average of 0.2 families. The average Jaccard accessory similarity was 0.701, and the weighted average protein identity was 98.36%.

Conclusion

The analyzed dataset displays a compact and strongly saturated pangenome, featuring broad core conservation and variability concentrated within a structured fraction of the accessory repertoire. The validation of rare genes and evolutionary inference require uniform reannotation and an independent core-genome-based phylogeny.

1. Introduction

Chlamydia trachomatis causes the most prevalent bacterial sexually transmitted infection (STI) in at least 86 countries the majority of which lack control schemes for this infection (Sameni et al., 2025) accounting for approximately 235 million annual cases, predominantly among young individuals (Hu et al., 2025) and socioeconomically vulnerable populations (Crichton et al., 2015). This underscores the need for control strategies tailored to different epidemiological contexts (Sameni et al., 2025). *C. trachomatis* belongs to the family *Chlamydiaceae*, which exhibits a biphasic developmental cycle and obligate intracellular energy parasitism. Owing to the evolutionary reduction of its biosynthetic repertoire,

C. trachomatis acquired the membrane proteins Npt1 and Npt2, which specialize in the import and transport of nucleotide triphosphates and ATP from the host cell to fulfill its energetic requirements (Tjaden et al., 1999; Fisher et al., 2013). These adaptive attributes stem from the genomic stability of *C. trachomatis* (Luu et al., 2023), which is conferred by its compact chromosome of approximately 1 Mb, high synteny, and a reduced gene repertoire. Despite the high degree of conservation and compactness of the *C. trachomatis* genome, variable regions exist such as the Plasticity Zone alongside variable genes that differ among lineages. Furthermore, genetic recombination is frequent in specific regions of its genome, including the *ompA* gene and proteins of the Pmp family (Joseph et al., 2011; Bowden et al., 2021). These variations can impact future therapeutic regimens and diagnostic targets, highlighting the urgent need to monitor *C. trachomatis* strains globally through whole-genome analysis (Collingro et al., 2011; Sigalova et al., 2019; Dharamshi et al., 2023). Historically, the serological and genetic detection of *C. trachomatis* has relied on the Major Outer Membrane Protein (MOMP) and its encoding gene, *ompA*, respectively. The *ompA* gene integrates conserved segments with four variable domains, whose diversity enables the discrimination of genotypes and serovars (Stephens et al., 1987; Lysén et al., 2004). Although this approach is useful for typing, recombination and the limited information inherent to a single locus can obscure relationships among lineages. Whole-genome analyses demonstrate that isolates clustered by conventional typing may exhibit distinct phylogenomic relationships, with diversity distributed heterogeneously across the chromosome (Harris et al., 2012; Hadfield et al., 2017).

Pangenome analysis thus provides a quantitative framework to delineate universal, persistent, and variable gene families, while evaluating how much observed diversity is introduced through expanded sampling. Within this context, the present study aimed to characterize the *C. trachomatis* pangenome using 70 complete chromosomal genomes available in the NCBI database. We quantified the core, soft-core, shell, and cloud components, evaluated repertoire saturation using accumulation curves, and described the presence-absence patterns of accessory gene families. Additionally, we examined the similarity among accessory profiles alongside the sequence identity distribution of the clustered proteins. The study design is comparative and descriptive; owing to the lack of standardized clinical and geographic metadata, epidemiological associations, relationships with clinical genotypes, and the performance of diagnostic assays were not evaluated.

2. Materials and Methods

2.1 Study design and genome selection

A comparative genomics study was conducted using complete genomes of *Chlamydia trachomatis* available in the National Center for Biotechnology Information (NCBI). Records classified as complete genomes and accompanied by gene annotation were included, to reduce artificial gene loss resulting from assembly fragmentation. Records were identified by their accession numbers and maintained as independent sampling units. The final dataset included 70 genomes, with accession numbers ranging from NC_000117 to NC_023060. Sequences lacking usable annotation or failing to meet the completeness criteria were not considered. For each accession, the chromosomal sequence and

corresponding annotation were retrieved. Sample traceability was preserved using the accession number across all steps. We selected all high-quality, complete chromosomal genomes covering all major serovars A–L. The comparative analysis included 70 chromosomal accessions; plasmids were neither treated as sampling units nor analyzed separately. This approach avoids attributing differences arising from extrachromosomal elements to the chromosome.

2.2 Standardization and Quality Control of Annotations

Prior to comparative analysis, we verified the alignment between sequences and their respective annotations, the uniqueness of identifiers, and the availability of coding regions alongside their translated products. The original annotations from public records were preserved without performing uniform *de novo* reannotation. Consequently, products lacking specific functional assignments were retained as hypothetical proteins or cluster identifiers, without introducing further functional inference. Annotations were structured into a Roary-compatible format and subjected to a uniform clustering threshold. Downstream consistency was evaluated based on the presence of 70 sample columns within the binary matrix, as well as the concordance among total gene families, frequency categories, and protein associations. Because the original annotations were not re-generated using a single predictor, an absence in the matrix was interpreted as the lack of a detectable family under the implemented pipeline, rather than an automatically confirmed chromosomal deletion.

2.3 Translation of Coding Sequences

Coding sequence (CDS) nucleotide sequences were translated using EMBOSS Transeq from the EMBOSS package v6.6.0 (Rice *et al.*, 2000), applying the bacterial genetic code and the strand orientation specified in the annotation. For CDSs annotated on the reverse strand, the reverse-complement sequence was used prior to translation. Translated sequences were inspected for internal stop codons and ambiguous amino acid residues. The resulting proteomes were used to verify protein sequences and conduct auxiliary similarity analyses. Pangenome reconstruction was based on standardized annotations and their corresponding CDS translations, avoiding indiscriminate six-frame translation across the complete genomes.

2.4 Pangenome Construction

The pangenome was reconstructed using Roary v3.13.0 (Page *et al.*, 2015) based on the standardized annotations. Coding sequences (CDSs) were extracted and translated into protein sequences, partial sequences were filtered, and highly similar proteins were pre-clustered. Subsequently, the reduced set was compared via BLASTp (Camacho *et al.*, 2009), clustered using the Markov Cluster algorithm (MCL; Enright *et al.*, 2002), and refined using gene synteny information to resolve paralogs. A minimum threshold of 95% protein sequence identity was applied, while maintaining default parameters and identical running conditions for all genomes. The analysis generated a binary presence/absence matrix, in which rows represented gene families and columns corresponded to the 70 genomes. A value of 1 indicated gene presence, and 0 indicated absence. The mapping between individual proteins and their respective families was preserved to enable cluster auditing, paralog identification, and verification of

families lacking specific functional annotations. Internal consistency was confirmed by verifying concordance between the total number of families and their corresponding protein associations.

2.5 Gene Family Classification

Gene families were classified based on their frequency across the set of 70 genomes: core genome, present in 99–100% of samples; soft-core, present in 95% to less than 99%; shell, present in 15% to less than 95%; and cloud, present in less than 15%. These threshold criteria were applied directly to the presence frequencies in each row of the binary matrix. For 70 genomes, this corresponded to core families present in 70 genomes, soft-core present in 67–69, shell present in 11–66, and cloud present in 1–10 genomes. As a quantitative control of the reconstruction, 979 gene families were identified: 772 core, 65 soft-core, 96 shell, and 46 cloud. Thus, 207 families comprised the accessory genome, corresponding to all non-strict core categories. Twenty-two families occurred in only a single genome. These figures were calculated directly from the presence and absence profiles and adopted as a reproducible baseline for subsequent analyses.

2.6 Pangenome, Conserved, New, and Unique Gene Curves

Pangenome dynamics were evaluated using accumulation curves constructed for the total number of gene families, conserved families, new families added with each inclusion, and unique families. Ten randomized orders of addition across the 70 genomes were employed. For each sample size n, the accumulated number of families in the pangenome, the number of conserved families, the number of new families introduced by genome n, and the number of unique families were summarized. The use of multiple permutations minimized the sensitivity of these estimates to sample input order. For each point along the curve, measures of central tendency and dispersion across the 10 permutations were calculated. Assessment of pangenome openness or closure should be grounded in the asymptotic behavior of the accumulation curve and, when fitting a model, on the parameter estimate and confidence interval of the decay parameter. An open or closed pangenome was not inferred solely from visual inspection or the observed count of singletons.

2.7 Accessory Genome Structure and Graphical Representation

The repertoire per genome was obtained by summing the presences in each column of the binary matrix; the accessory component was calculated after excluding the 772 universal families. For the 2,415 possible pairwise comparisons among 70 genomes, the similarity of accessory content was quantified using the Jaccard index. The matrix of 207 non-universal families was clustered by average linkage based on Jaccard distance, both across genomes and across families, and represented as a presence/absence heatmap. The gene content tree was derived from the binary alignment of accessory families. The topology, branch lengths, and recorded support values were preserved in the phylogenetic object. For graphical presentation, the tree was converted into a cladogram without scaling the horizontal axis to branch lengths, and nodes were rotated into increasing ladderization according to the number of descendant terminals, resolving ties by the lower accession number. Ladderization alters only

the visual ordering of branches. Because it is based on binary characters, the tree was interpreted as the structure of the accessory repertoire, rather than as a substitute for a phylogeny based on SNPs or single-copy orthologous genes.

2.8 Statistical Analyses and Visualization

Tabular processing, statistical summaries, and figures were generated in Python v3.12.13 using pandas v2.2.3, NumPy v2.3.5, SciPy v1.17.0, and Matplotlib v3.10.8. For each family, the frequency was calculated by dividing the sum of presences by 70. Continuous variables were described using mean, standard deviation, median, range, and quartiles, as applicable. Accumulation curves were summarized at each sample size across the ten available permutations. Visualizations included the frequency of pangenome categories, the accessory repertoire per genome, total and conserved family curves, marginal family gains, the clustered presence/absence heatmap, the accessory content tree, and the distribution of protein identity scores. In the latter analysis, the weighted mean identity was calculated as the sum of the products of identity and frequency, divided by the total number of associations. Given the absence of associated clinical or epidemiological metadata, all comparisons were treated as descriptive, without group-wise association testing.

2.9 Reproducibility and Quality Control of Outputs

Reproducibility was ensured by preserving accession numbers, software versions, identity thresholds, and the specific parameters applied at each stage. Audit records were maintained for family presence and absence, clustered proteins, accumulation curves, binary encodings, the tree, and identity distributions. Prior to interpretation, we verified that the analysis encompassed all 70 genomes, yielded a total of 979 families, and exhibited full concordance between the sum of the pangenome categories and the total number of identified families.

3. Results

3.1 Pangenome Size and Composition

The combined analysis of the 70 complete *Chlamydia trachomatis* genomes identified 979 non-redundant gene families. Of these, 772 were detected across all genomes and constituted the strict core genome, representing 78.9% of the total repertoire. An additional 65 families (6.6%) were classified as soft-core, occurring in 67–69 genomes. The intermediate-frequency component comprised 96 shell families (9.8%), present in 11–66 genomes, whereas 46 cloud families (4.7%) were restricted to 1–10 genomes. Considering all non-universal families as accessories, the accessory genome contained 207 families (21.1% of the pangenome), demonstrating that the vast majority of the gene repertoire remained highly conserved within the analyzed dataset.

Component	Presence criterion	Families	% of pangenome	Operational interpretation
Core estrito	70/70	772	78,9	Universally applicable
Soft-core	67–69/70	65	6,6	Almost universal
Shell	11–66/70	96	9,8	Intermediate frequency
Cloud	1–10/70	46	4,7	Low frequency
Acessório total	1–69/70	207	21,1	Not universal

Table 1. Pangenome composition according to the frequency of gene families across the 70 genomes.

3.2 Distribution of the gene repertoire among genomes

The number of gene families per genome ranged from 872 to 903, with a mean of 890.4 ± 6.6 and a median of 891. Since the 772 strict core families were present in all samples, intergenomic variation was driven by the accessory component, which ranged from 100 to 131 families per genome (mean of 118.4 ± 6.6; median of 119). On average, the accessory component represented 13.3% of the families detected in each genome. Accession NC_020935 exhibited the smallest repertoire (872 families; 100 accessory), whereas NC_020968 exhibited the largest (903 families; 131 accessory), representing a range of 31 families between the extremes. The narrow dispersion in the total number of families per genome, contrasted with differences concentrated within the non-universal component, is consistent with a predominantly conserved genomic architecture containing localized variability restricted to a limited subset of the repertoire. However, because the presence/absence matrix is derived from protein annotations, specific discrepancies should be interpreted as the absence of a detectable family under the applied clustering criteria, rather than automatically as a confirmed biological deletion.

3.3 Prevalence of accessory families

The frequency distribution was strongly skewed toward universal or nearly universal families. Among the 207 accessory families, 65 (31.4%) occurred in at least 67 genomes and, within this group, 54 were present in 69 of the 70 genomes. Thus, 837 families, equivalent to 85.5% of the pangenome, were detected in at least 95% of the samples. At the opposite extreme, the 46 cloud families included 22 singletons, six families present in two genomes, six in three, three in four, two in five, three in six, two in eight, and two in nine genomes. Singletons accounted for 2.2% of the total pangenome and 10.6% of the accessory genome. Of the 22 families exclusive to a single sample, three received explicit functional annotation: succinate dehydrogenase/fumarate reductase subunit B (*sdhB*), a protein with a YopT-type cysteine protease domain, and porphobilinogen synthase. A family annotated as a transcript cleavage factor was also recorded; the remaining 18 families were designated by generic cluster identifiers. The predominance of families lacking functional assignment among singletons limits direct biological inferences and reinforces the need for validation through inspection of genomic regions, coordinate comparison, and uniform reannotation before associating them with specific phenotypes.

3.4 Behavior of accumulation curves

The ten permutations of the genome inclusion order produced convergent trajectories. With one genome, the mean repertoire was 893.9 ± 6.7 families; after including 10 genomes, the pangenome reached 945.4 ± 4.0 families, corresponding to 96.6% of the final total. With 20 and 50 genomes, 956.4 ± 3.7 (97.7%) and 974.2 ± 3.0 families (99.5%) were recovered, respectively. At the end of the 70 inclusions, all permutations converged to 979 families. In the opposite direction, the mean number of conserved families decreased from 893.9 ± 6.7 with one sample to 836.8 ± 5.8 with 10, 817.4 ± 11.6 with 20, and 788.7 ± 9.5 with 50 genomes, stabilizing at 772 universal families. The marginal contribution of new genes fell rapidly as sampling increased. Excluding the first genome, which defines the initial repertoire, the second added an average of 20.9 ± 11.1 families and the tenth, 1.8 ± 1.9. At positions 35, 50, 60, and 70, the median was zero, with means of 1.1, 0.7, 0.3, and 0.2 new family, respectively. Consistently, the number of exclusive families converged to 22 across all permutations. This behavior indicates strong saturation of the detectable repertoire in this collection of complete genomes, although the definitive classification of the pangenome as open or closed would require formal fitting of accumulation models and expanding sampling diversity.

3.5 Similarity of accessory gene content

Pairwise comparison of the binary profiles of the 207 accessory families revealed a mean Jaccard index of 0.701 ± 0.158 across the 2,415 genome pairs, with a median of 0.688 and a range from 0.390 to 1.000. The quartiles of 0.559 and 0.831 indicated moderate heterogeneity: some genomes shared almost the entirety of their accessory repertoire, whereas other pairs differed more markedly. Twenty-one pairs exhibited identical accessory profiles, demonstrating that distinct accessions can be indistinguishable with respect to family presence/absence, without implying complete nucleotide identity. The gene content-based tree recovered multiple clusters of accessions with very short or topologically coincident branches, interspersed with lineages more differentiated in their accessory repertoire. Several internal nodes showed support equal to or greater than 0.70. Ladderizing in ascending order facilitated the reading of clusters without altering their topology. The tree should be interpreted as a representation of binary gene presence and absence characters, rather than an estimate of evolutionary relationships based on nucleotide substitutions. Even so, the coexistence of compact clades and topological separations shows that accessory variability is not uniformly distributed among the genomes (Figure 5).

3.6 Identity among clustered proteins

The distribution of protein comparisons recorded 1,860 associations in the identity strata between 95% and 100%. The weighted mean identity was 98.36%. Almost half of the associations (902; 48.5%) showed 100% identity, followed by 404 (21.7%) at 97%, 187 (10.1%) at 96%, 149 (8.0%) at 95%, 135 (7.3%) at 98%, and 83 (4.5%) at 99%. The concentration at high identity levels is consistent with the genomic proximity of the samples and supports the stability of the clusters for most of the proteins;

however, these values describe the comparisons recorded in this distribution and do not equal the mean identity of all core genome orthologs.

3.7 Synthesis of results

Taken together, the results define a compact and broadly conserved pangenome for the 70 evaluated genomes. The strict core accounted for approximately four-fifths of the families, 85.5% of the repertoire occurred in at least 95% of the samples, and the input of new families approached zero in the final stages of the permutations. The remaining diversity was driven by 207 accessory families, unevenly distributed among the samples and enriched in nearly universal or low-frequency families. This pattern highlights two main avenues for further investigation: nearly universal families absent in a few genomes, which may reflect gene loss, structural variation, or annotation inconsistencies; and cloud families, especially singletons, which require validation before being interpreted as lineage-specific acquisitions or determinants.

4. Discussion

The predominance of the strict core genome (772 families; 78.9%) and the presence of 837 families in at least 95% of the genomes demonstrate that the gene diversity of *C. trachomatis* possesses a broadly conserved genetic repertoire. The recovery of 99.5% of the observed pangenome with 50 genomes and the average gain of only 0.2 families at the seventieth inclusion reinforce the strong saturation within this collection. This pattern aligns with studies describing high conservation and genomic compaction in *Chlamydiaceae* (Collingro et al., 2011; Sigalova et al., 2019; Harris et al., 2012). However, the observed saturation should not be converted into a definitive classification of a closed pangenome, as the accumulation curves describe the 70 complete genomes actually sampled, and their shape depends on the diversity of lineages, geographic regions, and phenotypes represented. Without fitting an accumulation model, calculating the confidence interval of the openness parameter, and implementing stratified sampling, the most defensible conclusion is that of a compact and heavily saturated pangenome within the analyzed dataset. From a translational perspective, the characterization of a strict core genome spanning 78.9% of the pangenome and displaying extremely high protein identity conservation consolidates a robust foundation for universal target selection, validating the stability of biomarkers applied to nanobiotechnological diagnostics and pan-specific vaccine candidates against *C. trachomatis*.

Although quantitatively smaller, the accessory genome exhibited notable inter-strain variation (ranging from 100 to 131 families per genome; Jaccard similarity index: 0.390–1.000) and formed distinct co-occurrence blocks and clusters with variable topological support. Genome-wide analyses of *C. trachomatis* have established that sequence variation is non-uniformly distributed across the chromosome. Specifically, Jeffrey et al. (2010) demonstrated that inter-strain diversity is concentrated within localized hotspots—such as loci associated with tissue tropism and active homologous recombination, a pattern later corroborated by broader population genomics frameworks (Joseph et al.,

2012; Hadfield et al., 2017). These findings provide a compelling biological rationale for the accessory heterogeneity observed in our dataset. Nevertheless, without associated clinical or epidemiological metadata, attributing the identified accessory clades to specific genovars, tissue niches, or phenotypic traits remains speculative.

The presence of 21 genome pairs with identical accessory profiles suggests that gene presence/absence matrices offer limited resolution for discriminating closely related isolates, a pattern consistent with recent clonal expansion within specific lineages. Conversely, the marked divergence observed among less similar pairs demonstrates that a small subset of accessory families is sufficient to drive structural repertoire differentiation. In bacteria shaped by pervasive homologous recombination, such as *C. trachomatis*, this interplay likely reflects contemporary genetic mixing overlaid onto expanded clonal backbones. Consequently, while the accessory content cladogram effectively delineates functional repertoire partitioning, it cannot be used to infer common ancestry, evolutionary directionality of gene gain/loss, or transmission dynamics. Disentangling the relative contributions of contemporary mixing versus recent clonal expansion would require an independent core-genome SNP phylogeny, ideally constructed after masking recombinant regions to prevent topological distortion.

The weighted mean identity of 98.36% and the occurrence of 100% identity in 48.5% of the 1,860 associations are consistent with the high genomic conservation described for *C. trachomatis*. Whole-genome population analyses have also recovered closely related lineages over a conserved chromosomal backbone, even though recombination and clonal expansion produce localized differentiation (Harris et al., 2012; Hadfield et al., 2017; Sigalova et al., 2019). In the present study, however, the identity distribution is conditioned by the 95% clustering threshold and represents only the associations recorded between 95% and 100%; it does not equate to the mean identity of the core genome, average nucleotide identity (ANI) between genomes, or a direct measure of selective pressure.

The 22 unique families represent only 2.2% of the pangenome, with the majority retained under generic cluster identifiers. This finding may reflect true rare genes, sequence fragments, pseudogenization, sequence divergence exceeding the clustering threshold, or discrepancies among public annotations. The design of Roary inherently depends on the quality of the input annotations, and methodological studies demonstrate that assembly fragmentation, disrupted genes, and prediction errors can artificially inflate the accessory genome or reduce the core genome (Page et al., 2015; Denton et al., 2014; Tonkin-Hill et al., 2020). Although the exclusive use of complete genomes mitigates assembly fragmentation artifacts, it does not eliminate annotation discrepancies. Consequently, *sdhB*, the protein containing a YopT-type cysteine protease domain, porphobilinogen synthase, and the transcript cleavage factor constitute candidates for targeted inspection rather than isolated evidence of functional innovation. Confirmation should integrate uniform re-annotation, sequence searches across genomes where the family was marked as absent, local synteny inspection, and CDS integrity verification. This validation is especially critical for the 18 families lacking explicit functional assignments. Taken together, our data support a compact pangenome characterized by strong gene content conservation and variability concentrated within a limited fraction of the repertoire, confirming the classic genomic nature of

Chlamydiaceae and representing a positive impact on genome-based laboratory diagnostic methods, thereby facilitating future vaccination strategies.

Although clinical metadata for the 70 samples were not available, the presence of 207 accessory families and their segregation into distinct clades support the hypothesis that genomic plasticity in *C. trachomatis* is highly compartmentalized. This structured variation likely reflects tissue-specific niche pressures such as ocular, urogenital, and lymphogranuloma venereum-associated tropisms linked to the differential retention or loss of metabolic loci, such as the tryptophan operon located within the plasticity zone (Jeffrey et al., 2010). Understanding this compartmentalization is crucial for translational applications: while the highly conserved core genome (~ 79%) offers a stable, universal repertoire for broad-spectrum vaccine candidates and pan-specific diagnostic assays, the characterization of the accessory repertoire reveals strain-specific variations that must be considered to avoid diagnostic failures due to escape and off-target cross-reactivity (Freitas et al., 2022).

The main limitations of this study include the absence of standardized re-annotation, the lack of clinical/epidemiological metadata records, the absence of formal fitting to pangenome openness models, and the lack of an independent core-genome-based phylogeny. Such limitations do not invalidate our descriptive findings; rather, they define the scope of inference, establishing a solid genomic foundation for the 70 samples and providing concrete hypotheses for subsequent functional and phylogenomic validations.

5. Conclusion

The comparative analysis of 70 complete *C. trachomatis* chromosomes defined a compact pangenome of 979 gene families, overwhelmingly dominated by a strict core genome of 772 universal families (78.9%) and 65 soft-core families (6.6%). The rapid convergence of the accumulation curves, recovering 99.5% of the observed repertoire with 50 genomes and reaching a marginal gain close to zero, demonstrates strong pangenomic saturation within the analyzed collection. Intergenomic variability is strictly compartmentalized into 207 accessory families (21.1%), whose structured distribution reflects localized repertoire differentiation over a highly conserved chromosomal backbone, consistent with the pressures of an obligate intracellular lifestyle. The results support a model of high genomic conservation with localized diversity, coherent with the species obligate intracellular biology. However, the lack of uniform re-annotation and of clinical and geographical metadata limits the interpretation of singletons, gene losses, and clustering. Future studies should integrate standardized re-annotation, synteny validation, core-genome SNP phylogenies, and epidemiological metadata to determine the origin and functional significance of accessory variation

Declarations

CRediT authorship contribution statement

Conceptualization: L.M.S; E.C.S.J;

Methodology: E.C.S.J;

Software: E.C.S.J;

Validation: L.M.S; E.C.S.J

Formal analysis: L.M.S; E.C.S.J; L.F.A.M.; M.S.S.

Investigation: L.M.S; E.C.S.J

Resources: E.C.S.J; R.V.D.S.;

Data Curation: L.M.S; E.C.S.J; F.B.F.; R.I.

Writing – Original Draft: L.M.S; E.C.S.J.

Writing – Review & Editing: L.M.S; E.C.S.J; F.B.F.; R.I.

Visualization: L.M.S; E.C.S.J; J.V.S.R.; G.A.X.

Supervision: L.M.S; E.C.S.J; M.S.S.

Project administration: L.M.S; E.C.S.J; M.S.S.

Funding acquisition: E.C.S.J

Funding

The publication of this article was funded through an agreement with the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil.

Declaration of Competing Interest

The authors declare no conflicts of interest.

Acknowledgments

None

Data availability

All genome sequences analyzed in this study were retrieved from the public GenBank/RefSeq databases of the National Center for Biotechnology Information (NCBI). The complete list of genomic accessions, as well as all dataset scripts and output matrices generated during the pangenome and comparative analyses, are available from the corresponding author upon reasonable request.

References

1. Bowden KE, Joseph SJ, Cartee JC, Ziklo N, Danavall D, Raphael BH, et al (2021) Whole-genome enrichment and sequencing of *Chlamydia trachomatis* directly from patient clinical vaginal and rectal swabs. mSphere 6(2):e01302-20. https://doi.org/10.1128/mSphere.01302-20
2. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421
3. Collingro A, Tischler P, Weinmaier T, et al (2011) Unity in variety: the pan-genome of the Chlamydiae. Mol Biol Evol 28(12):3253–3270. https://doi.org/10.1093/molbev/msr161
4. Crichton J, Hickman M, Campbell R, Batista-Ferrer H, Macleod J (2015) Socioeconomic factors and other sources of variation in the prevalence of genital chlamydia infections: a systematic review and meta-analysis. BMC Public Health 15:729. https://doi.org/10.1186/s12889-015-2069-7
5. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW (2014) Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol 10(12):e1003998. https://doi.org/10.1371/journal.pcbi.1003998
6. Dharamshi JE, Köstlbacher S, Schön ME, Collingro A, Ettema TJG, Horn M (2023) Gene gain facilitated endosymbiotic evolution of Chlamydiae. Nat Microbiol 8(1):40–54. https://doi.org/10.1038/s41564-022-01284-9
7. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30(7):1575–1584. https://doi.org/10.1093/nar/30.7.1575
8. Fisher DJ, Fernández RE, Maurelli AT (2013) *Chlamydia trachomatis* transports NAD via the Npt1 ATP/ADP translocase. J Bacteriol 195(15):3381–3386. https://doi.org/10.1128/JB.00433-13
9. Freitas LS, Queiroz MAF, Machado LFA, Vallinoto ACR, Ishak MOG, Santos FAA, et al (2022) Bioprospecting by phage display of mimetic peptides of *Chlamydia trachomatis* for use in laboratory diagnosis. Infect Drug Resist 15:4935–4945. https://doi.org/10.2147/IDR.S369339
10. Hadfield J, Harris SR, Seth-Smith HM, Parmar S, Andersson P, Giffard PM, et al (2017) Comprehensive global genome dynamics of *Chlamydia trachomatis* show ancient diversification followed by contemporary mixing and recent lineage expansion. Genome Res 27(7):1220–1229. https://doi.org/10.1101/gr.212647.116
11. Harris SR, Clarke IN, Seth-Smith HMB, Solomon AW, Cutcliffe LT, Marsh P, et al (2012) Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. Nat Genet 44(4):413–419. https://doi.org/10.1038/ng.2214
12. Hu J, Lan Y, Zhang D, Lan G, Wei J (2025) Global, regional, and national burden of chlamydial infection: a systematic analysis of incidence, prevalence, deaths, and DALYs with projections to 2046. Front Public Health 13:1674277. https://doi.org/10.3389/fpubh.2025.1674277
13. Jeffrey BM, Suchland RJ, Quinn KL, Davidson JR, Stamm WE, Rockey DD (2010) Genome sequencing of recent clinical *Chlamydia trachomatis* strains identifies loci associated with tissue tropism and regions of apparent recombination. Infect Immun 78(6):2544–2553. https://doi.org/10.1128/IAI.01324-09

14. Joseph SJ, Didelot X, Gandhi K, Dean D, Read TD (2011) Interplay of recombination and selection in the genomes of *Chlamydia trachomatis*. Biol Direct 6:28. https://doi.org/10.1186/1745-6150-6-28
15. Joseph SJ, Didelot X, Rothschild J, de Vries HJ, Morré SA, Read TD, et al (2012) Population genomics of *Chlamydia trachomatis*: insights on drift, selection, recombination, and population structure. Mol Biol Evol 29(12):3933–3946. https://doi.org/10.1093/molbev/mss198
16. Luu LDW, Kasimov V, Phillips S, Myers GSA, Jelocnik M (2023) Genome organization and genomics in Chlamydia: whole genome sequencing increases understanding of chlamydial virulence, evolution, and phylogeny. Front Cell Infect Microbiol 13:1178736. https://doi.org/10.3389/fcimb.2023.1178736
17. Lysén M, Österlund A, Rubin CJ, Persson T, Persson I, Herrmann B (2004) Characterization of *ompA* genotypes by sequence analysis of DNA from all detected cases of *Chlamydia trachomatis* infections during 1 year of contact tracing in a Swedish county. J Clin Microbiol 42(4):1641–1647. https://doi.org/10.1128/JCM.42.4.1641-1647.2004
18. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31(22):3691–3693. https://doi.org/10.1093/bioinformatics/btv421
19. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16(6):276–277. https://doi.org/10.1016/S0168-9525(00)02024-2
20. Sameni F, Hajiarab T, Bayat S, et al (2025) Global mapping of sexually transmitted *Chlamydia trachomatis* disease: a systematic review and meta-analysis. BMC Infect Dis 25:1383. https://doi.org/10.1186/s12879-025-11809-2
21. Sigalova OM, Chaplin AV, Bochkareva OO, et al (2019) Chlamydia pan-genomic analysis reveals balance between host adaptation and selective pressure to genome reduction. BMC Genomics 20:710. https://doi.org/10.1186/s12864-019-6059-5
22. Stephens RS, Sanchez-Pescador R, Wagar EA, Inouye C, Urdea MS (1987) Diversity of *Chlamydia trachomatis* major outer membrane protein genes. J Bacteriol 169(9):3879–3885. https://doi.org/10.1128/JB.169.9.3879-3885.1987
23. Tjaden J, Winkler HH, Schwöppe C, Van Der Laan M, Möhlmann T, Neuhaus HE (1999) Two nucleotide transport proteins in *Chlamydia trachomatis*, one for net nucleoside triphosphate uptake and the other for transport of energy. J Bacteriol 181(4):1196–1202. https://doi.org/10.1128/JB.181.4.1196-1202.1999
24. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al (2020) Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol 21:180. https://doi.org/10.1186/s13059-020-02090-4

Figures

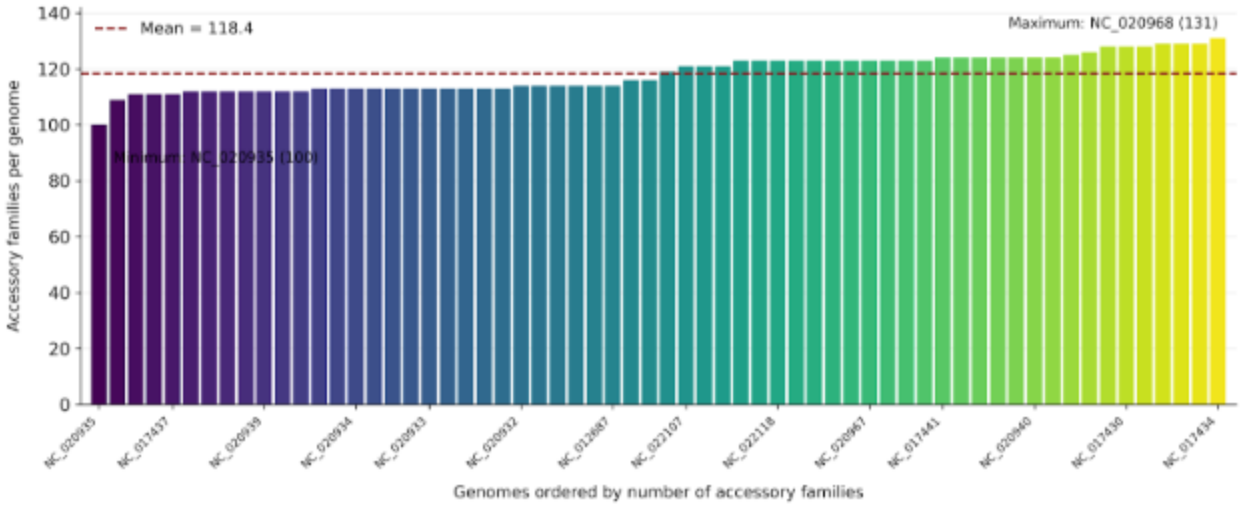


Figure 1

Distribution of the accessory repertoire across the 70 genomes. Samples are ordered by the number of detected non-universal families; the dashed line indicates the dataset mean. The extremes highlight that intergenomic variation is concentrated within the accessory component.

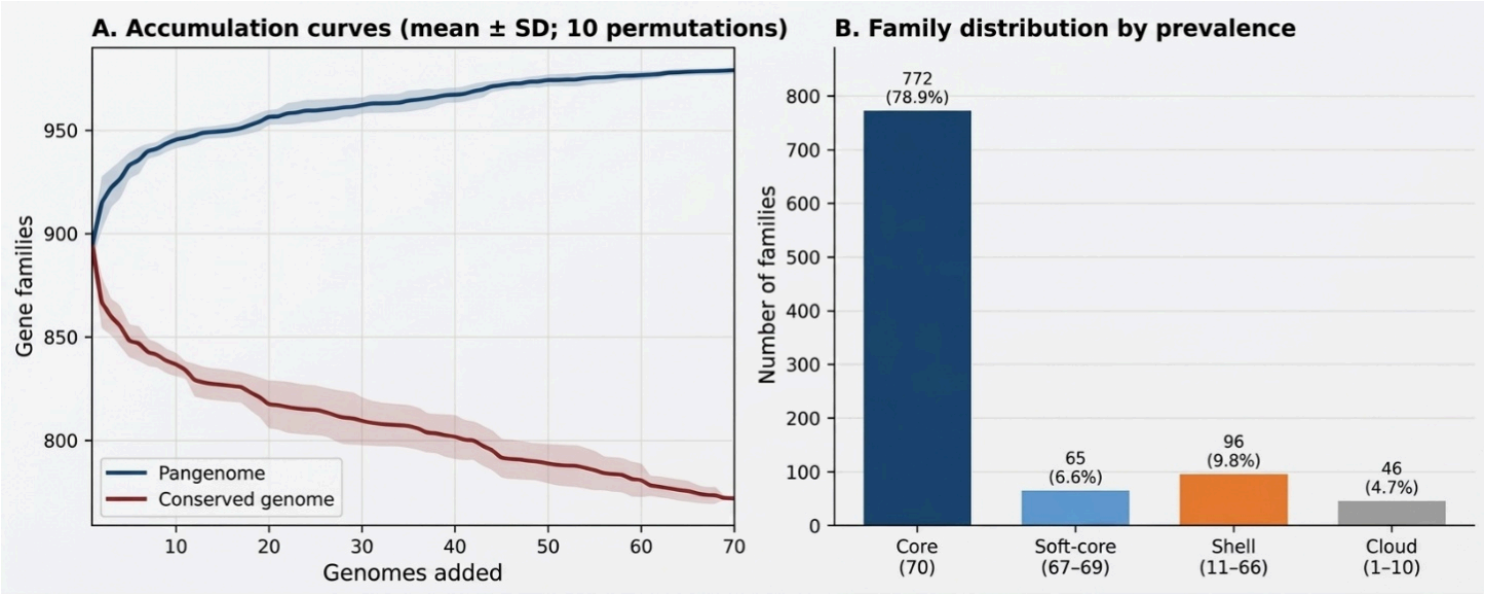


Figure 2

Dynamics of the *C. trachomatis* pangenome. (A) Mean and standard deviation of the cumulative number of total and conserved families across ten random inclusion orders of the 70 genomes. (B) Distribution of the 979 families according to their prevalence in the dataset.

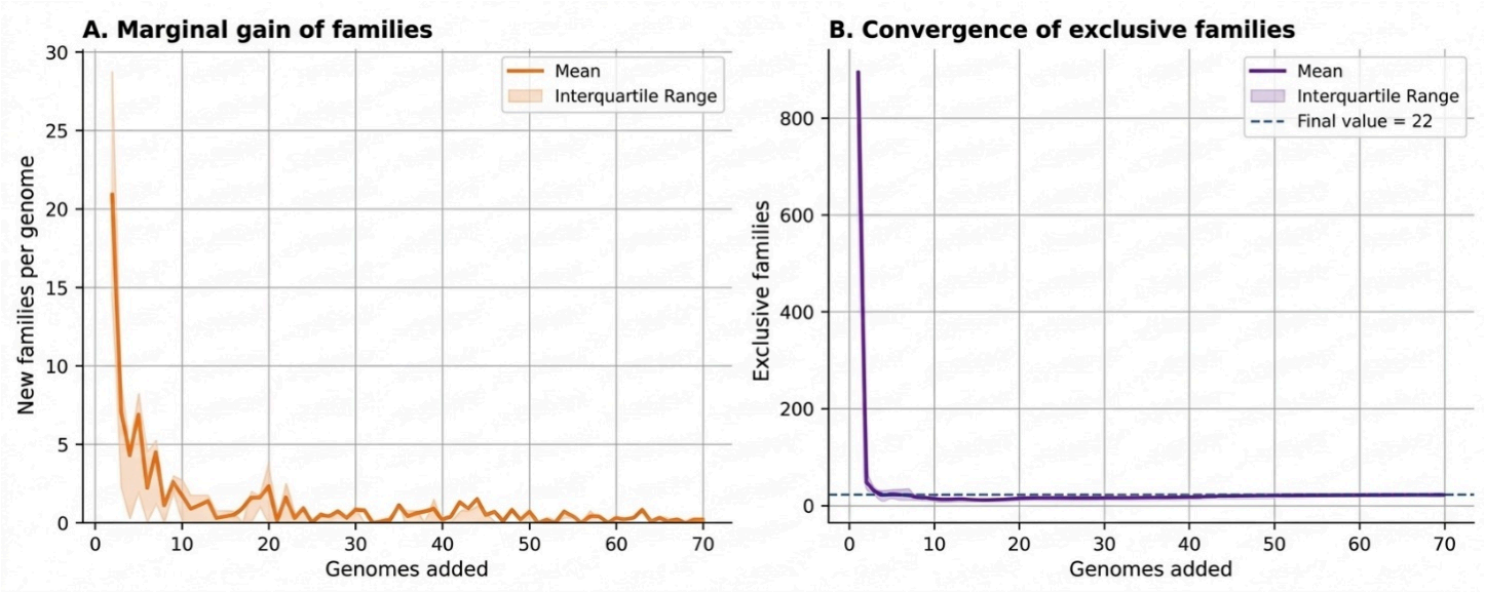


Figure 3

Marginal behavior of the pangenome across the ten permutations. (A) Mean number of new families incorporated with each additional genome, with interquartile range. (B) Convergence of the number of exclusive families to 22 at the end of sampling.

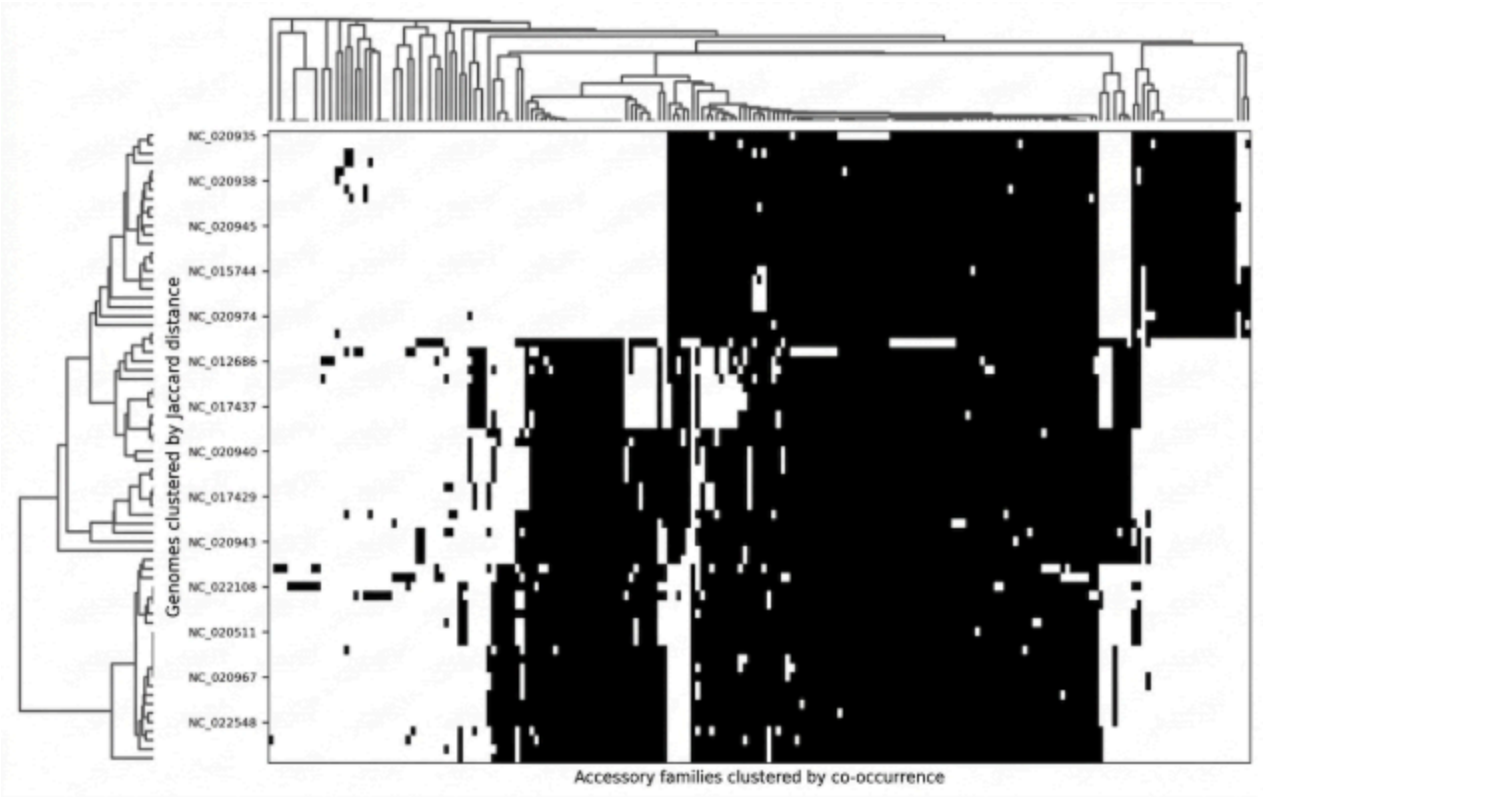


Figure 4

Structure of the accessory gene content**.** Heatmap of the 207 non-universal families across the 70 genomes, with presence in black and absence in white. Genomes and families were clustered by average

linkage based on Jaccard distance. The blocks highlight sets of co-occurring families and groups of genomes with similar repertoires; the clustering represents gene content, not a substitution-based phylogeny.

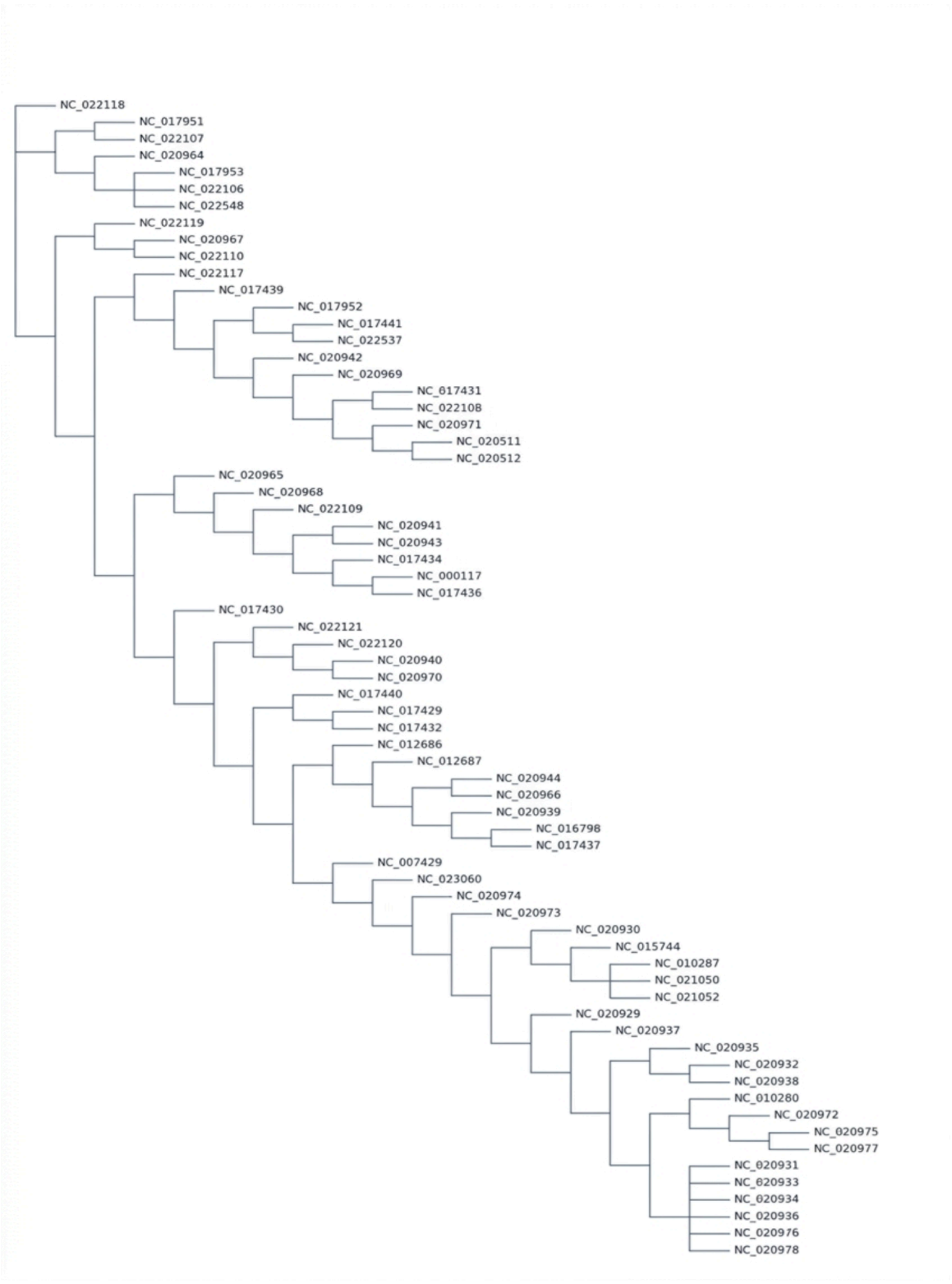


Figure 5

Tree based on the accessory gene content of the 70 *Chlamydia* genomes. The topology derives from the binary presence and absence states of non-universal families and is presented as a cladogram ladderized in ascending order by the number of descendant terminals. This rotation does not alter the topology; branch lengths are not drawn to scale in the figure. Support values ≥ 0.70 are indicated at the internal nodes. The tree represents accessory repertoire similarity and does not replace a phylogeny based on SNPs or the concatenated core genome alignment.

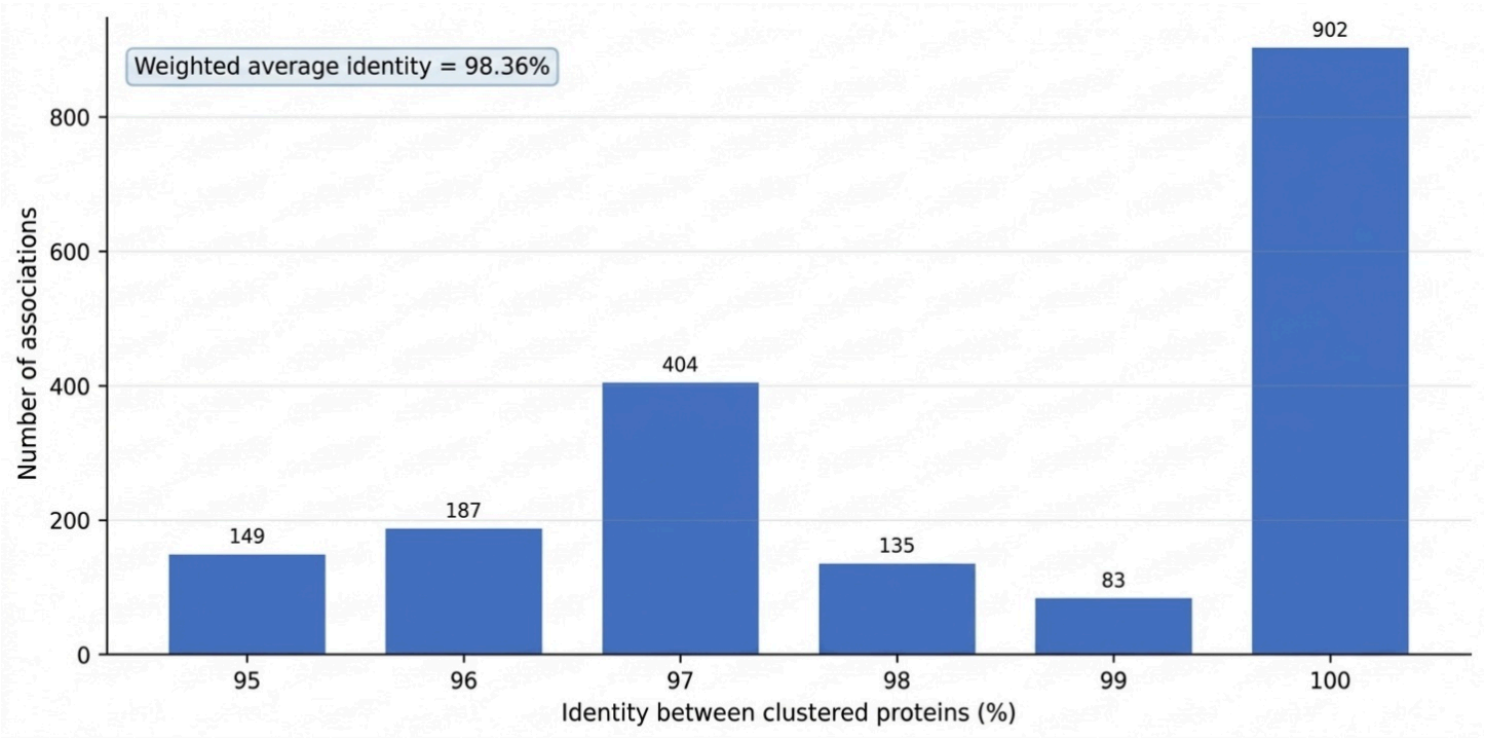


Figure 6

Distribution of protein associations by identity range. The concentration between 95% and 100%, with a higher frequency of identical sequences, demonstrates high conservation among the proteins included in the evaluated clusters.