nature microbiology

Article

https://doi.org/10.1038/s41564-022-01284-9

Gene gain facilitated endosymbiotic evolution of Chlamydiae

Received: 18 November 2021

Accepted: 7 November 2022

Published online: 5 January 2023

Check for updates

Jennah E. Dharamshi [1,5], Stephan Köstlbacher [2,3,4,5], Max E. Schön [1], Astrid Collingro [2], Thijs J. G. Ettema [1,4] ✉ & Matthias Horn [2] ✉

Chlamydiae is a bacterial phylum composed of obligate animal and protist endosymbionts. However, other members of the Planctomycetes–Verrucomicrobia–Chlamydiae superphylum are primarily free living. How Chlamydiae transitioned to an endosymbiotic lifestyle is still largely unresolved. Here we reconstructed Planctomycetes–Verrucomicrobia–Chlamydiae species relationships and modelled superphylum genome evolution. Gene content reconstruction from 11,996 gene families suggests a motile and facultatively anaerobic last common Chlamydiae ancestor that had already gained characteristic endosymbiont genes. Counter to expectations for genome streamlining in strict endosymbionts, we detected substantial gene gain within Chlamydiae. We found that divergence in energy metabolism and aerobiosis observed in extant lineages emerged later during chlamydial evolution. In particular, metabolic and aerobic genes characteristic of the more metabolically versatile protist-infecting chlamydiae were gained, such as respiratory chain complexes. Our results show that metabolic complexity can increase during endosymbiont evolution, adding an additional perspective for understanding symbiont evolutionary trajectories across the tree of life.

Symbioses are sustained interactions between different organisms that span the mutualism–parasitism spectrum[1,2]. Symbiotic associations between bacterial symbionts and both microbial (that is, protists) and multicellular eukaryotic hosts are ubiquitous[3,4] and play essential roles, from ecosystem functioning to the evolution of biological complexity[5–8]. Driven by small population sizes, lack of recombination and host dependence, obligate intracellular symbionts—that is, endosymbionts—tend to undergo genome reduction and metabolic streamlining[9–13]. Studying the origins of ancient endosymbiotic groups is necessary to unravel symbiont evolutionary trajectories and underlying evolutionary processes. Host association has evolved multiple times in the Planctomycetes–Verrucomicrobia–Chlamydiae (PVC) superphylum, a group of bacteria consisting of the aforementioned phyla alongside Lentisphaerae, Kirimatiellaeota and other potential members[14,15]. PVC bacteria represent an ideal case for investigation of symbiont evolution because they are ubiquitous, have large variations in lifestyle and metabolism and include members of ecological, medical and industrial importance[14,16,17]. While most PVC bacteria are free living, all described Chlamydiae are obligate endosymbionts of eukaryotes[16].

Chlamydiae are well known for the medically important human pathogen *Chlamydia trachomatis* and other Chlamydiaceae family members, which are animal pathogens with a high health burden and zoonotic potential[18–20]. Chlamydiae are also ubiquitous in environmental samples[21,22] as endosymbionts of a wide range of both protist and animal hosts[15,23,24]. Nevertheless, apart from roles as pathogens (for example, Chlamydiaceae), chlamydial host effects are understudied despite host interactions spanning the mutualism–parasitism spectrum. Protist-infecting chlamydiae (for example, Parachlamydiaceae)

[1]Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [2]University of Vienna, Centre for Microbiology and Environmental Systems Science, Vienna, Austria. [3]University of Vienna, Doctoral School in Microbiology and Environmental Science, Vienna, Austria. [4]Laboratory of Microbiology, Wageningen University and Research, Wageningen, The Netherlands. [5]These authors contributed equally: Jennah E. Dharamshi, Stephan Köstlbacher. ✉e-mail: thijs.ettema@wur.nl; matthias.horn@univie.ac.at

can act as mutualists that protect against host co-infection with *Legionella* and giant viruses[25,26]. Parachlamydiaceae have larger genomes and greater metabolic capacity than the animal-pathogenic Chlamydiaceae[27,28]. Despite contrasting genomic features, all described chlamydiae share a biphasic lifestyle with an intracellular replicative phase as reticulate bodies (RBs) and a nondividing extracellular phase as elementary bodies (EBs)[19]. Chlamydiae diverged from other PVC bacteria 1–2 billion years ago (Ga)[29,30], and their endosymbiotic lifestyle is proposed to have evolved early[29,31–33]. It was also thought that the chlamydial ancestor resembled extant protist-infecting lineages and had greater coding potential and metabolic versatility while other chlamydial groups underwent genome reduction. However, because initial studies included only the minimal chlamydial genomic diversity available from cultured representatives, little is known about the evolution of endosymbiosis in Chlamydiae. Through culture-independent genomics, numerous chlamydial lineages with unknown hosts have now been retrieved from various environments[34–39]. These groups fundamentally changed our understanding of chlamydial physiology by revealing genetic potential for motility[37,38] and anaerobic metabolism[39,40]. However, all isolated chlamydiae are still obligate endosymbionts.

We leveraged the culture-independent expansion in PVC bacteria genomic diversity to investigate endosymbiont evolution exemplified by Chlamydiae. We performed in-depth phylogenomic analyses to reconstruct PVC bacteria evolutionary relationships and gene-tree-aware ancestral state reconstruction. We reconstructed key endosymbiont genomic features in the Chlamydiae ancestor, suggesting an ancient capability to infect eukaryotic hosts. The Chlamydiae ancestor was inferred to have been a motile facultative anaerobe, indicating a lifestyle involving transitions between oxic and anoxic environments. Major shifts in chlamydial energy metabolism and oxygen tolerance were later mainly driven by gene gain. Counter to our expectations for genome streamlining during endosymbiont evolution, gene gain led to expanded metabolic potential in protist-infecting chlamydial groups.

Results

Establishment of a resolved Chlamydiae species phylogeny

To accurately resolve PVC bacteria species relationships we selected high-quality genomes of species (Chlamydiae) and genus (other PVC bacteria) representatives (Extended Data Fig. 1, Supplementary Fig. 1 and Supplementary Data 1 and 2). Maximum-likelihood (ML) and Bayesian species trees were inferred using 74 concatenated single-copy marker genes, using methods to account for compositional bias and long branch attraction (Fig. 1, Supplementary Figs. 2–5, Supplementary Data 3–5 and Supplementary Discussion 1). Chlamydiae monophyly was fully supported, and chlamydial families were consistently resolved and compatible with a 16S ribosomal RNA gene phylogeny (Fig. 1 and Supplementary Figs. 3–6). Two long-branching lineages consisting of four chlamydial genomes were removed due to unstable positions (Supplementary Figs. 3 and 4 and Supplementary Discussion 1). In the final dataset, deep evolutionary relationships were consistently resolved in both Bayesian and ML analyses when compositional bias was taken into account (Fig. 1 and Supplementary Fig. 5). These final 180 PVC representatives included 91 chlamydial species sequenced from nine environments with genome size and GC content ranging from 1.05 to 3.42 Mbp and 26.2 to 49.1%, respectively. With a robust species tree and comprehensive sampling, we here propose revision of chlamydial taxonomy (Fig. 1 and Supplementary Discussion 2). Consistent with recent work[37,39–41], an early divergence of Chlamydiae into two major groups, Group 1 (G1) and Group 2 (G2), is well supported. We can further subdivide G1 into two putative orders: Simkaniales (families Simkaniaceae and Rhabdochlamydiaceae) and Anoxychlamydiales (Anoxychlamydiaceae, formerly Anoxychlamydiales[37], and Chlamydiae Clade III). G1 members are primarily represented by metagenome-assembled genomes (MAGs) sequenced from diverse environments and with unknown hosts, although several distinct groups were obtained from invertebrate animal metagenomes. The G2 subdivision includes the classical chlamydial animal pathogens, Chlamydiaceae and *Clavichlamydia*, alongside Sororchlamydiaceae[42] in the previously established Chlamydiales order[43]. Other G2 families (Parachlamydiaceae, Criblamydiaceae, Waddliaceae and orphan lineages) comprise the here-defined order Amoebachlamydiales, which primarily infect protists.

The chlamydial endosymbiotic lifestyle is ancestral

To reconstruct PVC genome evolution we used the gene-tree-aware amalgamated likelihood estimation (ALE) approach[44] (Supplementary Figs. 7–10 and Supplementary Data 7–9). Protein-coding genes from PVC genomes were clustered into homologous gene families and phylogenies inferred for the 11,996 clusters with more than three sequences. Gene tree samples were then reconciled with the species tree to infer evolutionary events (speciations, originations, duplications, transfers and losses) and proteome size (that is, the sum of inferred gene copies corrected by gene extinction probability) across ancestral nodes (Extended Data Fig. 1). Genes inferred as present with high confidence were used to reconstruct ancestral gene content, and event frequencies to assess overall patterns in gene content evolution. Of note, we can reconstruct only those genes present in extant genomes and in the dataset. Thus, ancestral reconstructions are incomplete because we miss gene families that have gone extinct or were not sampled. ALE compensates for this by taking into account estimated gene extinction rates and genome completeness for inferring gene copy numbers. Originations can derive from either de novo gene birth or horizontal gene transfer (HGT) from outside the PVC genome dataset. For Chlamydiae originations, to differentiate between these we searched for homologues in protein databases. Where identified, phylogenetic trees were inferred to discern with which taxonomic groups chlamydial genes affiliated, indicating putative donor lineages of the horizontally transferred gene families.

The last common Chlamydiae ancestor (LCCA) was reconstructed with ~1,118 protein-coding genes (Figs. 2 and 3, Extended Data Fig. 2 and Supplementary Data 9). Of these, 401 were inferred as gene gains, many associated with metabolism ($n = 99$) and obtained through HGT (Extended Data Figs. 3 and 4). Hallmark endosymbiont genomic features for host interaction, energy parasitism and the chlamydial biphasic lifecycle were gained before LCCA. These genes included a type III secretion system (T3SS) (for example, genes *sctJ*, *sctT*, *sctS*, *sctV* and *sctW*), the adhesin Ctad1 and major outer membrane protein (MOMP), two nucleotide transporters (NTTs), DsbB, glycogen biosynthesis and degradation (for example, *glgC*, *glgP*, *malQ*) and the transcriptional regulator early upstream ORF (EUO) (Fig. 2, Extended Data Figs. 2 and 5 and Supplementary Data 9 and 10). The T3SS facilitates host cell entry through effector secretion[45], and Ctad1 and MOMP are pathogenicity factors in Chlamydiaceae involved in host invasion[19]. In the RB stage, NTTs facilitate energy parasitism and metabolite scavenging by importing ATP, nucleotides and NAD$^+$ from the host cytosol[28]. In the EB stage, DNA is condensed by histone-like proteins (for example, HctA), and the cell envelope rigidified to protect against osmotic and physical stress through disulfide crosslinking of outer membrane proteins by DsbB[46]. HctA was not reconstructed in LCCA, but in all early chlamydial ancestors (Extended Data Fig. 5 and Supplementary Data 10). Glycogen is used as a carbon source by EBs and enhances extracellular survival[47]. EUO is a master regulator that represses T3SS, DNA condensation and cell surface modification genes before RB-to-EB conversion[48,49], and is a putative chlamydial gene invention.

Many gene losses (31%, $n = 362$) were inferred between LCCA and the last common ancestor of Chlamydiae, Verrucomicrobia, Lentisphaerae and Kiritimatiellaeota (LVCCA) (Fig. 3), and were predominantly associated with metabolism (49%, Extended Data Fig. 3). Relative to LVCCA, de novo amino acid and nucleotide biosynthesis

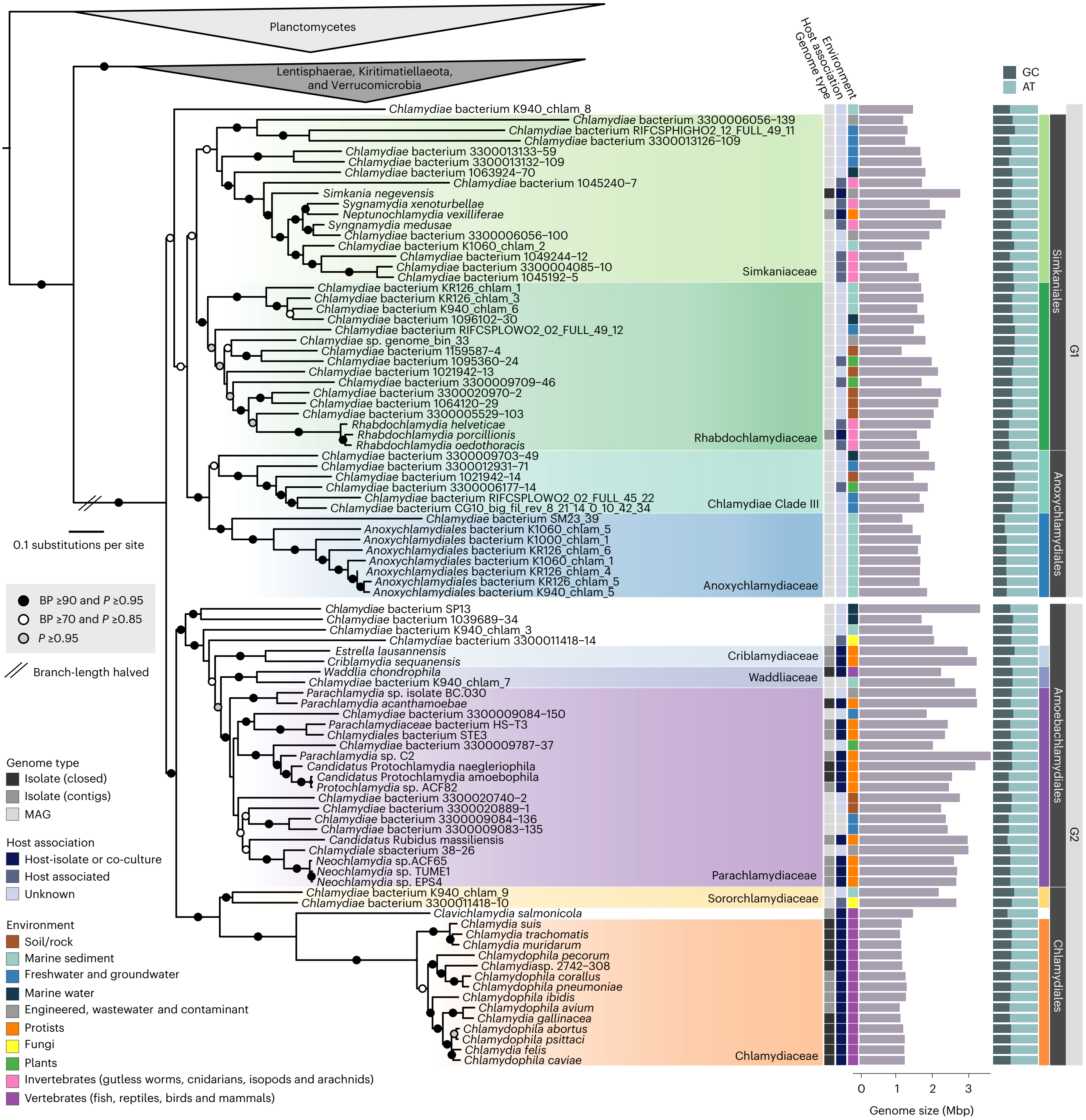

Planctomycetes
Lentisphaerae, Kiritimatiellaeota, and Verrucomicrobia
Environment
Host association
Genome type
GC
AT
Chlamydiae bacterium K940_chlam_8
Chlamydiae bacterium 3300006056–139
Chlamydiae bacterium RIFCSPHIGHO2_12_FULL_49_11
Chlamydiae bacterium 3300013126–109
Chlamydiae bacterium 3300013133–59
Chlamydiae bacterium 3300013132–109
Chlamydiae bacterium 1063924–70
Chlamydiae bacterium 1045240–7
Simkania negevensis
Sygnamydia xenoturbellae
Neptunochlamydia vexilliferae
Syngnamydia medusae
Chlamydiae bacterium 3300006056–100
Chlamydiae bacterium K1060_chlam_2
Chlamydiae bacterium 1049244–12
Chlamydiae bacterium 3300004085–10
Chlamydiae bacterium 1045192–5
Simkaniaceae
Chlamydiae bacterium KR126_chlam_1
Chlamydiae bacterium KR126_chlam_3
Chlamydiae bacterium K940_chlam_6
Chlamydiae bacterium 1096102–30
Chlamydiae bacterium RIFCSPLOWO2_02_FULL_49_12
Chlamydiae sp. genome_bin_33
Chlamydiae bacterium 1159587–4
Chlamydiae bacterium 1095360–24
Chlamydiae bacterium 1021942–13
Chlamydiae bacterium 3300009709–46
Chlamydiae bacterium 3300020970–2
Chlamydiae bacterium 1064120–29
Chlamydiae bacterium 3300005529–103
Rhabdochlamydia helveticae
Rhabdochlamydia porcillionis
Rhabdochlamydia oedothoracis
Rhabdochlamydiaceae
Chlamydiae bacterium 3300009703–49
Chlamydiae bacterium 3300012931–71
Chlamydiae bacterium 1021942–14
Chlamydiae bacterium 3300006177–14
Chlamydiae bacterium RIFCSPLOWO2_02_FULL_45_22
Chlamydiae bacterium CG10_big_fil_rev_8_21_14_0_10_42_34
Chlamydiae Clade III
Chlamydiae bacterium SM23_39
Anoxychlamydiales bacterium K1060_chlam_5
Anoxychlamydiales bacterium K1000_chlam_1
Anoxychlamydiales bacterium KR126_chlam_6
Anoxychlamydiales bacterium K1060_chlam_1
Anoxychlamydiales bacterium KR126_chlam_4
Anoxychlamydiales bacterium KR126_chlam_5
Anoxychlamydiales bacterium K940_chlam_5
Anoxychlamydiaceae
Simkaniales
Anoxychlamydiales
G1
0.1 substitutions per site
● BP ≥90 and P ≥0.95
○ BP ≥70 and P ≥0.85
◉ P ≥0.95
// Branch-length halved
Chlamydiae bacterium SP13
Chlamydiae bacterium 1039689–34
Chlamydiae bacterium K940_chlam_3
Chlamydiae bacterium 3300011418–14
Estrella lausannensis
Criblamydia sequanensis
Criblamydiaceae
Waddlia chondrophila
Chlamydiae bacterium K940_chlam_7
Waddliaceae
Parachlamydia sp. isolate BC.030
Parachlamydia acanthamoebae
Chlamydiae bacterium 3300009084–150
Parachlamydiaceae bacterium HS–T3
Chlamydiales bacterium STE3
Chlamydiae bacterium 3300009787–37
Parachlamydia sp. C2
Candidatus Protochlamydia naegleriophila
Candidatus Protochlamydia amoebophila
Protochlamydia sp. ACF82
Chlamydiae bacterium 3300020740–2
Chlamydiae bacterium 3300020889–1
Chlamydiae bacterium 3300009084–136
Chlamydiae bacterium 3300009083–135
Candidatus Rubidus massiliensis
Chlamydiale sbacterium 38–26
Neochlamydia sp.ACF65
Neochlamydia sp. TUME1
Neochlamydia sp. EPS4
Parachlamydiaceae
Chlamydiae bacterium K940_chlam_9
Chlamydiae bacterium 3300011418–10
Sororchlamydiaceae
Clavichlamydia salmonicola
Chlamydia suis
Chlamydia trachomatis
Chlamydia muridarum
Chlamydophila pecorum
Chlamydiasp. 2742–308
Chlamydophila corallus
Chlamydophila pneumoniae
Chlamydophila ibidis
Chlamydophila avium
Chlamydia gallinacea
Chlamydophila abortus
Chlamydophila psittaci
Chlamydia felis
Chlamydophila caviae
Chlamydiaceae
Amoebachlamydiales
Chlamydiales
G2
Genome type
■ Isolate (closed)
■ Isolate (contigs)
■ MAG
Host association
■ Host-isolate or co-culture
■ Host associated
■ Unknown
Environment
■ Soil/rock
■ Marine sediment
■ Freshwater and groundwater
■ Marine water
■ Engineered, wastewater and contaminant
■ Protists
■ Fungi
■ Plants
■ Invertebrates (gutless worms, cnidarians, isopods and arachnids)
■ Vertebrates (fish, reptiles, birds and mammals)
0
1
2
3
Genome size (Mbp)

Fig. 1 | Robust species phylogeny of PVC bacteria. Concatenated (74 marker genes) Bayesian phylogeny of 180 PVC bacteria with compositionally heterogeneous sites removed (8,151 amino acid sites). Circles indicate bipartition support from posterior probability (P) (CAT + GTR + Γ4 model), and nonparametric bootstraps (BP) (LG + C60 + F + Γ4-derived PMSF approximation). The tree is rooted by Planctomycetes. Reduced branch lengths are indicated by parallel lines, and substitutions per site by the scale bar. Genome type, evidence for host association and the environment from which the genome was obtained are indicated by coloured squares according to the legend. Genome size (Mbp in purple) and GC content (%GC in dark blue and %AT in light blue) are indicated by bars. Higher-level taxonomic classifications are indicated and chlamydial families outlined by coloured boxes. See also Supplementary Data 2 for genome characteristics and Supplementary Data 6 for the uncollapsed species phylogeny.

capabilities were strongly reduced in LCCA through loss of genes for histidine, arginine, tryptophan, methionine, valine, leucine, isoleucine, phenylalanine, threonine and purine (for example, *purC*, *purD* and *purH*) biosynthesis (Fig. 2, Extended Data Fig. 2 and Supplementary Data 9). However, a suite of amino acid and oligopeptide transporters was already inferred as being present in LVCCA and maintained in LCCA, with both able to acquire amino acids from external sources. NAD and NADP biosynthesis genes were inferred in LCCA. Pathways for biosynthesis of other cofactors, such as ferredoxin and cobalamin, were inferred as lost and LCCA probably depended on their uptake. Several Chlamydiaceae virulence factors associated with host metabolite degradation (for example, proteases and lipases) were inferred as

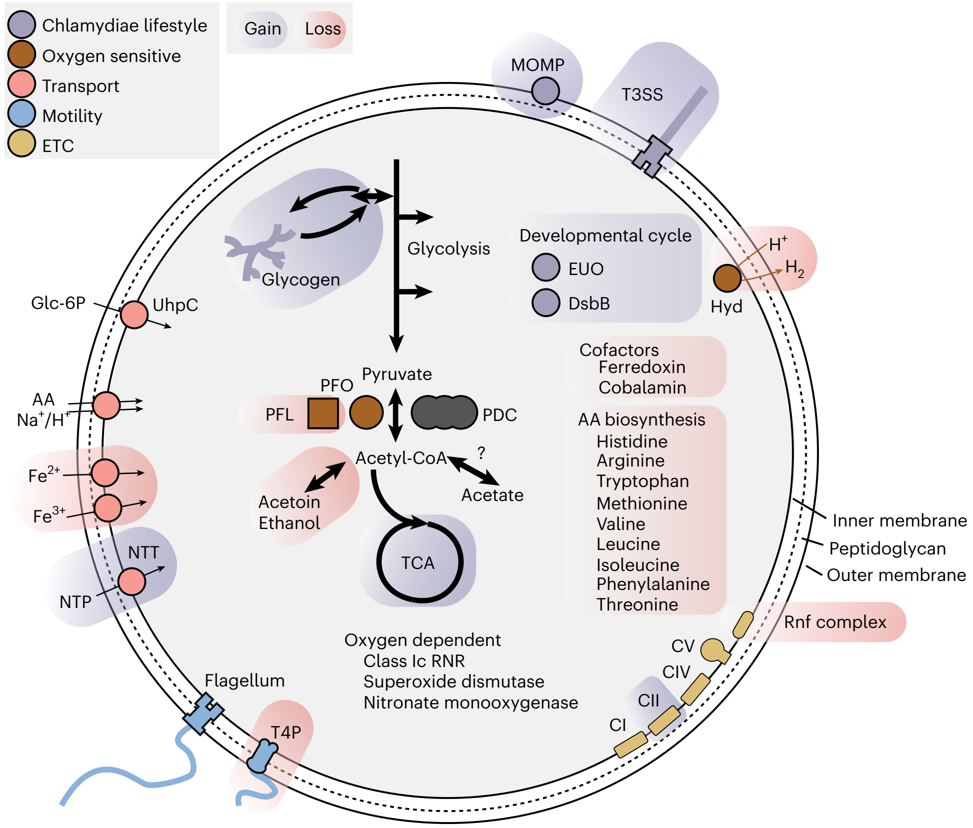


Fig. 2 | Schematic of gene content reconstructed in LCCA. Gains and losses in LCCA relative to LVCCA are indicated by blue and red backgrounds, respectively. The reconstructed presence of a peptidoglycan-based cell wall is indicated by the dashed line between inner and outer membranes. ADI, arginine deiminase pathway; CI–V, electron transport chain complexes I–V; Hyd, [FeFe]-hydrogenase; PDC, pyruvate dehydrogenase; T4P, type IV pilus. See also Extended Data Fig. 2 for a more detailed summary and Supplementary Data 9 for gene content reconstructed in LCCA and LVCCA.

present already in LVCCA and retained throughout chlamydial evolution (Extended Data Fig. 5 and Supplementary Data 10). LVCCA also already had the potential to import glucose-6-phosphate with the transporter UhpC, which is used to scavenge host glucose by chlamydiae and other endosymbionts[50] (Fig. 2 and Supplementary Data 9). While peptidoglycan biosynthesis genes are absent in some Planctomycetes[51], we reconstructed most key genes (for example, *murACDEFGJ*, *mreB* and *mraY*) in both LVCCA and LCCA (Fig. 2, Extended Data Fig. 2 and Supplementary Data 9 and 10).

Earlier hypotheses for Chlamydiae genome evolution were based only on gene presence patterns or included limited genomic diversity[32,33]. Our results provide the previously missing support that LCCA already had the genetic toolkit for an endosymbiotic and biphasic lifestyle. Furthermore, we have shown that key genes were gained before LVCCA, and that LCCA also evolved through a reduction in pathways involved in de novo biosynthesis and hence a dependence on uptake of essential metabolites. Chlamydiae diverged from other PVC bacteria between 1 and 2 Ga, coinciding with estimates for the evolution of eukaryotes (1.2–2.1 Ga)[29,30,32,33]. Our reconstruction demonstrates that LCCA was already an obligate endosymbiont, indicating a billion-year-old history of chlamydiae infecting eukaryotic hosts as they evolved.

A facultative anaerobic origin of Chlamydiae

Although most Chlamydiae are aerobes, groups with anaerobic metabolism (for example, Anoxychlamydiaceae) were recently identified[39,40]. To unravel the evolutionary history of aerobiosis in Chlamydiae, we investigated metabolic genes reconstructed in LCCA and LVCCA. Core metabolic genes conserved in most extant chlamydiae[28] were inferred alongside genes indicating a facultatively anaerobic lifestyle (Figs. 2 and 4, Extended Data Figs. 5 and 6 and Supplementary Data 9 and 10). LCCA had the potential to use glycolysis to generate ATP with both glucose (glucokinase) and glucose-6-phosphate (UhpC). The resulting pyruvate could be converted to acetyl-CoA using pyruvate dehydrogenase or the oxygen-sensitive pyruvate:ferredoxin oxidoreductase (PFO) under oxic and anoxic conditions, respectively. Acetyl-CoA could then be directed into the tricarboxylic acid (TCA) cycle or fermented. The TCA cycle was reconstructed as missing citrate synthase and malate dehydrogenase in LCCA, but as complete in many early chlamydial ancestors (Extended Data Fig. 6 and Supplementary Data 10). LCCA could probably perform oxidative phosphorylation, because we inferred a complete respiratory electron transport chain (ETC) including sodium-transporting NADH dehydrogenase (Nqr; Complex I, CI), succinate dehydrogenase (Sdh; CII), terminal oxidases (CIV) cytochrome *bd* ubiquinol oxidase (CydA-B) and cytochrome *c* oxidase cbb_3-type (CcoO/N), and sodium-driven ATP synthase (Ntp; CV) (Fig. 4a,b and Suplementary Data 9 and 10). These terminal oxidases have high oxygen affinity and could have been used to respire oxygen under micro-aerophilic conditions, or to provide oxidative stress protection for oxygen-sensitive enzymes like PFO[52,53]. The LCCA ETC was probably used to generate a sodium motive force (SMF), as demonstrated in *C. trachomatis*[54]. The same central metabolism was reconstructed in LVCCA, except for succinate dehydrogenase (Extended Data Fig. 6). LVCCA could also oxidize pyruvate using the oxygen-sensitive pyruvate formate lyase (PFL), and couple pyruvate oxidation to H_2 production with the oxygen-sensitive [FeFe]-hydrogenase (HydA) (Fig. 2, Extended Data Fig. 2 and Supplementary Data 9). More extensive fermentative capabilities were also reconstructed in LVCCA, which could ferment pyruvate to acetate, acetoin and, potentially,

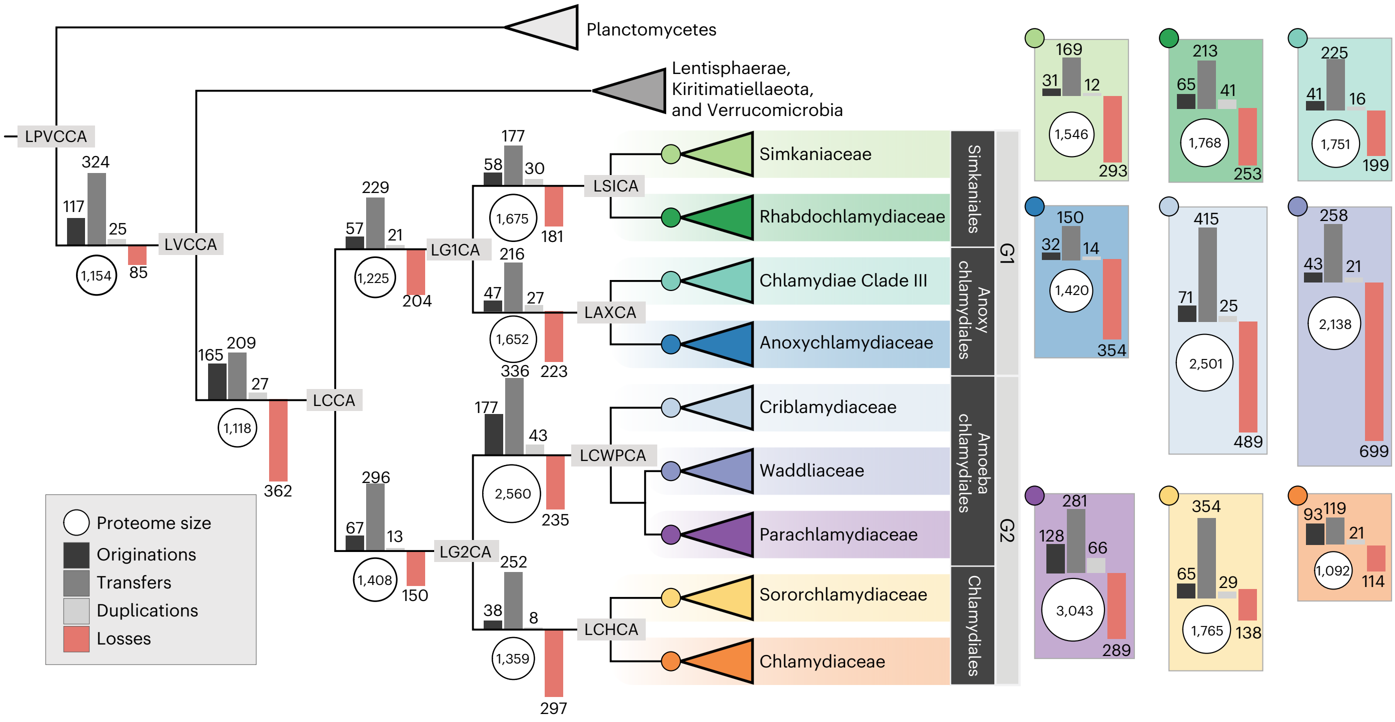


Fig. 3 | Ancestral reconstruction of Chlamydiae gene content evolution. Schematic Chlamydiae tree indicating ancestral proteome size by white circles to the left of each node. Branches are annotated with bars representing the inferred sum of origination, transfer and duplication events leading to each node (see legend). The sum of loss events is indicated negatively by a red bar. Terminal nodes represent chlamydial family ancestors, with corresponding events shown to the right. Orphan lineages are excluded from the schematic. Abbreviations of ancestors not defined in the text: all PVC bacteria (LPVCCA), Anoxychlamydiales (LAXCA) and Chlamydiales (LCHCA). See also Supplementary Data 7 for event sums across all nodes and event numbers with different reconciliation frequency cutoffs. See Supplementary Fig. 11 for events inferred at reconciliation frequencies ≥0.3.

ethanol. An Rnf complex (sodium ion-translocating ferredoxin:NAD$^+$ oxidoreductase), which is strictly linked to sodium energetics and strongly associated with anaerobes[55], was also reconstructed as part of the LVCCA ETC. PFL, HydA, the Rnf complex and some fermentative capabilities were lost between LVCCA and LCCA.

Additional genes encoding proteins with central functions and varying oxygen tolerances were reconstructed in LCCA and LVCCA, and differentially retained across early chlamydial ancestors (Fig. 2, Extended Data Figs. 2 and 5 and Supplementary Data 9 and 10). Complementary copies of ribonucleotide reductase (RNR), the key enzyme for ribonucleotide-to-deoxyribonucleotide interconversion, were reconstructed in LCCA. Under anoxic conditions LCCA and LVCCA could use a class III RNR, which is highly oxygen sensitive[56], and under oxic conditions LCCA could use a class Ic RNR, which is oxygen dependent[57]. Similarly, haem biosynthesis can occur through both anaerobic and aerobic routes. The anaerobic-route, oxygen-independent coproporphyrinogen III oxidase was reconstructed in LCCA and LVCCA and retained in all chlamydial family ancestors. The aerobic-route, oxygen-dependent protoporphyrinogen III oxidase was reconstructed in most early chlamydial ancestors, but not in LCCA and LVCCA. The oxygen-dependent superoxide dismutase and nitronate monooxygenase, which detoxify oxygen radicals and oxidize alkyl nitronates, were also reconstructed in LCCA and LVCCA. Transport systems for both primary iron species under anoxic (Fe^{2+}) and oxic (Fe^{3+}) conditions[58,59] were reconstructed in LVCCA but not in LCCA. LVCCA and LCCA were both probably motile, based on an inferred flagellar apparatus and an additional type IV pilus in LVCCA. Contrary to our expectations and previous hypotheses[32,33], in LCCA we did not reconstruct the extensive aerobic and energy metabolism found in modern Amoebachlamydiales.

LCCA and LVCCA were reconstructed as facultative anaerobes that encoded oxygen-sensitive and -dependent metabolic genes and pathways associated with both anaerobic and aerobic lifestyles. LVCCA would have lived 2 Ga, soon after the great oxidation event (2.1–2.4 Ga)[30], when environments with transient oxygen and oxic microclines would have been common. Extant facultative anaerobes in analogous environments (for example, tidal zones, sediments and animal tissues) regulate aerobic and anaerobic gene expression with oxygen exposure[60]. LVCCA may have used motility to transit oxic microclines and adjusted metabolism accordingly, and potentially had a biphasic lifestyle based on oxic–anoxic transitions rather than host invasion as in extant chlamydiae. The eukaryotic intracellular environment can provide a refuge from oxygen, and strict anaerobes can survive and divide within amoeba vacuoles when exposed to high oxygen[61]. A possible scenario that drove the evolution of chlamydial endosymbiosis was a coinciding increase in oxygen and the emergence of a niche suited to a facultative anaerobe within early eukaryotic hosts. Thus, Chlamydiae evolution may have been facilitated by both endosymbiosis-related gene gains and a facultatively anaerobic ancestor.

Gene gain facilitated oxygen tolerance and respiratory chain expansion in Chlamydiae

Chlamydiae later diversified into two major groups, G1 and G2, divergent in oxygen tolerance. Genes reconstructed in the last common ancestors of G1 (LG1CA) and G2 (LG2CA) and their descendents are indicative of lifestages in anoxic and oxic environments, respectively. The arginine deiminase pathway, known as anaerobic substrate-level phosphorylation, was reconstructed in LG1CA (Extended Data Fig. 6 and Supplementary Data 9 and 10). Similarly, the iron transporter FeoAB for the primary species under anoxic conditions (Fe^{2+}) and the hydrogen-producing and oxygen-sensitive HydA were reconstructed in Anoxychlamydiales and Anoxychlamydiaceae ancestors,

https://doi.org/10.1038/s41564-022-01284-9

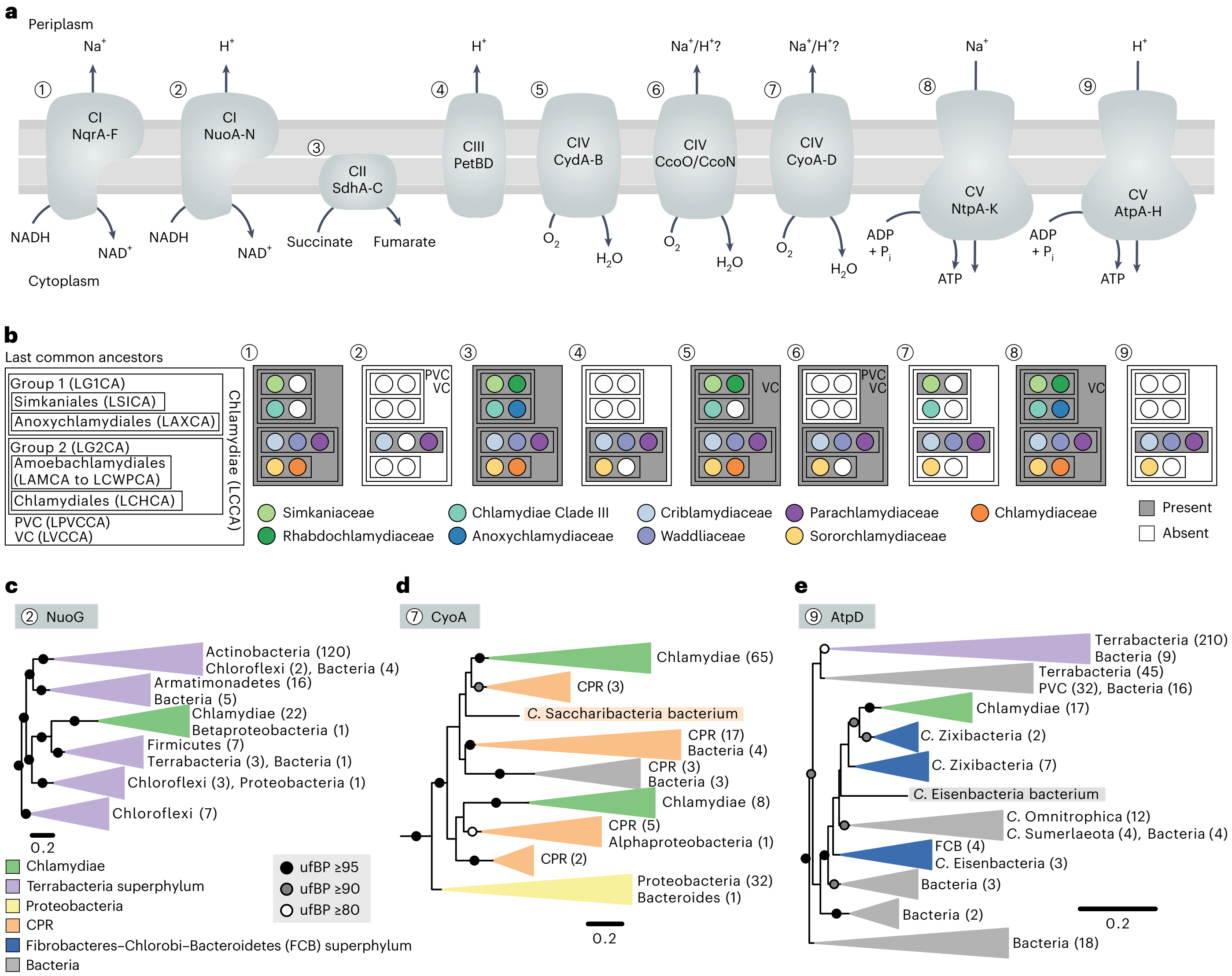


Fig. 4 | The ancestral chlamydial respiratory chain has undergone expansion in Amoebachlamydiales. a, Schematic of respiratory complexes found in Chlamydiae, with substrates and complexes. **b**, Presence and absence of respiratory complexes across chlamydial ancestors. Presence is defined as ≥50% of complex subunits. **c**–**e**, ML phylogenies of the NuoG subunit of proton-transporting NADH dehydrogenase (NuoA-N) (**c**), the CyoA subunit of cytochrome *o* ubiquinol oxidase (CyoA-D) (**d**) and the AtpD subunit of proton-driven ATP synthase (AtpA-H) (**e**), inferring with the LG model of evolution. Circles indicate ufBP bipartition support. Collapsed clades are annotated and coloured if most sequences have the same taxonomy. Scale bars indicate the number of substitutions per amino acid position in the alignment. See also Supplementary Data 6 for trees of all complex subunits and Supplementary Data 9 for inferred presence across all Chlamydiae ancestors.

respectively (Extended Data Fig. 5). Oxygen-utilizing enzymes, such as cytochrome *bd* ubiquinol oxidase, were also lost before the Anoxychlamydiaceae ancestor (Fig. 4a,b and Supplementary Data 9 and 10). In contrast, the oxygen-sensitive PFO was lost before LG2CA (Extended Data Fig. 6). Unexpected in G2 was the gain of an extensive suite of aerobiosis-associated pathways and oxygen-dependent enzymes between the last common ancestors of Amoebachlamydiales (LAMCA) and Criblamydiaceae, Waddliaceae and Parachlamydiaceae (LCWPCA), which indicates adaptation to higher-oxygen environments during early Amoebachlamydiales evolution (Extended Data Fig. 5 and Supplementary Data 10). These included genes encoding coproporphyrinogen III oxidase for aerobic-route haem biosynthesis, iron complex transporters (for example, siderophores) for the primary species under oxic conditions (Fe^{3+}), catalase and superoxide dismutases for oxidative stress response, a bacterial globin for nitric oxide detoxification, and the glyoxylate shunt, a TCA cycle bypass almost exclusive to aerobes[62]. Overall, early Amoebachlamydiales ancestors were probably better adapted to oxic environments, suggesting oxygen tolerance as a driving force in chlamydial evolution.

Amoebachlamydiales also expanded energy metabolism by gaining complexes for generation of a proton motive force (PMF) alongside the ancestral SMF. PMF has a larger redox gap than SMF and can result in greater ATP generation[63]. We reconstructed several PMF-associated complexes in LCWPCA, including a proton-transporting NADH dehydrogenase (NuoA-N; CI), cytochrome *bc* complex (PetBD; CIII), cytochrome *o* ubiquinol oxidase (CyoA-D) and proton-driven F-type ATP synthase (AtpA-H; CV) (Fig. 4a,b). In phylogenetic trees of NuoA-H subunits, chlamydial sequences consistently branch with members of the Terrabacteria superphylum (NuoG; Fig. 4c and Supplementary Data 6), supporting gain before LCWPCA. Such physiologically coupled proteins from multisubunit complexes are often gained as a functional unit[64]. PetBD was reconstructed in LG2CA and retained in most descendants, but lost in Chlamydiaceae (Fig. 4a,b). The evolutionary history of CyoA-D in Chlamydiae is unclear. It was reconstructed in

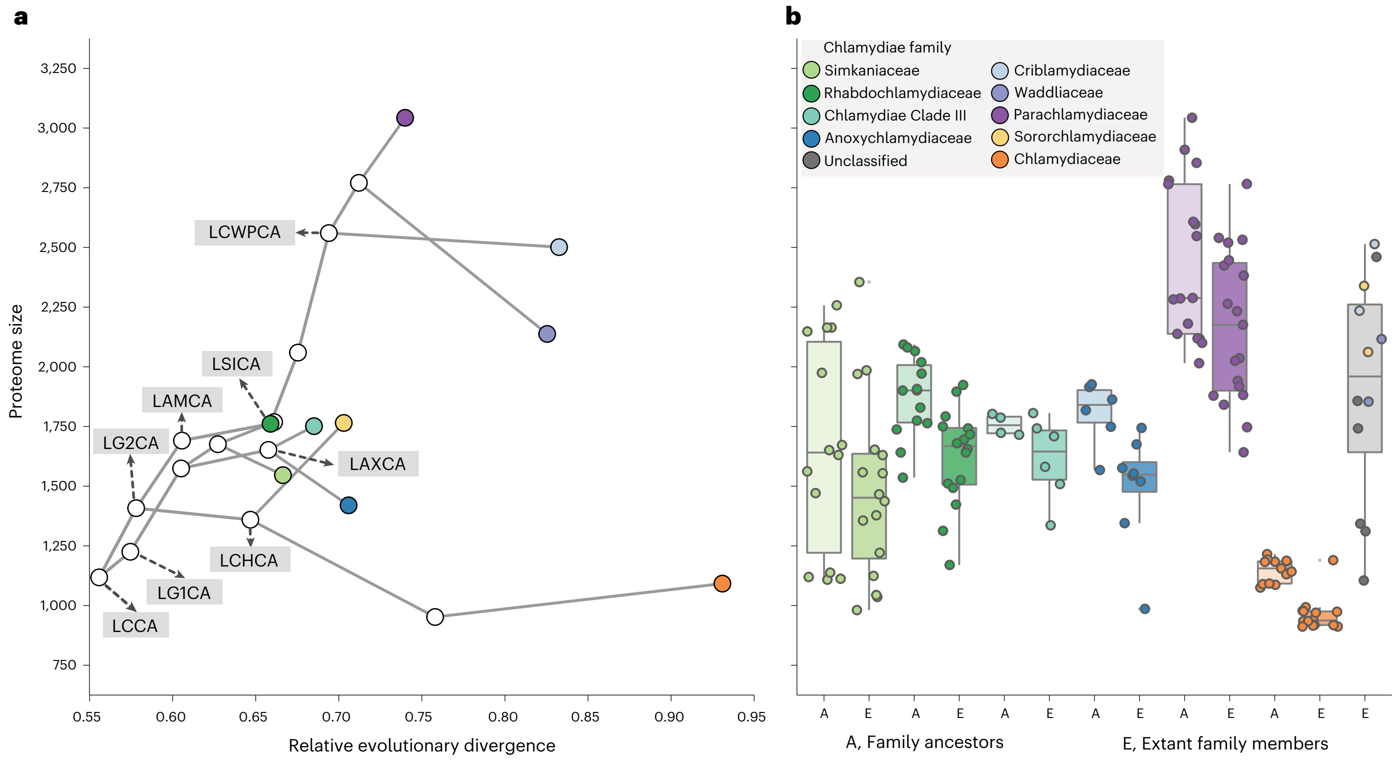


Fig. 5 | The proteome size of Amoebachlamydiales ancestors increased relative to other chlamydiae. a,**b**, Proteome size inferred for chlamydial ancestors, and protein-coding gene copies present in extant chlamydial genomes. **a**, Inferred proteome size of early chlamydial ancestors scaled to relative evolutionary divergence, from LCCA to extant taxa. **b**, Proteome size comparison between inferred ancestors and extant members within each chlamydial family. Chlamydiae without in-family ancestors are grouped together in the grey boxplot (that is, Criblamydiaceae, Waddliaceae, Sororchlamydiaceae and unclassified). Points and boxplots are coloured according to the legend. Centre lines in the box-and-whisker plot represent median values, box limits represent upper and lower quartile values and whiskers represent 1.5 times the interquartile range above the upper quartile and below the lower quartile. Number of ancestors and extant members per family depicted from left to right: n = 14, 16, 14, 16, 4, 6, 6, 8, 17, 19, 12, 14 and 12. Abbreviations of key ancestors are labelled as in Fig. 3. See also Supplementary Data 7 for proteome sizes.

Simkaniales (LSICA) and LAMCA order ancestors, but also in Chlamydiae Clade III and Sororchlamydiaceae family ancestors (Fig. 4a,b). CyoA-D gene trees show an affiliation with Candidate Phyla Radiation (CPR) bacteria members and suggest at least one HGT event (CyoA; Fig. 4d and Supplementary Data 6). CyoA-D has lower oxygen affinity and is associated with higher oxygen levels than other terminal oxidases in Chlamydiae[65]. Haem O synthase (CyoE), which generates the CyoA-D haem O cluster, was also reconstructed in LCWPCA and the Chlamydiae Clade III ancestor but not in other early ancestors (Extended Data Fig. 5). In phylogenetic trees of AtpA-H subunits, chlamydial sequences affiliate with *Candidatus* Zixibacteria and the complex was probably gained before LCWPCA in a single HGT event (AtpD; Fig. 4e and Supplementary Data 6). Thus, an extended ETC with mosaic origins was gained before LCWPCA through additive HGT of several complexes from different bacterial groups (Fig. 4). The more extensive metabolic capabilities in protist-infecting Amoebachlamydiales compared with animal pathogen Chlamydiaceae had previously been noted[27,28,66]. However, it had been presumed that differences were a result of gene loss in Chlamydiaceae and other lineages, with LCCA having had the more flexible and branched ETC[33,36]. Our analyses instead indicate that the extended Amoebachlamydiales ETC was gained after divergence from LCCA.

Gene content expansion as a mode of evolution in endosymbionts

Amoebachlamydiales aerobiosis-associated gene expansion was accompanied by additional metabolic gene gains (Extended Data Fig. 3), in line with the extended metabolic capabilities and larger gene repertoires of extant members relative to other chlamydiae[27,28,66]. Our results provide evidence that these key genes were not present in LCCA as expected, but were instead gained later through HGT leading to the characteristic Amoebachlamydiales metabolic complexity. Although we cannot accurately reconstruct the evolution of all genes, such as those gone extinct or rapidly evolving, it is possible to investigate general patterns in relative reconstructed proteome sizes. Our results indicate a shift towards larger proteome sizes between LAMCA and LCWPCA relative to other early chlamydial ancestors (Fig. 5a). The upward trend in reconstructed Amoebachlamydiales proteome sizes was corroborated using a gene presence/absence method (Supplementary Fig. 12). However, proteome sizes of extant taxa and in-family ancestors were consistent only when using the gene-tree-aware method (Fig. 5b and Supplementary Fig. 12b). Proteome size was reconstructed as having expanded from 1,691 in LAMCA to 2,560 in LCWPCA, nearly double that inferred for LG2CA (n = 1,408) and indicating genome expansion in Amoebachlamydiales. In contrast, the reconstructed proteome size of the Chlamydiaceae ancestor (n = 1,092) suggests genome reduction. In other early chlamydial ancestors, reconstructed proteome sizes are consistent with genome maintenance (Fig. 5).

Despite a conserved endosymbiotic lifestyle, divergent patterns in genome evolution were found across Chlamydiae. Genome reduction in obligate endosymbionts is associated with vertical transmission and a strict host range. The resulting small intracellular population sizes and genetic isolation lead to gene loss through genetic drift, the accumulation of slightly deleterious mutations and a lack of recombination[9–12]. We observed genome reduction leading to the pathogenic Chlamydiaceae, which are typically horizontally transmitted but have strict animal hosts. Genome maintenance in obligate endosymbionts is associated with horizontal transmission and a wider host range. For example, marine bivalve endosymbionts maintain intermediate genome sizes due to horizontal transmission and recombination[67]. Protists have been referred to as 'melting pots' of evolution, because their endosymbionts tend to undergo less genome reduction due to HGT with host prey and co-infecting endosymbionts[68–70]. In most chlamydial lineages we observe patterns consistent with genome maintenance, although we lack information about host ranges and transmission. Genome reduction and maintenance are in line with previous work on the effects of transmission mode and host variability on endosymbiont genome evolution[71,72]. While genome expansion has been shown in facultative symbionts with free-living lifestages[73], it has not previously been shown in obligate endosymbionts. We observed extensive gene gain as having led to the larger proteome sizes and increased metabolic complexity characteristic of extant protist-infecting Amoebachlamydiales. Our findings challenge existing paradigms by providing evidence that obligate endosymbionts can counter genome reduction processes and undergo genome expansion. Given that Chlamydiae is an ancient endosymbiotic phylum, we suggest that endosymbiont genomic and metabolic complexity can increase over long evolutionary time scales.

Gene exchange is common among chlamydiae

In our analysis, 70% of PVC gene families were found to have evolved vertically, closely mirroring previous reconstructions of bacterial evolution[74]. The remaining 30% represent horizontal gene transmission from within (that is, transfers) or outside the PVC dataset (that is, HGT-derived originations). HGT is known to occur in both horizontally and vertically transmitted endosymbionts (for example, *Wolbachia*), although it is more prevalent in the former[67,71,75–79]. Nevertheless, obligate endosymbionts are expected to be limited in HGT relative to free-living bacteria[9]. A large number of gene originations were reconstructed in Chlamydiae ($n = 1{,}458$), many of which were probably HGTs from diverse bacterial groups (Extended Data Fig. 4 and Supplementary Data 11). For example, chlamydial sequences affiliate with bacterial groups for 94% of LCCA HGT-derived originations. For many gene originations, chlamydial sequences affiliate with taxonomic groups known for host association, such as Proteobacteria, Bacteroides and CPR bacteria[80]. This pattern is suggestive of HGT facilitated by co-occurring symbionts or phagocytosed prey bacteria, which is common in protists[68–70,81]. We further examined putative HGTs within the PVC dataset (that is, transfer events) and approximated gene transfer rates between ancestors. In Chlamydiae we found lower, but not statistically significant ($P = 0.068$), transfer rates than in other PVC bacteria (Extended Data Fig. 7). Between chlamydial families, Parachlamydiaceae and Rhabdochlamydiaceae had significantly higher transfer rates than Chlamydiaceae ($P = 4.8 \times 10^{-3}$ and 8.5×10^{-3}) and Anoxychlamydiaceae ($P = 7.1 \times 10^{-3}$ and 1.1×10^{-3}). Overall, gene exchange rates in some chlamydial groups do not differ from free-living PVC bacteria.

Interchlamydial HGT was visualized by testing for statistically over-represented gene transfers between chlamydial nodes where donor and acceptor lineages could be assigned ($n = 5{,}937$, $P \leq 0.05$). The resulting network reveals gene transfer highways indicating probable shared environmental niches, such as shared hosts (Extended Data Fig. 7). Genome sequence divergence is a major barrier to HGT[82]. Despite this, 59% ($n = 3{,}493$, $P \leq 0.05$) of significant transfers occured between members of different chlamydial families. Elevated HGT frequency between more distantly related chlamydiae could be explained by ecological overlap in host or environment[82,83]. Chlamydiaceae, Anoxychlamydiaceae and *Neochlamydia* were under-represented in the network and have isolated transfer highways. These groups have convergently lost central metabolic pathways, including TCA cycle and ETC components (Fig. 4, Extended Data Figs. 5 and 6 and Supplementary Data 10), suggesting adaptation to specialized environments. Extensive gene transfer highways between Amoebachlamydiales members also support divergent genome evolution in this group (Extended Data Fig. 7). Chlamydial gene exchange could be facilitated by the presence of ancestral plasmids, which encode conjugative elements in some lineages including several Parachlamydiaceae[41]. While HGT is well recognized in endosymbiotic bacteria[67,71,75–79], our study provides a systematic view on the pervasiveness of HGT and the role of intersymbiont transfers in endosymbiont genome evolution.

Conclusions

In our study we present a comprehensive view of evolution in the Chlamydiae phylum. We found that the Chlamydiae ancestor was already adapted to an endosymbiotic lifestyle and probably infected eukaryotic hosts. We also found that Chlamydiae did not evolve from a metabolically versatile aerobe as expected but rather from a facultative anaerobe. Energy metabolism and oxygen tolerance gene gain later shaped diversification within the phylum. Counter to expectations for obligate endosymbionts, our results show that the protist-infecting Amoebachlamydiales underwent genome expansion and only later gained their characteristic aerobic and metabolic versatility. Together, our results lay a foundation for further investigation of the complex, and perhaps varied, evolutionary trajectories of bacterial endosymbionts.

Methods

See Extended Data Fig. 1 for an overview of key steps for the reconstruction of gene content evolution in PVC bacteria.

Selection of representative genomes

A representative dataset of PVC bacteria genomes was selected using genome quality to obtain species-level Chlamydiae representatives and genus-level representatives of other PVC bacteria from the genome taxonomy database (GTDB) and the National Center for Biotechnology Information (NCBI). GTDB is continually updated as genomes are released on NCBI and thus naming structures are nonstationary[84]. Here Chlamydiae were initially classified as a phylum, but in the version used were classified as a class of Verrucomicrobiota (that is, Chlamydiia). All genomes from GTDB v.86 (2018 database) classified as Planctomycetota and Verrucomicrobiota were selected ($n = 1{,}183$). Non-chlamydial PVC genomes ($n = 773$; Supplementary Data 1) with completeness ≥90% and contamination ≤2%, based on GTDB metadata, were downloaded from NCBI ($n = 182$; 3 April 2019). For Chlamydiae, genomes from GTDB class 'c__Chlamydiia' were downloaded from NCBI ($n = 410$; 3 April 2019) and supplemented with recently acquired MAGs and isolate genomes ($n = 216$) for a total of 626 chlamydial genomes. We used miComplete[85] v.1.1.1 to estimate the quality of chlamydiae genomes using a specific marker gene set[37] and selected those with completeness ≥0.9 and redundancy ≤1.02 for downstream analysis ($n = 460$; Supplementary Data 1).

To reduce dataset redundancy, all genomes were dereplicated with dRep[86] v.1.4.3 using previously proposed cutoffs for strain-level delineation[87]—that is, an average nucleotide identity of 96.5% and genome alignment fraction of at least 60%, resulting in 224 genomes (Supplementary Fig. 1 and Supplementary Data 1). Non-chlamydial PVC genomes were further dereplicated by comparing genome quality scores (GQS) per GTDB genus level (Supplementary Data 1). GQS was calculated as described in ref. 88—that is, GQS = completeness (%) − 5 × contamination (%). The highest GQS-scoring genome per genus was selected as a representative and, when two genomes had an equal score one was manually selected (Supplementary Fig. 1). The final dataset included 184 PVC genomes with 95 species-level Chlamydiae representatives and 89 genus-level non-Chlamydiae PVC representatives (47 Planctomycetes, 34 Verrucomicrobia, 5 Lentisphaerae and 3 Kiritimatiellaeota) (Supplementary Fig. 1 and Supplementary Data 2). Genome characteristics were calculated using miComplete[85] v.1.1.1 (Fig. 1 and Supplementary Data 2). Putative uncharacterized PVC phyla were not included, such as *Candidatus* Omnitrophica[16] due to its conflicting position in some large-scale species trees of Bacteria[15,88].

Phylogenomic analyses

PVC species relationships were inferred using phylogenomic datasets of concatenated single-copy marker genes (Supplementary Fig. 2 and Supplementary Data 3) for the initial 184 selected taxa (Supplementary Fig. 3 and Supplementary Data 2). Additional species phylogenies were inferred after removal of genomes with unresolved phylogenetic positions, resulting in datasets with 183 taxa (removal of *Chlamydiae* bacterium 1070360-7; Supplementary Fig. 4) and 180 taxa (further removal of the 3 Parilichlamydiaceae genomes; Supplementary Fig. 5). ML and Bayesian phylogenies were inferred with and without the removal of compositionally heterogeneous sites for all three datasets (184, 183 and 180 taxa) as outlined below (Supplementary Data 4–6). Species phylogenies were rooted with Planctomycetes based on its phylogenetic position in recent large-scale phylogenomic analyses of bacterial species relationships[88–90]. The 180-taxa dataset was selected for further analyses and the converged Bayesian species phylogeny (consensus of chains 1 and 3), with compositionally heterogeneous sites removed, was used for ancestral state reconstruction (Fig. 1 and Supplementary Discussion 1).

Identification of single-copy marker genes. Candidate single-copy marker genes were identified using nonsupervised orthologous groups (NOGs) from eggNOG[91] v.4.5.1. Protein-coding gene sequences from all PVC bacteria genomes were mapped to NOGs at the last universal common ancestor level (that is, root-level '-d NOG') using emapper[92] v.1.0.1. Resulting NOGs where 95% of taxa were found in a single copy were identified as candidate markers for further investigation ($n = 116$; Supplementary Data 3).

Sequences from each gene family NOG were aligned using MAFFT L-INS-i[93] v.7.427 and manually inspected, with poorly aligned and short sequences removed. Alignments were trimmed using BMGE[94] v.1.12 (entropy score cutoff or '-h' of 0.6). IQ-TREE[95] v.1.6.11 was used to infer phylogenetic trees, with model selection by ModelFinder[96] from empirical profile mixture models[97] combined with the LG exchangeability matrix[98] (that is, LG + C10 to LG + C60), and with 1,000 ultrafast bootstraps (ufBP)[99]. Resulting trees (available in repository) were manually examined for patterns indicative of vertical inheritance and sufficient phylogenetic signal, and markers were removed that did not generally resolve PVC phyla (Supplementary Data 3). Sequences were removed that could represent HGT events, distant paralogues or contamination (Supplementary Data 3). Where multiple sequences per taxon were present, if they overlapped both were removed (duplicates) and, if they were partial and nonoverlapping, the longer sequence was retained (Supplementary Data 3). A second round of sequence alignment and tree inference was performed as above, with further markers removed resulting in 79 marker genes (Supplementary Data 3).

Discordance filtering[100] was then performed to remove markers with the most anomalous phylogenetic signal relative to the majority (that is, the most discordant trees). NOGs (all of which were clusters of orthologous groups, that is, COGs) were ranked by discordance score and the top-scoring fraction was removed, leaving 74 single-copy marker genes for phylogenomic analyses (Supplementary Fig. 2). Amino acid sequences for each selected marker gene were realigned and trimmed, as above, after removal of taxa with unresolved phylogenetic positions (that is, datasets with 183 and 180 taxa). Trimmed amino acid alignments were concatenated into a supermatrix for each of the three datasets.

Phylogenomic inferences. Heterogenous site removal was performed using χ^2-trimming[101], with the most compositionally heterogeneous sites removed from each concatenated alignment in incremental steps of 1% of alignment sites. Site removal continued until no taxa significantly heterogeneous in their amino acid composition remained (based on the χ^2 test score statistic; significance $P \leq 0.05$; Supplementary Figs. 3–5 and Supplementary Data 4).

Using IQ-TREE[95] v.1.6.10 with model selection[96], ML phylogenies were inferred for the initial unrefined alignment, for alignments in 10% increments of total sites removed based on χ^2-trimming (up to 50%; Supplementary Data 4), and for the alignment with no significantly heterogeneous taxa (fully refined alignment). ML trees were then reconstructed using the posterior mean site frequency (PMSF) approximation of the LG + C60 + F + Γ4 model (selected in all initial trees) with 100 nonparametric bootstraps. Transfer bootstrap expectation[102] bipartition support was also inferred for the initial unrefined alignment and for the fully refined alignment using IQ-TREE[103] v.2.0.

Bayesian phylogenies were reconstructed for these two alignments for all three taxa datasets. In each case, four independent Markov chain Monte Carlo chains were run using PhyloBayes-MPI v.1.7b[104] with the CAT + GTR + Γ4 model[97,105], for at least 10,000 iterations. CAT, a site-heterogeneous model, performs more robustly against long-branch attraction artefacts[106]. If at least 10,000 iterations had been run but no chains had begun to converge (maximum difference <1), all chains were stopped. The number of generations, burn-in and any chain convergence (maximum difference <0.3) can be found in Supplementary Figs. 3–5 alongside a consensus tree of all four chains with posterior probability (P) indicating branch support. Posterior predictive checks were also performed with PhyloBayes-MPI v.1.7b[104], with configurations sampled every ten generations after burn-in. The resulting range of z-scores for maximum heterogeneity and diversity across chains can be found in Supplementary Figs. 3–5. See Supplementary Data 6 for all uncollapsed species phylogenies and Supplementary Data 5 for a summary of the number of taxa, alignment lengths, inference methods, bootstrap supports and model of evolution for each phylogeny.

16S rRNA gene species phylogeny

Near-full-length 16S rRNA gene sequences from chlamydiae ($n = 233$) and other PVC members ($n = 205$) were downloaded from SILVA[107] v.138 SSU Ref NR 99, 79 near-full-length chlamydial 16S rRNA gene sequences (97% identity operational taxonomic unit representatives) retrieved from Schulz et al.[22] and 142 sequences from our reference genome dataset. Sequences ($n = 659$) were clustered at 90% sequence identity to reduce redundancy using USEARCH[108] v.11.0.667 with '-cluster_smallmem'. The resulting family-level sequence representatives ($n = 177$) were aligned with SINA[109] and the alignment trimmed with trimAl[110] v.1.4.1 '-gappyout' (1,533 aligned positions). Bayesian tree samples with four Markov chain Monte Carlo chains in parallel ($n = 100,000$ each) were inferred under the CAT + GTR + Γ4 model[97,105] in PhyloBayes v.4.1c[111] (Supplementary Fig. 6). Convergence was assumed once maximum difference dropped below 0.3 and effective sample sizes for continuous parameters were >100 (according to the commands 'bpcomp' and 'tracecomp' in PhyloBayes, respectively) after burn-in ($n = 25,000$).

Generation of gene families and trees

NOG clustering. PVC gene sequences from the 180-taxa dataset ($n = 445,591$) were mapped against eggNOG[91] v.4.5.1 using emapper[92] v.1.0.1 at root-level '-d NOG'. Of these, 326,083 (73%) gene sequences were assigned to 17,935 NOGs (Supplementary Fig. 7).

De novo clustering. For the remaining 119,508 gene sequences in the 180-taxa dataset with no homologue in eggNOG v.4.5.1, we performed pairwise sequence alignment in an all-against-all fashion with DIAMOND[112] v.0.9.21 using the parameter '–more-sensitive'. Subsequently, de novo clustering with SiLiX[113] v.1.2.9 was performed with default overlap of 80% and identity thresholds ranging from 5 to 40% in 5% increments (Supplementary Fig. 7). To select an appropriate identity threshold we (1) inspected the number of singleton clusters per threshold and (2) assigned TIGRFAM[114] v.15.0 domains with InterProScan[115] v.5.36-75.0 to gene sequences. Using the assigned TIGRFAMs, true positive rate (sensitivity) and true negative rate (specificity) were calculated for clusters, with different clusterings evaluated using the balanced accuracy measure ((specificity + sensitivity)/2) as suggested[113]. A 25% identity cutoff performed best, yielding 10,548 de novo gene families with at least two members (75,218 singletons).

Gene family phylogenetic trees. We performed phylogenetic analysis on all gene families (both NOG and de novo clusters) with at least four members (n = 11,996). Sequences were aligned with MAFFT[116] v.7.427 using the strategy '–localpair'. Alignments were then trimmed using BMGE[94] v.1.12 with default parameters and an entropy cutoff of 0.6. The permitted gap rate for alignment positions was increased to 0.5 for 94 gene families with <50 informative aligned positions using the initial parameters. Gene trees were then inferred with IQ-TREE[95] v.1.6.11 using the best-fit model identified by ModelFinder[96], with '-m TESTNEW', '-madd LG + C10, LG + C20, LG + C30, LG + C40, LG + C50, LG + C60' and 1,000 improved ufBPs[99] '-bnni'. Two specific gene families were later excluded (COG3119 and COG0457) due to poor alignment (probably caused by repeat regions) and subsequent difficulties inferring a ML tree, bootstraps from which are required for ALE. These gene families primarily occur in non-Chlamydiae PVC members and thus should not impact chlamydial ancestor proteome sizes (Supplementary Data 7). The remaining 91,705 gene families had three or fewer sequences. Thus, phylogenetic trees could not be inferred. However, because ALE requires gene families in a tree format we thus provided mock trees for those with two or three sequences.

Annotation of gene families. We assigned protein domain annotations to gene families using InterProScan[115] v.5.36-75.0 to identify the domains of Protein Families (Pfam)[117], TIGRFAM[114] and InterPro (IPR)[118]. We assigned Kyoto Encyclopedia of Genes and Genomes orthology and enzyme commision numbers using GhostKOALA[119] and inferred eggNOG[91] functional annotation as described above, and also at the bacterial level ('-d BACT').

Ancestral state reconstruction

To gain a more complete perspective on PVC genome evolution we used the complete genome dataset outlined above, which includes MAGs from uncultured lineages that would otherwise be missed. However, this was restricted to high-quality MAGs as outlined above (completeness ≥90% and contamination ≤2%).

Gene-tree-unaware method, Count. For gene-tree-unaware ancestral gene content reconstruction, we ran Count[120] v.10.04 with the gain–loss–duplication model of evolution with Poisson distribution to model gene family size at the root. We used the same gain–loss and duplication–loss ratios for all lineages and inferred ancestral gene content using the Wagner maximum parsimony framework with default costs.

Gene-tree-aware method, ALE. ML tree bootstrap samples of gene families identified in the PVC dataset were reconciled with the species tree to reconstruct their gene family histories. We computed conditional clade probabilities from bootstrap samples (ALEobserve) and sampled 100 reconciliations with the species tree (ALEml_undated) using ALE[121] v.0.449, implemented as a computational pipeline (https://github.com/maxemil/ALE-pipeline). We added singleton gene families as originations at the corresponding species node to the reconstructions. Furthermore, the estimated fraction of missing gene content per genome was provided to ALE because it uses this to correct for potentially missing data—that is, in the MAGs included. In addition, the specific implementation of ALE corrects ancestor gene copy number estimates using modelled gene extinction probability rates[44], which has previously been employed to estimate ancestral proteome sizes[122].

Comparison of ancestral state reconstruction methods and selection of ALE cutoff. ALE improves on earlier methods by direct estimation of rates of gene duplication, transfer and loss from data, as well as incorporating the uncertainty in gene trees while exploring a larger gene tree space[121]. The accuracy of reconstructions can be negatively influenced by an inaccurate species tree and imbalanced taxon sampling[44,121]. Here, these risks are minimized due to our extensive taxon sampling and species tree reconstruction efforts (Supplementary Figs. 1–6 and Supplementary Data 1–6).

ALE reports relative frequencies for ancestral events and gene family copy frequencies that express their statistical support. This support accumulates the uncertainty introduced by alignment, tree reconstruction and reconciliation and should therefore not be set at a standard-level cutoff. We therefore aimed to identify a suitable threshold by investigating density distribution per inferred event type and transfer ratio per gene family (Supplementary Fig. 8), which indicated a cutoff of 0.3 for a candidate with high signal-to-noise ratio. The identified cutoff is in accordance with recent similar analyses that selected 0.3 as a frequency cutoff[123,124]. The transfer ratio represents the proportion of horizontal events over all events per gene tree. We further compared the tree-aware reconstructions generated by ALEml with the thresholds 0.3 (sensitive), 0.5 (specific) and 0.7 (very specific) using gene-content-only-aware Count reconstructions (Supplementary Fig. 9). The highest number of consensus gene families obtained with gene-tree-aware and -unaware methods was reached with a threshold of 0.3 (Supplementary Fig. 9a). Based on this analysis and event density distributions, we selected a frequency cutoff of 0.3 for inferring evolutionary events in our ancestral state reconstruction analysis (Supplementary Figs. 8 and 9). Gene families were thus inferred as present when reconstructing ancestor gene content if they had a copy frequency of at least 0.3. In addition, we used this cutoff to calculate confident-event frequencies. This meant that an event frequency ≥0.3 and <1.3 was counted as 1, a frequency ≥1.3 and <2.3 was counted as 2 and so on. These confident-event frequencies correspond to the gene content and events used in our ancestral reconstructions (Figs. 2 and 4, Supplementary Figs. 10–12 and Extended Data Figs. 2–6). However, for estimation of genome content evolution dynamics and ancestral proteome size, raw reconciliation frequencies were used to avoid potential underestimation of transfers and losses (Figs. 3 and 5 and Extended Data Fig. 7).

Inference of transfer rates and gene transfer highways

We approximated the rate of intra-PVC HGT (that is, transfer rate) for nodes in the species tree by calculating the inferred gene transfers in our reconstructions divided by the number of substitutions in the species tree along the given branch. Based on shared gene families between two extant or ancestral genomes, we tested whether more HGT events occured between genomes than the median of transferred gene families within chlamydiae members. We used a one-sided binomial test ('binom.test') with 'alternative = greater' in the R base package[125], and false discovery rate corrected P values for multiple testing with 'p.adjust' to identify enriched transfer routes ('gene transfer highways') with P ≤0.05. Significant gene transfer highways were visualized with Cytoscape[126] v.3.7.0.

Identification of non-PVC gene transfer donors

For genes inferred as originations within Chlamydiae, to distinguish bona fide HGTs from outside the PVC dataset and candidate de novo gene families, we performed a homology search against the NCBI nonredundant database. If no homologous protein sequences could be identified, gene families were referred to as de novo candidates, otherwise we inferred gene trees to identify affiliated taxonomic groups and thus potential donor lineages of the horizontally transferred gene (workflow: https://github.com/jennahd/HGT_trees). For each gene family a DIAMOND v.0.9.36.137 blastp[112] search (with 'max-target-seqs 2000' and 'more-sensitive') was performed using all sequences against NCBI's nonredundant database[127] (v.5, accessed 8 October 2020). Unique hits per gene family were compiled and clustered using CD-HIT[128] v.4.8.1 at 80% sequence identity. NCBI's taxonomy database[129] was used for taxonomic classification. Protein sequences from each gene family and any database hits were aligned with MAFFT[93] v.7.471 ('–auto') and trimmed with trimAl[110] v.1.4.rev15 ('gappyout'). Sequences that covered <40% of the trimmed alignment were removed, followed by inference of an initial phylogenetic tree using FastTree[130] 2 v.2.1.11. Long-branching taxa were identified as having outlier terminal branch lengths (third quartile + 1.5× interquartile range) relative to others in each tree, and were removed before reinferring trees as above.

These initial trees were prohibitively large for performing ML analyses and smaller subtrees were therefore selected using the above workflow. Here, clades comprising ≥25% chlamydiae with at least two chlamydial sequences were identified. To account for multiple HGT events, per gene family up to three clades with the largest number of chlamydial sequences were identified although, in the majority of cases, only one was found. Subtrees including these clades were selected by finding nodes at least three further up the tree hierarchy and that included at least 150 additional taxa and up to 400 additional taxa with bipartition support ≥0.7. Where a subtree fulfilling these conditions was identified, but with a larger number of taxa, the number of taxa was reduced to ≤400 by removal of more distant sequences (support ≥0.7, at least five sequences and at least six steps until a common ancestor with the chlamydial clade). Between 20 and 50 outgroup sequences were randomly selected from the clade with a position sister to the selected subtree (moving to the next subtending clade when there were <20 initial outgroup sequences). Selected subtrees were subsequently aligned and trimmed and an initial phylogenetic tree was inferred as above. ML trees were then inferred for each subtree using the trimmed alignment with IQ-TREE[95] v.1.6.12 under the LG model of evolution[98], with 1,000 ufBP. The clade sister to chlamydial sequences, and that subtending this clade ('nested') with ufBP ≥80, were identified. Taxonomic labels of sister and nested taxa were each compared at domain, superphylum and phylum levels. The lowest-level shared taxonomy at cutoffs of 75% (Extended Data Fig. 7), 90% and 100% of taxa was selected as the affiliated taxonomic group and hence putative gene donor (Supplementary Data 11). Originations were identified as de novo in the case where no non-chlamydial hits were found or when no sister group to chlamydial sequences was inferred.

Reconstruction of metabolic pathways

The evolutionary history of ETC components was investigated using the workflow described above. We reconstructed ancestral gene family repertoires from ALE by selecting all families predicted to be present at a given node with a relative frequency ≥0.3. We assessed the metabolic capabilities of ancestral genomes using either the Kyoto Encyclopedia of Genes and Genomes Module tool[131] or MetaCyc pathways[132].

Statistics and data visualization

Phylogenetic trees and protein domains were visualized using Figtree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree), iTOL[132] and the ETE3 Toolkit[133] v.3.1.2. Relative evolutionary divergence of chlamydiae ancestors in the species tree was calculated using PhyloRank[84] v.0.1.10 (https://github.com/dparks1134/PhyloRank). Plots were generated using Cytoscape[126] v.3.7.0 and the R v.3.6.2 base package[125] alongside the packages ggplot2 (ref. 134) v.3.3.3, ggtree[135] v.2.5.0.991 and treeio[136] v.1.10.0.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Genome data was obtained either from NCBI Genbank (https://www.ncbi.nlm.nih.gov/nucleotide), the JGI portal (https://portal.nersc.gov/GEM) or a zenodo repository (https://doi.org/10.5281/zenodo.4318714). 16S rRNA gene data used in this study are available via the SILVA database (https://www.arb-silva.de). Genbank accessions and database links for genomes used in the ancestral state reconstruction are provided in Supplementary Data 2. Additional raw data files are hosted on the online repository figshare (https://doi.org/10.6084/m9.figshare.17033417). These include sequences, alignments, trimmed alignments and trees for single-copy marker genes used for species phylogenies (both those selected and not selected), the 16S rRNA gene alignment and tree, as well as concatenated alignments and trees for all three species datasets (of 184, 183 and 180 taxa). Both NOG and de novo gene families used for the ancestral state reconstruction are also provided along with alignments, trimmed alignments, trees and bootstrap trees (ufboot) provided to ALE. The raw ALE results with all events are also included, along with gene annotations and events, and events for each gene family mapped to the species tree. Protein sequence datasets, alignments and trees inferred as part of the analysis to determine HGT donors for chlamydiae gene originations are provided. In addition, PDFs of metabolic reconstructions of LVCCA, LG1CA and LG2CA can be found in the repository files.

Code availability

The Nextflow pipeline for running the ALE ancestral state reconstruction (https://github.com/maxemil/ALE-pipeline) and the Snakemake pipeline for identification of putative HGT donor groups (https://github.com/jennahd/HGT_trees) are available on GitHub.

References

1. De Bary, A. *Die Erscheinung der Symbiose* (De Gruyter, 1879).
2. Drew, G. C., Stevens, E. J. & King, K. C. Microbial evolution and transitions along the parasite-mutualist continuum. *Nat. Rev. Microbiol.* **19**, 623–638 (2021).
3. McFall-Ngai, M. et al. Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl Acad. Sci. USA* **110**, 3229–3236 (2013).
4. Husnik, F. et al. Bacterial and archaeal symbioses with protists. *Curr. Biol.* **31**, R862–R877 (2021).
5. Sogin, E. M., Maggie Sogin, E., Leisch, N. & Dubilier, N. Chemosynthetic symbioses. *Curr. Biol.* **30**, R1137–R1142 (2020).
6. López-García, P., Eme, L. & Moreira, D. Symbiosis in eukaryotic evolution. *J. Theor. Biol.* **434**, 20–33 (2017).
7. Wilkins, L. G. E. et al. Host-associated microbiomes drive structure and function of marine ecosystems. *PLoS Biol.* **17**, e3000533 (2019).
8. McLean, A. H. C., Parker, B. J., Hrček, J., Henry, L. M. & Godfray, H. C. J. Insect symbionts in food webs. *Phil. Trans. R. Soc. B* **371**, 20150325 (2016).
9. Moran, N. A. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**, 583–586 (2002).
10. Wernegreen, J. J. For better or worse: genomic consequences of intracellular mutualism and parasitism. *Curr. Opin. Genet. Dev.* **15**, 572–583 (2005).
11. Toft, C. & Andersson, S. G. E. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat. Rev. Genet.* **11**, 465–475 (2010).
12. Moran, N. A. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA* **93**, 2873–2878 (1996).
13. McCutcheon, J. P., Boyd, B. M. & Dale, C. The life of an insect endosymbiont from the cradle to the grave. *Curr. Biol.* **29**, R485–R495 (2019).

14. Rivas-Marín, E. & Devos, D. P. The paradigms they are a-changin': past, present and future of PVC bacteria research. *Antonie Van Leeuwenhoek* **111**, 785–799 (2018).
15. Collingro, A., Köstlbacher, S. & Horn, M. Chlamydiae in the environment. *Trends Microbiol.* **28**, 877–888 (2020).
16. Wagner, M. & Horn, M. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr. Opin. Biotechnol.* **17**, 241–249 (2006).
17. van Niftrik, L. & Devos, D. P. Editorial: Planctomycetes-Verrucomicrobia-Chlamydiae bacterial superphylum: new model organisms for evolutionary cell biology. *Front. Microbiol.* **8**, 1458 (2017).
18. Bachmann, N. L., Polkinghorne, A. & Timms, P. *Chlamydia* genomics: providing novel insights into chlamydial biology. *Trends Microbiol.* **22**, 464–472 (2014).
19. Elwell, C., Mirrashidi, K. & Engel, J. *Chlamydia* cell biology and pathogenesis. *Nat. Rev. Microbiol.* **14**, 385–400 (2016).
20. Borel, N., Polkinghorne, A. & Pospischil, A. A review on chlamydial diseases in animals: still a challenge for pathologists? *Vet. Pathol.* **55**, 374–390 (2018).
21. Lagkouvardos, I. et al. Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. *ISME J.* **8**, 115–125 (2014).
22. Schulz, F. et al. Towards a balanced view of the bacterial tree of life. *Microbiome* **5**, 140 (2017).
23. Horn, M. Chlamydiae as symbionts in eukaryotes. *Annu. Rev. Microbiol.* **62**, 113–131 (2008).
24. Taylor-Brown, A., Vaughan, L., Greub, G., Timms, P. & Polkinghorne, A. Twenty years of research into *Chlamydia*-like organisms: a revolution in our understanding of the biology and pathogenicity of members of the phylum Chlamydiae. *Pathog. Dis.* **73**, 1–15 (2015).
25. Ishida, K. et al. Amoebal endosymbiont *Neochlamydia* genome sequence illuminates the bacterial role in the defense of the host amoebae against *Legionella pneumophila*. *PLoS ONE* **9**, e95166 (2014).
26. Arthofer, P., Delafont, V., Willemsen, A., Panhölzl, F. & Horn, M. Defensive symbiosis against giant viruses in amoebae. *Proc. Natl Acad. Sci. USA* **119**, e2205856119 (2022).
27. Collingro, A. et al. Unity in variety—the pan-genome of the Chlamydiae. *Mol. Biol. Evol.* **28**, 3253–3270 (2011).
28. Omsland, A., Sixt, B. S., Horn, M. & Hackstadt, T. Chlamydial metabolism revisited: interspecies metabolic variability and developmental stage-specific physiologic activities. *FEMS Microbiol. Rev.* **38**, 779–801 (2014).
29. Kamneva, O. K., Knight, S. J., Liberles, D. A. & Ward, N. L. Analysis of genome content evolution in PVC bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. *Genome Biol. Evol.* **4**, 1375–1390 (2012).
30. Betts, H. C. et al. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556–1562 (2018).
31. Greub, G. & Raoult, D. History of the ADP/ATP-translocase-encoding gene, a parasitism gene transferred from a Chlamydiales ancestor to plants 1 billion years ago. *Appl. Environ. Microbiol.* **69**, 5530–5535 (2003).
32. Horn, M. et al. Illuminating the evolutionary history of chlamydiae. *Science* **304**, 728–730 (2004).
33. Subtil, A., Collingro, A. & Horn, M. Tracing the primordial Chlamydiae: extinct parasites of plants? *Trends Plant Sci.* **19**, 36–43 (2014).
34. Taylor-Brown, A., Spang, L., Borel, N. & Polkinghorne, A. Culture-independent metagenomics supports discovery of uncultivable bacteria within the genus *Chlamydia*. *Sci. Rep.* **7**, 10661 (2017).
35. Taylor-Brown, A., Madden, D. & Polkinghorne, A. Culture-independent approaches to chlamydial genomics. *Micro. Genom.* **4**, e000145 (2018).
36. Pillonel, T., Bertelli, C. & Greub, G. Environmental metagenomic assemblies reveal seven new highly divergent chlamydial lineages and hallmarks of a conserved intracellular lifestyle. *Front. Microbiol.* **9**, 79 (2018).
37. Dharamshi, J. E. et al. Marine sediments illuminate Chlamydiae diversity and evolution. *Curr. Biol.* **30**, 1032–1048.e7 (2020).
38. Collingro, A. et al. Unexpected genomic features in widespread intracellular bacteria: evidence for motility of marine chlamydiae. *ISME J.* **11**, 2334–2344 (2017).
39. Köstlbacher, S. et al. Pangenomics reveals alternative environmental lifestyles among chlamydiae. *Nat. Commun.* **12**, 4021 (2021).
40. Stairs, C. W. et al. Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. *Sci. Adv.* **6**, eabb7258 (2020).
41. Köstlbacher, S., Collingro, A., Halter, T., Domman, D. & Horn, M. Coevolving plasmids drive gene flow and genome plasticity in host-associated intracellular bacteria. *Curr. Biol.* **31**, 346–357 (2021).
42. Dharamshi, J. E. et al. Genomic diversity and biosynthetic capabilities of sponge-associated chlamydiae. *ISME J.* https://doi.org/10.1038/s41396-022-01305-9 (2022).
43. Gupta, R. S., Naushad, S., Chokshi, C., Griffiths, E. & Adeolu, M. A phylogenomic and molecular markers based analysis of the phylum Chlamydiae: proposal to divide the class Chlamydiia into two orders, Chlamydiales and Parachlamydiales ord. nov., and amended description of the class Chlamydiia. *Antonie Van Leeuwenhoek* **108**, 765–781 (2015).
44. Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).
45. Peters, J., Wilson, D. P., Myers, G., Timms, P. & Bavoil, P. M. Type III secretion à la *Chlamydia*. *Trends Microbiol.* **15**, 241–251 (2007).
46. Aistleitner, K. et al. Conserved features and major differences in the outer membrane protein composition of chlamydiae. *Environ. Microbiol.* **17**, 1397–1413 (2015).
47. Gehre, L. et al. Sequestration of host metabolism by an intracellular pathogen. *eLife* **5**, e12552 (2016).
48. Rosario, C. J., Hanson, B. R. & Tan, M. The transcriptional repressor EUO regulates both subsets of *Chlamydia* late genes. *Mol. Microbiol.* **94**, 888–897 (2014).
49. Rosario, C. J. & Tan, M. The early gene product EUO is a transcriptional repressor that selectively regulates promoters of *Chlamydia* late genes. *Mol. Microbiol.* **84**, 1097–1107 (2012).
50. Passalacqua, K. D., Charbonneau, M.-E. & O'Riordan, M. X. D. Bacterial metabolism shapes the host-pathogen interface. *Microbiol. Spectr.* **4**, 10.1128/microbiolspec.VMBF-0027-2015 (2016).
51. Wiegand, S. et al. Cultivation and functional characterization of 79 planctomycetes uncovers their unique biology. *Nat. Microbiol.* **5**, 126–140 (2020).
52. Giuffrè, A., Borisov, V. B., Arese, M., Sarti, P. & Forte, E. Cytochrome bd oxidase and bacterial tolerance to oxidative and nitrosative stress. *Biochim. Biophys. Acta* **1837**, 1178–1187 (2014).
53. Buschmann, S. et al. The structure of cbb3 cytochrome oxidase provides insights into proton pumping. *Science* **329**, 327–330 (2010).
54. Liang, P. et al. Dynamic energy dependency of *Chlamydia trachomatis* on host cell metabolism during intracellular growth: role of sodium-based energetics in chlamydial ATP generation. *J. Biol. Chem.* **293**, 510–522 (2018).
55. Kuhns, M., Trifunović, D., Huber, H. & Müller, V. The Rnf complex is a Na^+ coupled respiratory enzyme in a fermenting bacterium, *Thermotoga maritima*. *Commun. Biol.* **3**, 431 (2020).

56. Torrents, E. Ribonucleotide reductases: essential enzymes for bacterial life. *Front. Cell. Infect. Microbiol.* **4**, 52 (2014).
57. Jiang, W. et al. A manganese(IV)/iron(III) cofactor in *Chlamydia trachomatis* ribonucleotide reductase. *Science* **316**, 1188–1191 (2007).
58. Lau, C. K. Y., Krewulak, K. D. & Vogel, H. J. Bacterial ferrous iron transport: the Feo system. *FEMS Microbiol. Rev.* **40**, 273–298 (2016).
59. Miethke, M. Molecular strategies of microbial iron assimilation: from high-affinity complexes to cofactor assembly systems. *Metallomics* **5**, 15–28 (2013).
60. Sawers, G. & Böck, A. Anaerobic regulation of pyruvate formate-lyase from *Escherichia coli* K-12. *J. Bacteriol.* **170**, 5330–5336 (1988).
61. Tomov, A. T., Tsvetkova, E. D., Tomova, I. A., Michailova, L. I. & Kassovski, V. K. Persistence and multiplication of obligate anaerobe bacteria in amebae under aerobic conditions. *Anaerobe* **5**, 19–23 (1999).
62. Ahn, S., Jung, J., Jang, I.-A., Madsen, E. L. & Park, W. Role of glyoxylate shunt in oxidative stress response. *J. Biol. Chem.* **291**, 11928–11938 (2016).
63. Mulkidjanian, A. Y., Dibrov, P. & Galperin, M. Y. The past and present of sodium energetics: may the sodium-motive force be with you. *Biochim. Biophys. Acta* **1777**, 985–992 (2008).
64. Pál, C., Papp, B. & Lercher, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* **37**, 1372–1375 (2005).
65. Degli Esposti, M., Mentel, M., Martin, W. & Sousa, F. L. Oxygen reductases in alphaproteobacterial genomes: physiological evolution from low to high oxygen environments. *Front. Microbiol.* **10**, 499 (2019).
66. Bertelli, C. et al. The *Waddlia* genome: a window into chlamydial biology. *PLoS ONE* **5**, e10890 (2010).
67. Russell, S. L. et al. Horizontal transmission and recombination maintain forever young bacterial symbiont genomes. *PLoS Genet.* **16**, e1008935 (2020).
68. Moliner, C., Fournier, P.-E. & Raoult, D. Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. *FEMS Microbiol. Rev.* **34**, 281–294 (2010).
69. Wang, Z. & Wu, M. Comparative genomic analysis of *Acanthamoeba* endosymbionts highlights the role of amoebae as a 'Melting Pot' shaping the Rickettsiales evolution. *Genome Biol. Evol.* **9**, 3214–3224 (2017).
70. Bertelli, C. & Greub, G. Lateral gene exchanges shape the genomes of amoeba-resisting microorganisms. *Front. Cell. Infect. Microbiol.* **2**, 110 (2012).
71. Perreau, J. & Moran, N. A. Genetic innovations in animal–microbe symbioses. *Nat. Rev. Genet.* **23**, 23–39 (2022).
72. Husnik, F. & Keeling, P. J. The fate of obligate endosymbionts: reduction, integration, or extinction. *Curr. Opin. Genet. Dev.* **58–59**, 1–8 (2019).
73. Medina, M. & Sachs, J. L. Symbiont genomics, our new tangled bank. *Genomics* **95**, 129–137 (2010).
74. Coleman, G. A. et al. A rooted phylogeny resolves early bacterial evolution. *Science* **372**, eabe0511 (2021).
75. Nakabachi, A. et al. Horizontal gene acquisition of *Liberibacter* plant pathogens from a bacteriome-confined endosymbiont of their psyllid vector. *PLoS ONE* **8**, e82612 (2013).
76. Bordenstein, S. R. & Wernegreen, J. J. Bacteriophage flux in endosymbionts (*Wolbachia*): infection frequency, lateral transfer, and recombination rates. *Mol. Biol. Evol.* **21**, 1981–1991 (2004).
77. Pinto-Carbó, M. et al. Evidence of horizontal gene transfer between obligate leaf nodule symbionts. *ISME J.* **10**, 2092–2105 (2016).
78. Scholz, M. et al. Large scale genome reconstructions illuminate *Wolbachia* evolution. *Nat. Commun.* **11**, 5235 (2020).
79. Tsai, Y.-M., Chang, A. & Kuo, C.-H. Horizontal gene acquisitions contributed to genome expansion in insect-symbiotic *Spiroplasma clarkii*. *Genome Biol. Evol.* **10**, 1526–1532 (2018).
80. Jaffe, A. L. et al. Patterns of gene content and co-occurrence constrain the evolutionary path toward animal association in candidate phyla radiation bacteria. *mBio* **12**, e0052121 (2021).
81. Lagkouvardos, I., Shen, J. & Horn, M. Improved axenization method reveals complexity of symbiotic associations between Bacteria and Acanthamoebae. *Environ. Microbiol. Rep.* **6**, 383–388 (2014).
82. Popa, O., Hazkani-Covo, E., Landan, G., Martin, W. & Dagan, T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.* **21**, 599–609 (2011).
83. Smillie, C. S. et al. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).
84. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
85. Hugoson, E., Lam, W. T. & Guy, L. miComplete: weighted quality evaluation of assembled microbial genomes. *Bioinformatics* **36**, 936–937 (2020).
86. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
87. Varghese, N. J. et al. Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* **43**, 6761–6771 (2015).
88. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol* **2**, 1533–1542 (2017).
89. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
90. Zhu, Q. et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).
91. Huerta-Cepas, J. et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–D293 (2016).
92. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
93. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
94. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
95. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
96. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
97. Quang, L. S., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).
98. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
99. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

100. Williams, K. P. et al. Phylogeny of Gammaproteobacteria. *J. Bacteriol.* **192**, 2305–2314 (2010).
101. Viklund, J., Ettema, T. J. G. & Andersson, S. G. E. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol. Biol. Evol.* **29**, 599–615 (2012).
102. Lemoine, F. et al. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).
103. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
104. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615 (2013).
105. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
106. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7** (Suppl 1), S4 (2007).
107. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
108. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
109. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).
110. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
111. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
112. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
113. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12**, 116 (2011).
114. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
115. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
116. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).
117. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
118. Blum, M. et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
119. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
120. Csurös, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
121. Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V. & Boussau, B. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Phil. Trans. R. Soc. B* **370**, 20140335 (2015).
122. Williams, T. A. et al. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl Acad. Sci. USA* **114**, E4602–E4611 (2017).
123. Martijn, J. et al. Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nat. Commun.* **11**, 5490 (2020).
124. Huang, W.-C. et al. Comparative genomic analysis reveals metabolic flexibility of Woesearchaeota. *Nat. Commun.* **12**, 5281 (2021).
125. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2018).
126. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
127. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **44**, D7–D19 (2016).
128. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
129. Schoch, C. L. et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, baaa062 (2020).
130. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
131. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
132. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes – a 2019 update. *Nucleic Acids Res.* **48**, D445–D453 (2020).
133. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
134. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
135. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
136. Wang, L.-G. et al. Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* **37**, 599–603 (2020).

Acknowledgements

We thank the IMG/M Data Consortium for contributing metagenomic data, and L. Alteo and J. Blanchard for early access to Harvard forest soil MAGs. We thank D. Tamarit and C. Stairs for advice and discussions. T.J.G.E. received funding from the European Research Council (grant nos. 310039 and 817834), the Dutch Research Council (grant no. VI.C.192.016), the Swedish Research Council (grant no. 2015-04959), Volkswagen Foundation (grant no. 96725) and the European Union (grant no. H2020-MSCA-ITN-2015-675752). M.H. and A.C. received funding from the European Research Council ERC (grant no. 281633) and the Austrian Science Fund FWF (FunChlam, grant no. P32112 and doc.funds MAINTAIN, grant no. DOC 69-B). S.K. received a KWA (Kurzfristige Wissenschaftliche Auslandsstipendien) travel grant from the University of Vienna. Computational resources were provided by the Life Science Compute Cluster (https://lisc.univie.ac.at) at the University of Vienna and the Swedish National Infrastructure for Computing (SNIC) at UPPMAX (project no. SNIC 2020/15-158), and at PDC (project nos. SNIC 2019/3-474 and SNIC 2020/5-473).

Author contributions

T.J.G.E., M.H., J.E.D. and S.K. conceptualized the project. S.K. and J.E.D. were responsible for data curation. J.E.D., S.K., A.C. and M.E.S. conducted formal analysis. S.K. and J.E.D. carried out the investigation. All authors were responsible for validation. T.J.G.E. and M.H. supervised the project. J.E.D. and S.K. were responsible for visulization. S.K. and J.E.D. wrote the original draft. All authors reviewed and edited the article. These authors contributed equally: J.E.D. and S.K., T.J.G.E. and M.H.

Funding

Open access funding provided by Uppsala University.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-022-01284-9.

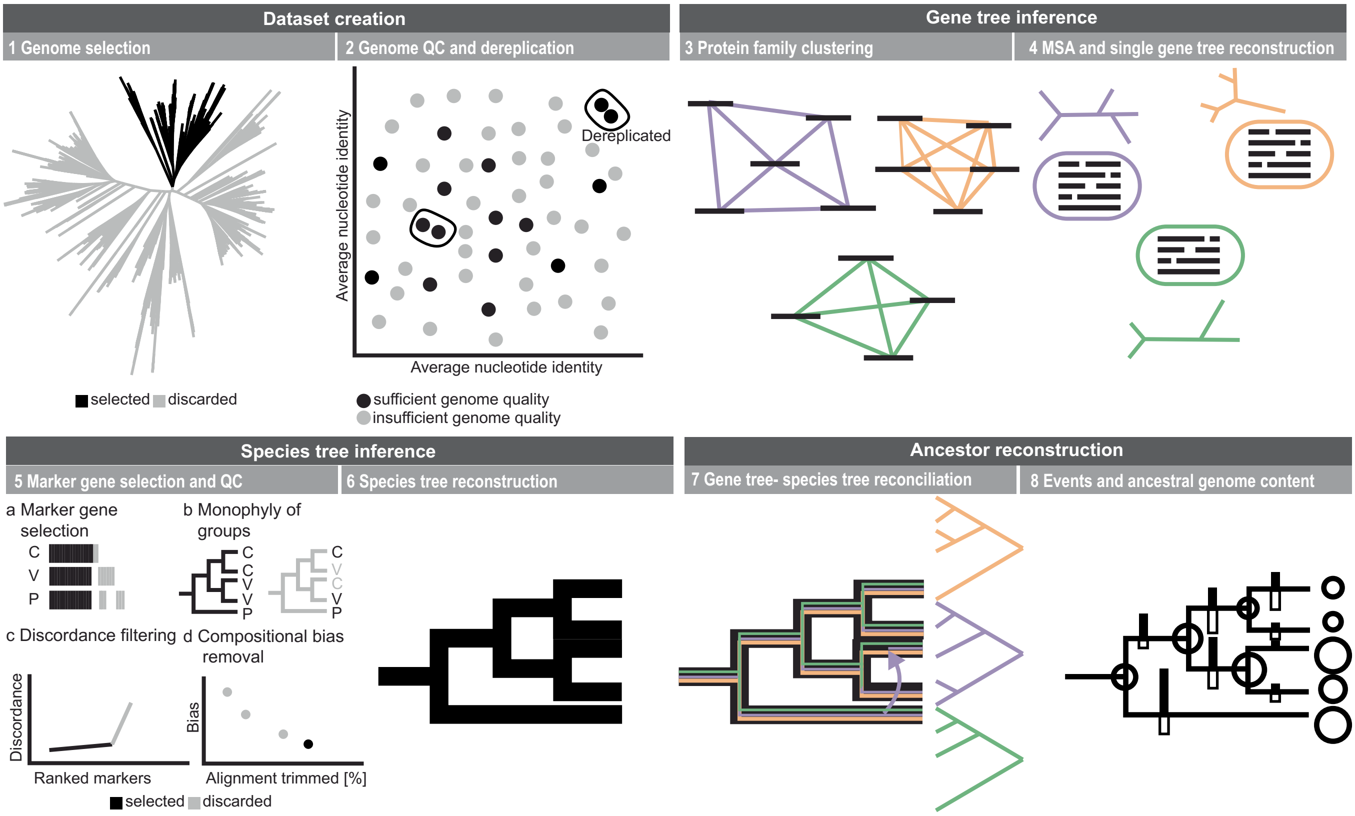
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-022-01284-9.

Correspondence and requests for materials should be addressed to Thijs J. G. Ettema or Matthias Horn.

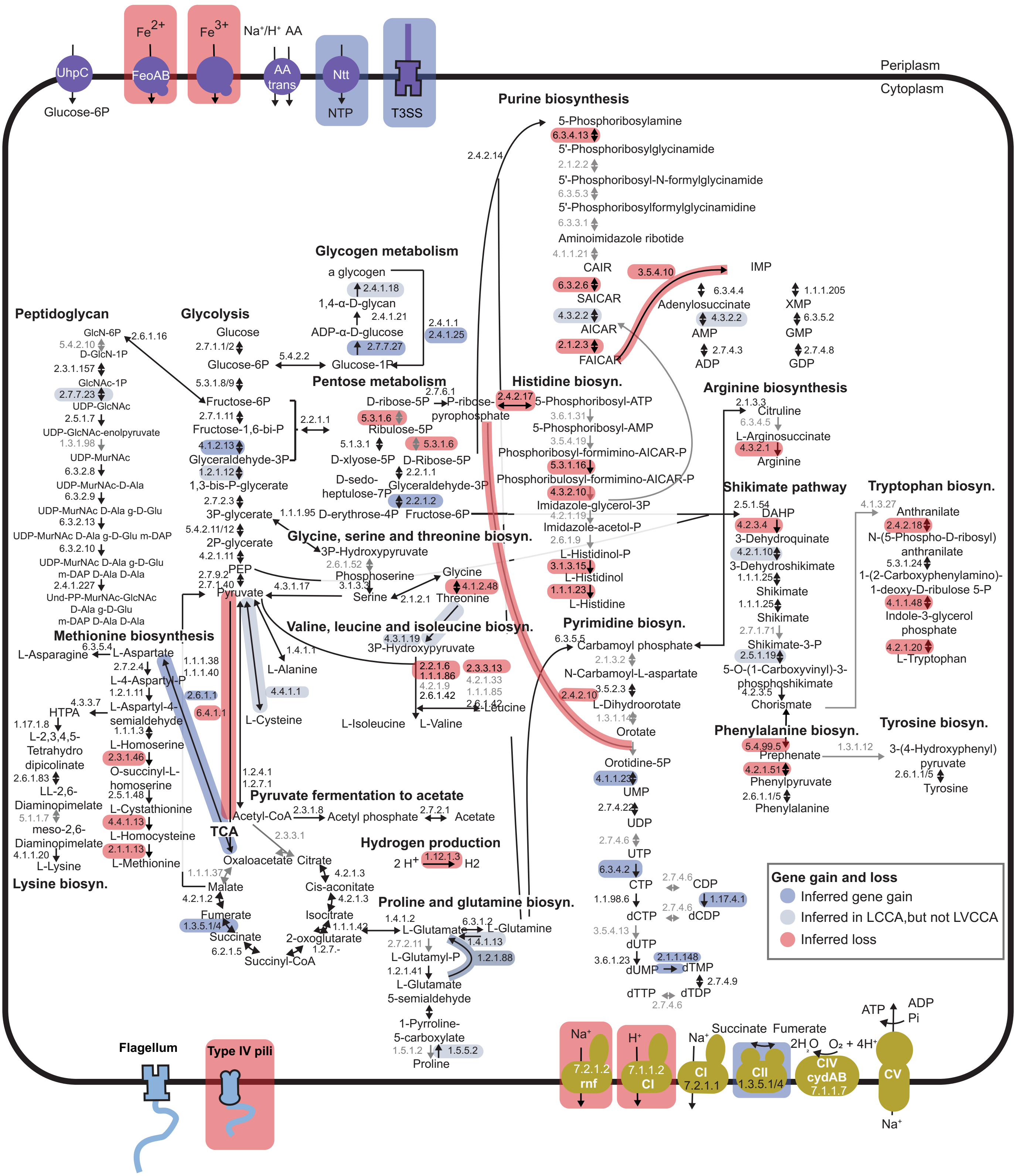
Peer review information *Nature Microbiology* thanks Christian Jogler, Peter Timms and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

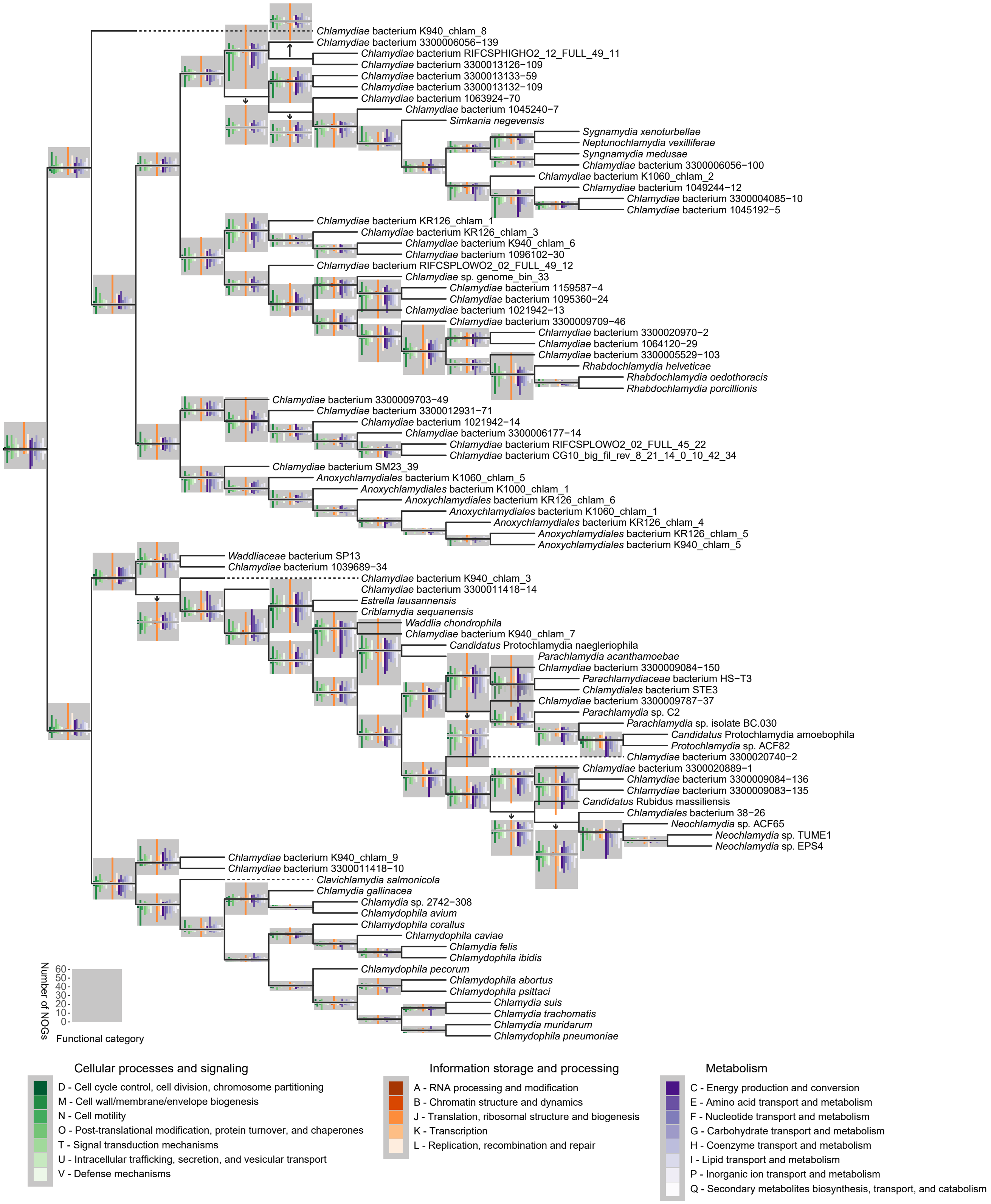
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


Dataset creation
1 Genome selection
2 Genome QC and dereplication
Gene tree inference
3 Protein family clustering
4 MSA and single gene tree reconstruction
Average nucleotide identity
Average nucleotide identity
Dereplicated
■ selected ■ discarded
● sufficient genome quality
● insufficient genome quality
Species tree inference
5 Marker gene selection and QC
6 Species tree reconstruction
Ancestor reconstruction
7 Gene tree- species tree reconciliation
8 Events and ancestral genome content
a Marker gene selection
b Monophyly of groups
C
V
P
C
C
V
V
P
C
V
C
V
P
c Discordance filtering
d Compositional bias removal
Discordance
Ranked markers
Bias
Alignment trimmed [%]
■ selected ■ discarded

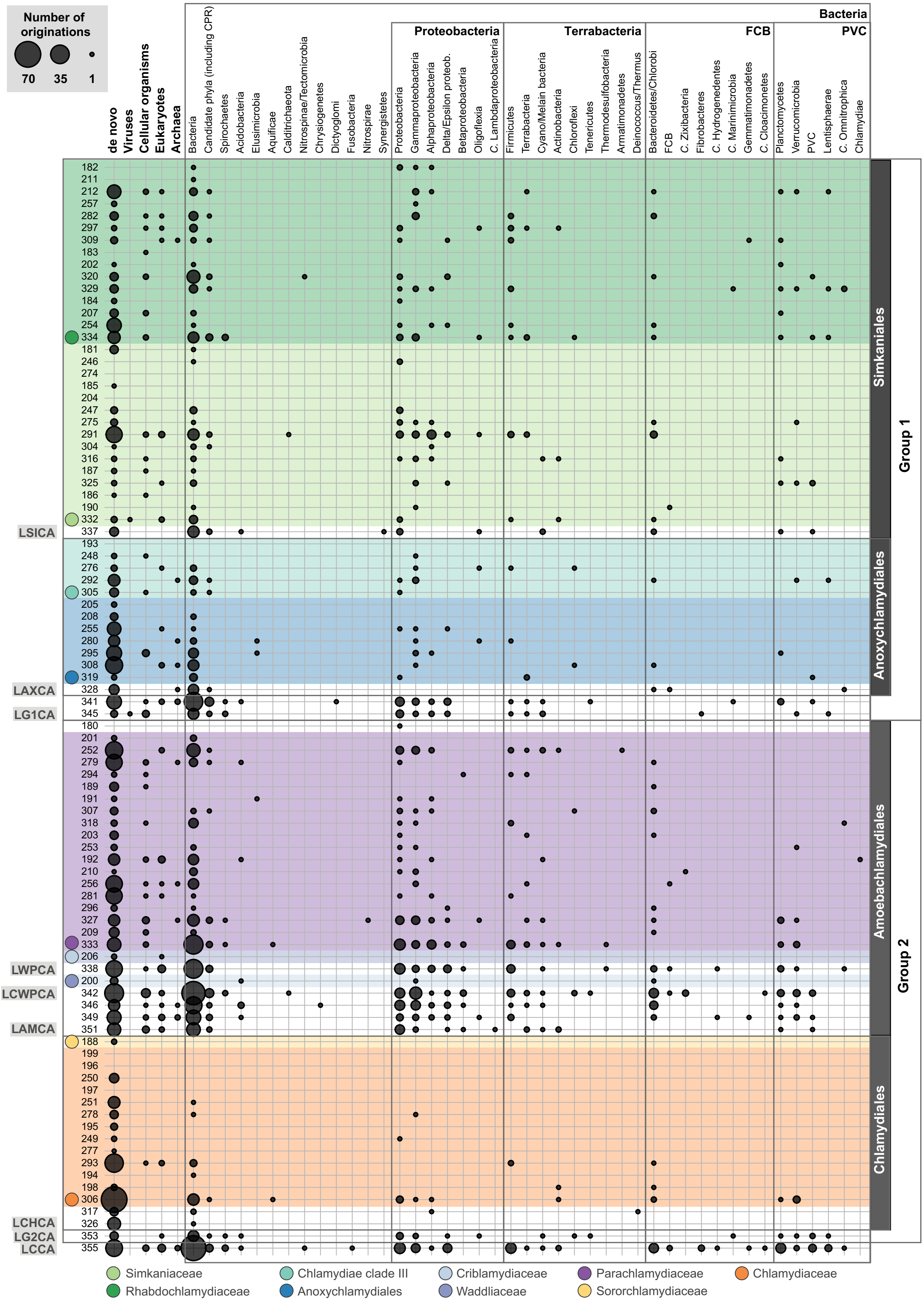
Extended Data Fig. 1 | Workflow for ancestral state reconstruction of Chlamydiae. Dataset creation: PVC bacteria representatives from public repositories were selected with completeness ≥90% and redundancy ≤2%. Species and genus-level representatives were selected for Chlamydiae and non-Chlamydiae PVC members, respectively. Gene tree inference: Protein sequences from the selected dataset were clustered into NOG gene families at the last universal common ancestor level. Unmapped protein sequences were de novo clustered. Protein sequences from each resulting gene family were aligned into a multiple sequence alignment (MSA) and ML single-gene trees inferred. Species tree inference: Gene families found in a single-copy in at least 95% of dataset taxa were selected as potential marker genes. ML single-gene trees were inferred and manually curated, with marker genes that well-resolved PVC phyla retained; further marker genes were removed through discordance filtering, while distant homologs, paralogs, and redundant sequences were removed for each retained marker gene. Individually aligned protein sequences from each marker gene were then concatenated into a supermatrix alignment that was used for both ML and Bayesian phylogenetic inference, with compositionally heterogeneous sites sequentially removed to reduce bias. Ancestor reconstruction: ancestral states were reconstructed using gene-tree species-tree reconciliation. See Methods for details.



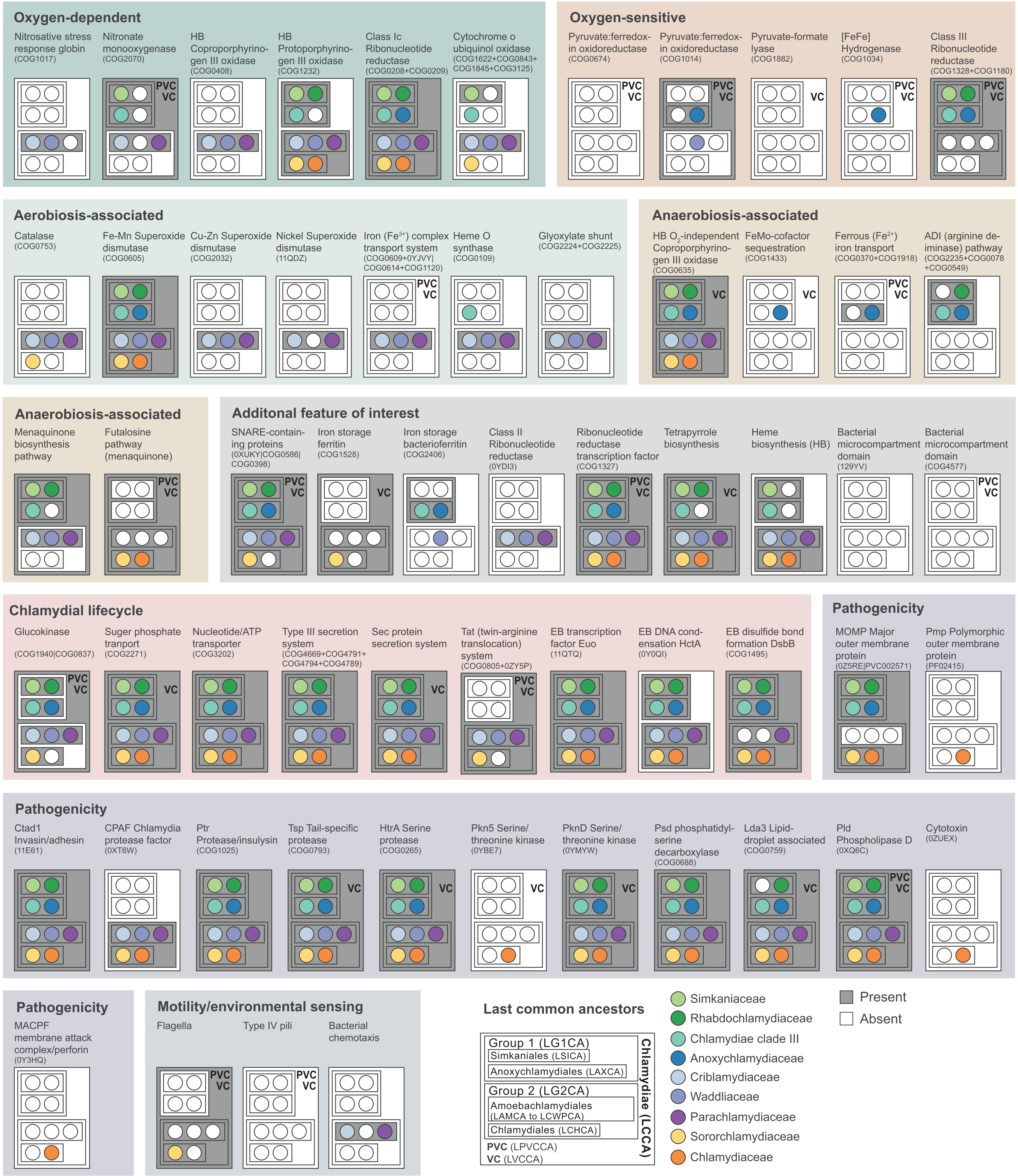
Extended Data Fig. 2 | Chlamydiae last common ancestor reconstructed central metabolism. Genes were reconstructed as present if they had an inferred relative frequency ≥0.3 in ALE. Enzyme complexes were inferred as present if at least half of the necessary genes were present. Enzymes are annotated with enzyme commission numbers, arrows indicate the directionality of reactions and are coloured in black or grey if inferred as present or absent, respectively. Highlights show changes in inferred gene content between LVCCA and LCCA. Purple, light purple and red highlights indicate inferred gains, presence (while not inferred in LVCCA), and losses LVCCA, respectively. See Data S9 for gene families inferred in LCCA and LVCCA and corresponding annotations.


Chlamydiae bacterium K940_chlam_8
Chlamydiae bacterium 3300006056−139
Chlamydiae bacterium RIFCSPHIGHO2_12_FULL_49_11
Chlamydiae bacterium 3300013126−109
Chlamydiae bacterium 3300013133−59
Chlamydiae bacterium 3300013132−109
Chlamydiae bacterium 1063924−70
Chlamydiae bacterium 1045240−7
Simkania negevensis
Sygnamydia xenoturbellae
Neptunochlamydia vexilliferae
Syngnamydia medusae
Chlamydiae bacterium 3300006056−100
Chlamydiae bacterium K1060_chlam_2
Chlamydiae bacterium 1049244−12
Chlamydiae bacterium 3300004085−10
Chlamydiae bacterium 1045192−5
Chlamydiae bacterium KR126_chlam_1
Chlamydiae bacterium KR126_chlam_3
Chlamydiae bacterium K940_chlam_6
Chlamydiae bacterium 1096102−30
Chlamydiae bacterium RIFCSPLOWO2_02_FULL_49_12
Chlamydiae sp. genome_bin_33
Chlamydiae bacterium 1159587−4
Chlamydiae bacterium 1095360−24
Chlamydiae bacterium 1021942−13
Chlamydiae bacterium 3300009709−46
Chlamydiae bacterium 3300020970−2
Chlamydiae bacterium 1064120−29
Chlamydiae bacterium 3300005529−103
Rhabdochlamydia helveticae
Rhabdochlamydia oedothoracis
Rhabdochlamydia porcillionis
Chlamydiae bacterium 3300009703−49
Chlamydiae bacterium 3300012931−71
Chlamydiae bacterium 1021942−14
Chlamydiae bacterium 3300006177−14
Chlamydiae bacterium RIFCSPLOWO2_02_FULL_45_22
Chlamydiae bacterium CG10_big_fil_rev_8_21_14_0_10_42_34
Chlamydiae bacterium SM23_39
Anoxychlamydiales bacterium K1060_chlam_5
Anoxychlamydiales bacterium K1000_chlam_1
Anoxychlamydiales bacterium KR126_chlam_6
Anoxychlamydiales bacterium K1060_chlam_1
Anoxychlamydiales bacterium KR126_chlam_4
Anoxychlamydiales bacterium KR126_chlam_5
Anoxychlamydiales bacterium K940_chlam_5
Waddliaceae bacterium SP13
Chlamydiae bacterium 1039689−34
Chlamydiae bacterium K940_chlam_3
Chlamydiae bacterium 3300011418−14
Estrella lausannensis
Criblamydia sequanensis
Waddlia chondrophila
Chlamydiae bacterium K940_chlam_7
Candidatus Protochlamydia naegleriophila
Parachlamydia acanthamoebae
Chlamydiae bacterium 3300009084−150
Parachlamydiaceae bacterium HS−T3
Chlamydiales bacterium STE3
Chlamydiae bacterium 3300009787−37
Parachlamydia sp. C2
Parachlamydia sp. isolate BC.030
Candidatus Protochlamydia amoebophila
Protochlamydia sp. ACF82
Chlamydiae bacterium 3300020740−2
Chlamydiae bacterium 3300020889−1
Chlamydiae bacterium 3300009084−136
Chlamydiae bacterium 3300009083−135
Candidatus Rubidus massiliensis
Chlamydiales bacterium 38−26
Neochlamydia sp. ACF65
Neochlamydia sp. TUME1
Neochlamydia sp. EPS4
Chlamydiae bacterium K940_chlam_9
Chlamydiae bacterium 3300011418−10
Clavichlamydia salmonicola
Chlamydia gallinacea
Chlamydia sp. 2742−308
Chlamydophila avium
Chlamydophila corallus
Chlamydophila caviae
Chlamydia felis
Chlamydophila ibidis
Chlamydophila pecorum
Chlamydophila abortus
Chlamydophila psittaci
Chlamydia suis
Chlamydia trachomatis
Chlamydia muridarum
Chlamydophila pneumoniae
Number of NOGs
60
50
40
30
20
10
0
Functional category
Cellular processes and signaling
D - Cell cycle control, cell division, chromosome partitioning
M - Cell wall/membrane/envelope biogenesis
N - Cell motility
O - Post-translational modification, protein turnover, and chaperones
T - Signal transduction mechanisms
U - Intracellular trafficking, secretion, and vesicular transport
V - Defense mechanisms
Information storage and processing
A - RNA processing and modification
B - Chromatin structure and dynamics
J - Translation, ribosomal structure and biogenesis
K - Transcription
L - Replication, recombination and repair
Metabolism
C - Energy production and conversion
E - Amino acid transport and metabolism
F - Nucleotide transport and metabolism
G - Carbohydrate transport and metabolism
H - Coenzyme transport and metabolism
I - Lipid transport and metabolism
P - Inorganic ion transport and metabolism
Q - Secondary metabolites biosynthesis, transport, and catabolism

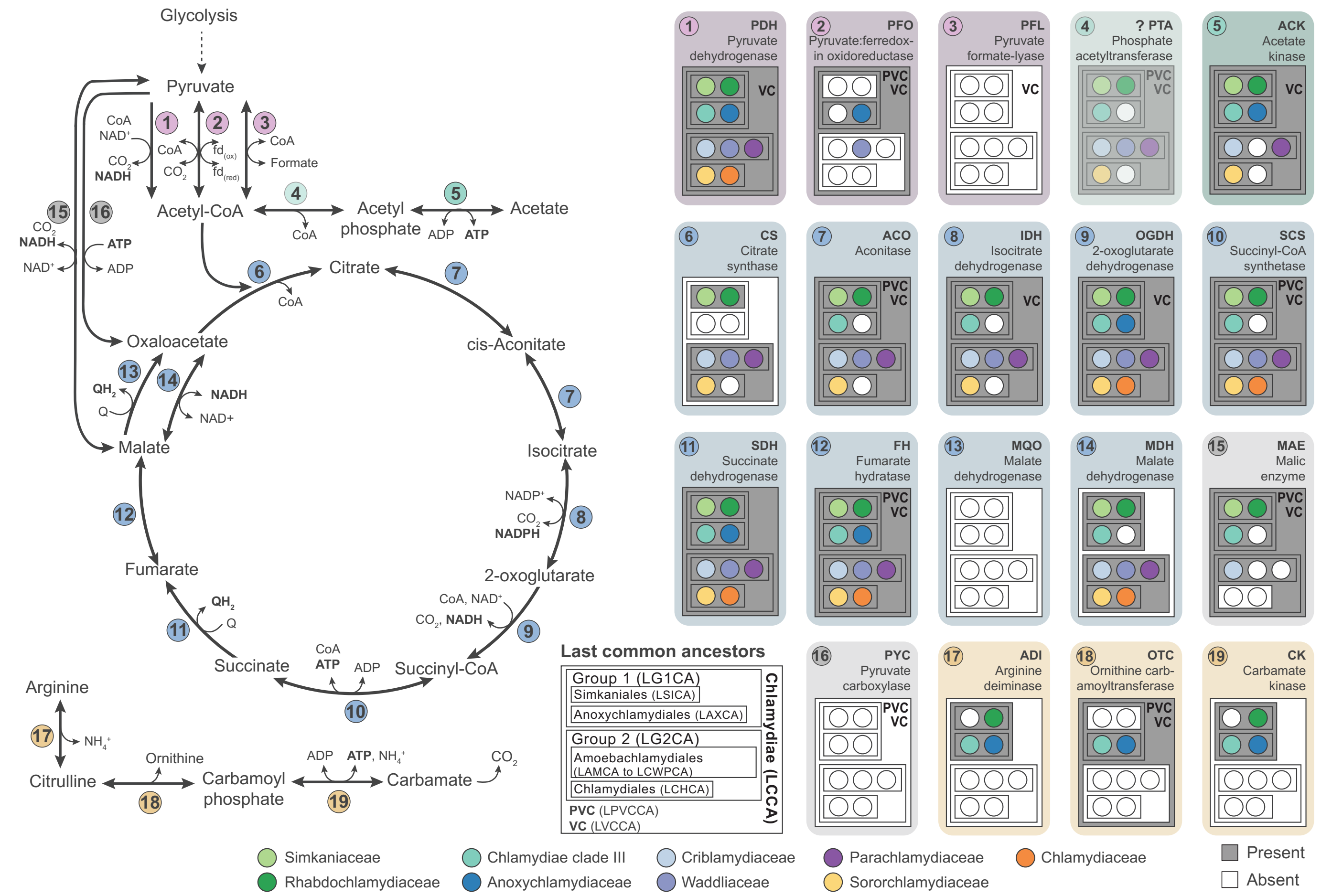
Extended Data Fig. 3 | Gain and loss events per functional category across Chlamydiae. Cladogram of Chlamydiae phylogenetic relationships showing gain (originations and transfers) and loss events of eggNOG NOG gene families with reconciliation frequencies ≥0.3 (see Methods) across COG categories. The barplots mapped onto each branch indicate the events that have occurred leading to the ancestral node to their right. The number of NOGs gained are indicated in the positive direction, and those lost in the negative direction, with the number corresponding to the bar height (see grey box scale). Bars are sorted and coloured according to the COG category. NOGs assigned to poorly characterized COG categories (R: general function prediction only, S: function unknown, X: Mobilome), multiple categories, and de novo gene families were excluded. See Data S7 for gene gains and losses inferred per functional category.



Extended Data Fig. 4 | Taxonomic affiliation of inferred gene originations across Chlamydiae. Number of gene originations affiliated with different taxonomic groups across Chlamydiae ancestral nodes. For all chlamydial originations (that is, transfers from outside the PVC bacteria dataset used), gene trees were inferred with sequences from public databases and the affiliated taxonomic group identified (consensus of 75% of taxa in the sister clade), which represents the putative donor group. Chlamydiae taxonomy and ancestors are indicated to the right and left, respectively; see Data S7 for ancestor abbreviations and Figure S10 for node numbers mapped to chlamydial phylogeny. See Data S11 for chlamydial originations and affiliated taxonomic groups.

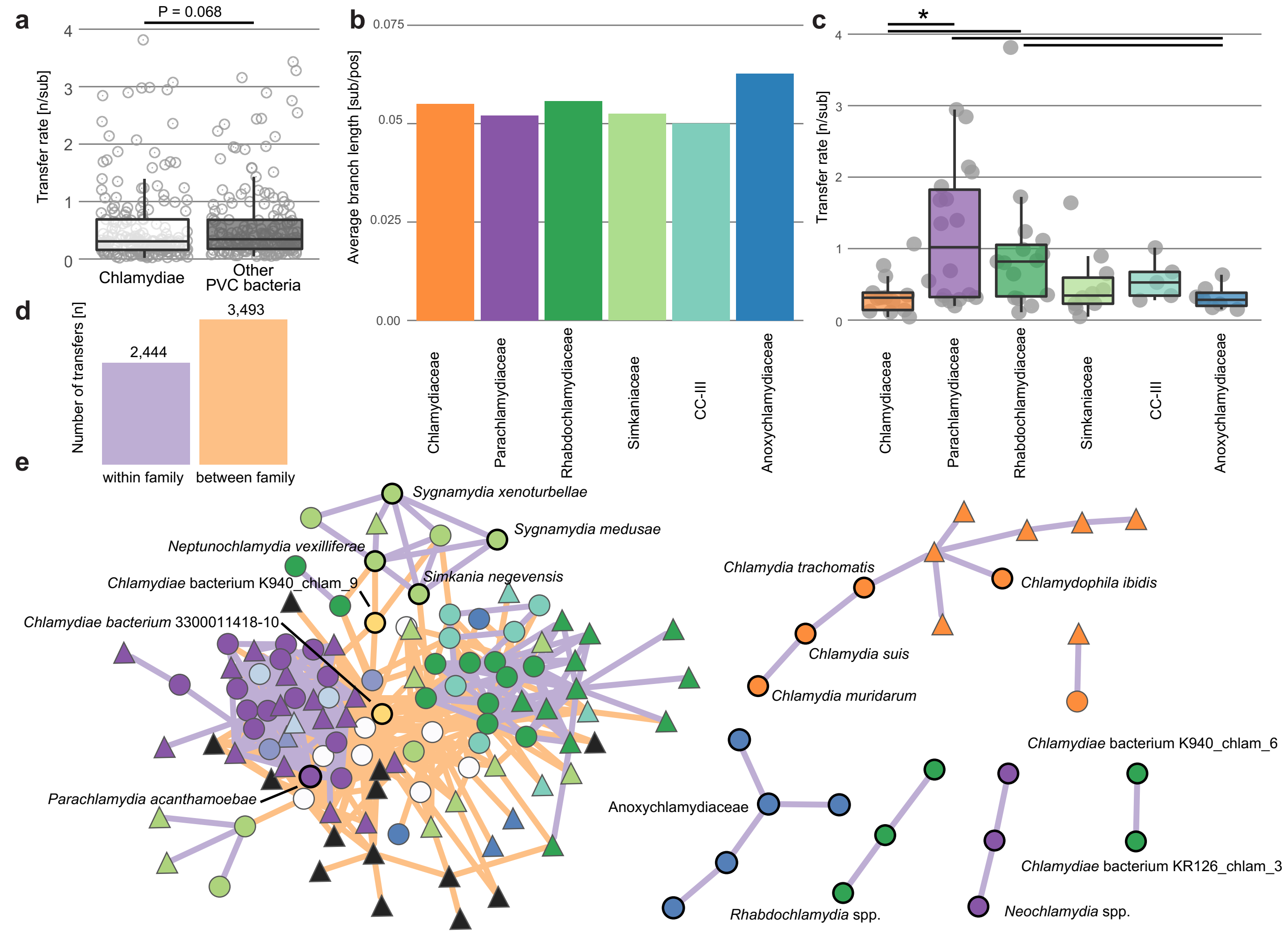

Oxygen-dependent
Nitrosative stress response globin (COG1017)
Nitronate monooxygenase (COG2070)
HB Coproporphyrinogen III oxidase (COG0408)
HB Protoporphyrinogen III oxidase (COG1232)
Class Ic Ribonucleotide reductase (COG0208+COG0209)
Cytochrome o ubiquinol oxidase (COG1622+COG0843+ COG1845+COG3125)
Oxygen-sensitive
Pyruvate:ferredoxin oxidoreductase (COG0674)
Pyruvate:ferredoxin oxidoreductase (COG1014)
Pyruvate-formate lyase (COG1882)
[FeFe] Hydrogenase (COG1034)
Class III Ribonucleotide reductase (COG1328+COG1180)
Aerobiosis-associated
Catalase (COG0753)
Fe-Mn Superoxide dismutase (COG0605)
Cu-Zn Superoxide dismutase (COG2032)
Nickel Superoxide dismutase (11QDZ)
Iron (Fe3+) complex transport system (COG0609+0YJVY| COG0614+COG1120)
Heme O synthase (COG0109)
Glyoxylate shunt (COG2224+COG2225)
Anaerobiosis-associated
HB O2-independent Coproporphyrinogen III oxidase (COG0635)
FeMo-cofactor sequestration (COG1433)
Ferrous (Fe2+) iron transport (COG0370+COG1918)
ADI (arginine deiminase) pathway (COG2235+COG0078 +COG0549)
Anaerobiosis-associated
Menaquinone biosynthesis pathway
Futalosine pathway (menaquinone)
Additonal feature of interest
SNARE-containing proteins (0XUKY|COG0586| COG0398)
Iron storage ferritin (COG1528)
Iron storage bacterioferritin (COG2406)
Class II Ribonucleotide reductase (0YDI3)
Ribonucleotide reductase transcription factor (COG1327)
Tetrapyrrole biosynthesis
Heme biosynthesis (HB)
Bacterial microcompartment domain (129YV)
Bacterial microcompartment domain (COG4577)
Chlamydial lifecycle
Glucokinase (COG1940|COG0837)
Suger phosphate tranport (COG2271)
Nucleotide/ATP transporter (COG3202)
Type III secretion system (COG4669+COG4791+ COG4794+COG4789)
Sec protein secretion system
Tat (twin-arginine translocation) system (COG0805+0ZY5P)
EB transcription factor Euo (11QTQ)
EB DNA condensation HctA (0Y0QI)
EB disulfide bond formation DsbB (COG1495)
Pathogenicity
MOMP Major outer membrane protein (0Z5RE|PVC002571)
Pmp Polymorphic outer membrane protein (PF02415)
Pathogenicity
Ctad1 Invasin/adhesin (11E61)
CPAF Chlamydia protease factor (0XT6W)
Ptr Protease/insulysin (COG1025)
Tsp Tail-specific protease (COG0793)
HtrA Serine protease (COG0265)
Pkn5 Serine/threonine kinase (0YBE7)
PknD Serine/threonine kinase (0YMYW)
Psd phosphatidylserine decarboxylase (COG0688)
Lda3 Lipid-droplet associated (COG0759)
Pld Phospholipase D (0XQ6C)
Cytotoxin (0ZUEX)
Pathogenicity
MACPF membrane attack complex/perforin (0Y3HQ)
Motility/environmental sensing
Flagella
Type IV pili
Bacterial chemotaxis
PVC
VC
Last common ancestors
Group 1 (LG1CA)
Simkaniales (LSICA)
Anoxychlamydiales (LAXCA)
Group 2 (LG2CA)
Amoebachlamydiales (LAMCA to LCWPCA)
Chlamydiales (LCHCA)
Chlamydiae (LCCA)
PVC (LPVCCA)
VC (LVCCA)
Simkaniaceae
Rhabdochlamydiaceae
Chlamydiae clade III
Anoxychlamydiaceae
Criblamydiaceae
Waddliaceae
Parachlamydiaceae
Sororchlamydiaceae
Chlamydiaceae
Present
Absent

Extended Data Fig. 5 | Reconstructed presence of genes and pathways of interest across Chlamydiae ancestors. Gene family IDs (COG, NOG, or de novo identifier) are indicated for genes and complexes, and for pathways can be found in Data S10. See Data S10 for corresponding gene family annotations and inferred presence across all Chlamydiae ancestors.



Extended Data Fig. 6 | Reconstructed presence of TCA cycle components and fermentative metabolism across Chlamydiae ancestors. Presence and absence of genes encoding enzymes involved in converting pyruvate to acetyl-CoA (pink), pyruvate fermentation to acetate (green), the TCA cycle (blue), pyruvate to TCA intermediate interconversion (grey), and the fermentative arginine deiminase pathway (yellow). A schematic overview of reactions performed by each numbered enzyme is shown to the left, alongside key metabolites (water, bicarbonate and protons are excluded). To the right the presence and absence of each component across key ancestors is indicated according to the legend. See Data S10 for gene annotations and inferred presence across all Chlamydiae ancestors.

https://doi.org/10.1038/s41564-022-01284-9



Extended Data Fig. 7 | Transfer rates in Chlamydiae and other PVC bacteria, and network of enriched transfer routes in Chlamydiae. a. Boxplot depicting median lower transfer rates (one-sided Wilcoxon signed rank test, $P = 6.8 \times 10^{-2}$) in Chlamydiae (n = 181, median 0.31, IQR 0.16–0.69) than in other PVC bacteria (n = 172, median 0.34, IQR 0.18–0.68). **b**. The average branch length per family without terminal leaves. **c**. Transfers per substitution in the species tree per chlamydial node for each family. To account for heterogeneous genome sampling in our dataset, we only evaluated transfer rates of chlamydial families with at least two ancestors reconstructed in our analysis, excluding terminal nodes. From left to right, *n* = (14, 18, 15, 10, 5, 7). Transfer rates are significantly higher in ancestors of the Parachlamydiaceae and Rhabdochlamydiaceae than in the Chlamydiaceae ($P = 4.8 \times 10^{-3}$ and 8.5×10^{-3}) and in Anoxychlamydiaceae ($P = 7.1 \times 10^{-3}$ and 1.1×10^{-3}), respectively. Asterisk and lines indicate families with $P \leq 0.05$ based on a one-sided Wilcoxon signed rank test with Bonferroni correction. **d**. Barplot showing the number of significant transfer events (binomial test, significance $P \leq 0.05$; see Methods) in gene highways within (purple) and between (orange) chlamydial families and **e**. corresponding networks depicting chlamydiae (nodes, colour-coded by family; families represented by only one genome are white, see Fig. 1) sharing significant gene highways (edges). Edges are coloured based on gene highways within and between families in purple and orange, respectively. Center lines in the box-and-whisker plots represent median values, box limits represent upper and lower quartile values, and whiskers represent 1.5 times the interquartile range above the upper quartile and below the lower quartile. See also Figure S10 for events across all PVC nodes, and the repository for raw data.

nature portfolio

Corresponding author(s): Matthias Horn & Thijs Ettema

Last updated by author(s): October 31, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection
dRep v1.4.3; GTDB v86; miComplete 1.1.1;

Data analysis
emapper v1.0.1; MAFFT L-INS-i v7.427; BMGE vl.12; IQ-TREE v1.6.10, 1.6.11, 1.6.12, 2.0; PhyloBayes-MPI v1.7b; USEARCH v11.0.667; trimAl v1.4.1; PhyloBayes v4.1c; DIAMOND v0.9.21; SiliX v1.2.9; interproscan v5.36-75.0; Count v10.04; ALE v0.449; Cytoscape v3.7.0; Snakemake v5.19.3; MAFFT v7.471; FastTree 2 v2.1.11; ETE3 Toolkit v3.1.2; PhyloRank v0.1.10; R v3.6.2; ggplot2 v3.3.3; ggtree v2.5.0.991; treeio v1.10.0; Figtree v1.4.4; CD-HIT v4.8.1; trimAl v1.4.rev15

Custom code:
https://github.com/jennahd/HGT_trees,
https://github.com/maxemil/ALE-pipeline

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Genome data was obtained either from NCBI Genbank (https://www.ncbi.nlm.nih.gov/nucleotide/), the JGI portal (https://portal.nersc.gov/GEM/) or a zenodo repository (https://doi.org/l0.5281/zenodo.4318714). Small subunit rRNA gene data used in this study are available via the SILVA database (https://www.arb-silva.de/). Genbank accessions and database links for genomes used in the ancestral state reconstruction are provided in Data S2. Additional raw data files are hosted on the online repository figshare (https://doi.org/10.6084/m9.figshare.17033417). These include sequences, alignments, trimmed alignments, and trees for single-copy marker genes used for species phylogenies (both those selected and not selected), the 16S rRNA gene, and concatenated alignments and trees for all three species datasets (of 184, 183, and 180 taxa). Both NOG and de novo gene families used for the ancestral state reconstruction are also provided alongside alignments, trimmed alignments, trees, and bootstrap trees (ufboot) provided to ALE. The raw ALE results with all events are also included, alongside gene annotations together with events, and events for each gene family mapped to the species tree. Protein sequence datasets, alignments and trees inferred as part of the analysis to determine HGT donors for chlamydiae gene originations are provided. In addition, pdfs of metabolic reconstructions of LVCCA, LG1CA, and LG2CA can be found in the repository files.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Phylogenomic analyses of Chlamydiae and other Planctomycetes-Verrucomicrobia-Chlamydiae (PVC) superphylum bacteria. Gene-tree species tree reconciliation and ancestral reconstruction of the last common ancestors of PVC bacteria with a focus on Chlamydiae evolution. Analysis of the origin of gained gene content in the Chlamydiae phylum for key ancestors.
Research sample	Available genomes of cultured and uncultured PVC bacteria with a focus on recently published Chlamydiae draft genomes.
Sampling strategy	We selected genomes based on their phylogenetic affiliation with PVC bacteria. Representatives where then selected based on the highest genome quality score per taxonomic unit: approximately species for Chlamydiae, and genus (sensu GTDB) level for other PVC bacteria.
Data collection	N/A, as primary data collection (i.e., DNA sequencing, genome assembly, and quality control) was performed by other parties (sequence contributors to JGI and NCBI). PVC genomes were downloaded based on their taxonomy from GTDB, additional chlamydiae from more recent studies were additionally downloaded from the JGI.
Timing and spatial scale	Genomes were downloaded all at once on April 3rd, 2019.
Data exclusions	We excluded Chlamydiae genomes with a miComplete specific marker gene set estimated completeness smaller than 0.9 and a redundancy larger than 1.02. Quality of other PVC genomes was based on GTDB provided CheckM quality scores, we excluded genomes with an estimated completeness smaller than 90% and a contamination larger than 2%.

Reproducibility	All results of this study can be reproduced given the same original source data and the methods provided in this manuscript.
Randomization	N/A because randomization was not required for the purposes of this study, as we based all ancestral reconstructions on the complete set of highest quality available PVC genomes to infer ancestral state reconstruction.
Blinding	N/A because blinding was not required for the purposes of this study, as the taxonomic and evolutionary context was of great importance for interpretation of the findings.

Did the study involve field work? □ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging