The evolutionary origin of host association in the Rickettsiales

Max E. Schönn, Joran Martijnn, Julian Vossebergn, Stephan Köstlbachern and Thijs J. G. Ettema

Obligate host-associated bacteria include pathogens that represent a leading cause of human, livestock and crop disease, resulting in considerable economic loss worldwide. While the molecular and cellular underpinnings of host association have been described in considerable detail, their evolutionary origin remains generally poorly understood. The Rickettsiales represent a widespread and diverse order of obligate host-associated alphaproteobacteria that have been estimated to have originated >1.7 Ga, similar to a conservative estimated age of eukaryotes. Rickettsiales infect a wide variety of eukaryotic species, including protists, leeches, cnidarians, arthropods and mammals. Well-known examples include Rickettsia prowazekii, the causative agent of epidemic typhus in humans, and Wolbachia, a genus of bacteria infecting over two-thirds of arthropods and nearly all ferial nematodes. Genomes of Rickettsiales are shaped by ongoing reductive evolution, leading to small, AT-rich genomes with limited metabolic capacities. Here we uncover eleven deep-branching alphaproteobacterial metagenome assembled genomes from aquatic environments, including data from the Tara Oceans initiative and other publicly available datasets, distributed over three previously undescribed Rickettsiales-related clades. Phylogenomic analyses reveal that two of these clades, Mitibacteraceae and Athabascaceae, branch sister to all previously sampled Rickettsiales. The third clade, Gamibacteraceae, branch sister to the recently identified ectosymbiotic Candidatus Deianirae vastatrix. Comparative analyses indicate that the gene complement of Mitibacteraceae and Athabascaceae is reminiscent of that of free-living and biofilm-associated bacteria. Ancestral genome content reconstruction across the Rickettsiales species tree further suggests that the evolution of host association in Rickettsiales was a gradual process that may have involved the repurposing of a type IV secretion system.

Results

Metagenomic identification of previously undescribed Rickettsiales. To shed light on the early evolution of Rickettsiales and the emergence of host association within this clade, we screened publicly available metagenomic repositories for deep-branching Rickettsiales (Supplementary Data 1 and 2). We reconstructed three metagenome assembled genomes (MAGs) from the Tara Oceans metagenome data and identified another eight MAGs from public data derived from aquatic (marine, lake, aquifer and tailings water) environments (Supplementary Fig. 1 and Data 3). To assess their phylogenetic affiliation with other alphaproteobacteria and mitochondrial genomes, we inserted them into a previously established dataset of publicly available metagenomic repositories for deep-branching Rickettsiales. Metagenomic placement of obtained clades. To pinpoint the phylogenetic position of the retrieved Rickettsiales-associated clades

1Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. 2Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden. 3Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Canada. 4Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Utrecht, The Netherlands. 5Laboratory of Microbiology, Wageningen University and Research, Wageningen, The Netherlands. 6These authors contributed equally: Max E. Schönn, Joran Martijn. 7E-mail: thijs.ettema@wur.nl
more confidently, we performed in-depth phylogenomic analyses with a dataset comprising 116 marker genes highly conserved among Rickettsiales\(^{16}\) (Supplementary Fig. 2). These analyses confirmed that two of the three clades (hereafter Mitibacteraceae and Athabascaceae; Supplementary Text) branched as distinct groups sister to all other sampled Rickettsiales (Supplementary Figs. 3 and 4, and Data 1). Of the two, the Mitibacteraceae represented the deepest branching clade. The third clade (hereafter Gamibacteraceae; Supplementary Text) branched sister to the recently described ‘Candidatus Deianiraea vastatrix’ (Fig. 2a), a host-associated bacterium able to replicate outside their host cells\(^{17}\). The placement of the clade comprising Gamibacteraceae and Deianiraeaceae within the Rickettsiales species tree was not fully resolved. In agreement with Castelli et al.\(^{17}\), this clade branched sister to the Anaplasmataceae with maximum branch support in both Bayesian (Supplementary Fig. 3) and maximum likelihood (Supplementary Fig. 4) trees. However, alphaproteobacterial and possibly rickettsial phylogenies\(^{25,27–29}\) are known to be affected by long branch and compositional bias artefacts. We tested for such artefacts by separately removing the long-branched ‘Ca. D. vastatrix’ and the most heterogeneous sites (Methods and Supplementary Text). The overall topology was robust to either treatment (Supplementary Figs. 5 and 6), except for the placement of the Gamibacteraceae-Deianiraeaceae clade. When removing the most heterogeneous sites, they branched sister to a clade comprising the Midichloriaceae and Anaplasmataceae with near maximum branch support in the Bayesian tree (Fig. 2a, Supplementary Fig. 7 and Table 1) but with insignificant branch support in the maximum likelihood tree (Supplementary Text and Fig. 8). On the basis of the resolved Bayesian phylogeny, we suggest that this topology more probably reflects Rickettsiales evolutionary history and that the placement of the Gamibacteraceae-Deianiraeaceae clade sister to Anaplasmataceae is the result of a phylogenetic artefact.

**Environmental distribution of added Rickettsiales.** To assess the environmental distribution of the newly obtained alphaproteobacterial clades, we used the 16S ribosomal RNA gene sequences of Gamibacteraceae and Mitibacteraceae MAGs to query public sequence databases. Athabascaceae MAGs did not contain 16S rRNA genes and could therefore not be included in this analysis. Highly similar sequences were exclusively found in datasets obtained from aquatic habitats. Sequences closely related to the Gamibacteraceae 16S rRNA gene were associated with a diverse set of environments, including lakes and aquifers as well as marine systems. Interestingly, two recovered 16S rRNA gene sequences were obtained from sequencing datasets of cells of the freshwater cnidarian *Hydra vulgaris* and of the marine ciliate *Hemigastrostyla elongata* (Fig. 2b and Supplementary Fig. 9)\(^{30,31}\). The two sequences are unlikely to be contaminants as rigorous efforts were made to minimize contamination\(^{30,31}\). Their detection in a ciliate microbiome could hint at a conserved lifestyle of Gamibacteraceae and their closest characterized relative ‘Ca. D. vastatrix’, which was described to colonize the cell surface of its ciliate host *Paramecium primaurelia*\(^{17}\). Apart from Gamibacteraceae, Deianiraeaceae-related 16S rRNA gene sequences were also detected in the *Hydra vulgaris* microbiome (Supplementary Fig. 9)\(^{30,31}\). Most probably, the observed interactions of Gamibacteraceae and Deianiraeaceae with *Hydra* can be explained by the presence of host ciliates in the polyp microbiome. Alternatively, *Hydra* could represent a direct host for these lineages as well. We could not detect potential host organisms for Mitibacteraceae, as all ‘Ca. Mitibacter marchionensis’-related 16S rRNA gene sequences were exclusively associated with marine, non-host-associated environments (Fig. 2c and Supplementary Fig. 9).

**Inferred physiology of Mitibacteraceae and Athabascaceae.** The reconstructed genomes of Mitibacteraceae and Athabascaceae are on average larger in size, display a higher GC content and a
higher protein-coding density (Fig. 2d–f and Supplementary Data 3) compared with those of the known obligate host-associated Rickettsiales (hereafter referred to as ‘classical Rickettsiales’, including Gambobacteraeae). Mitibacteraceae and Athabascaceae, like most classical Rickettsiales, encode an incomplete glycolysis pathway and a complete tricarboxylic acid cycle (Extended Data Fig. 2 and Supplementary Data 4). However, unlike all classical Rickettsiales, they encode complete pathways to synthesize all amino acids and nucleotides (Fig. 3a, Extended Data Fig. 2 and Supplementary Data 4). Even the ectosymbiotic Ca. D. vastatrix, by far the richest classical Rickettsiales described so far in terms of amino acid biosynthesis potential, can only synthesize 15 amino acids and lacks the genetic potential for de novo nucleotide biosynthesis (Fig. 3a, Extended Data Fig. 2 and Supplementary Data 4)12. The two lineages further encode nearly all subunits of the main oxidative phosphorylation complexes and a complete glycolylate cycle. The latter may allow them to use substrates such as acetate as a sole carbon source (Extended Data Fig. 2 and Supplementary Data 4). They further uncharacteristically encode functions more commonly found in free-living bacteria, such as a sulfate uptake system and a complete (Athabascaceae) or partial (Mitibacteraceae) assimilatory sulfate reduction pathway, with the potential to provide sulfide for the de novo biosynthesis of cysteine (Extended Data Fig. 2 and Supplementary Data 4)13,14. Two Mitibacteraceae and one Athabascaceae MAG also feature arcs gene clusters (Supplementary Data 4), which are lacking in all classical Rickettsiales. These clusters facilitate arsenite export (Extended Data Fig. 2) and so confer arsenic resistance. Arsenic is the most prevalent toxic element in the environment14 and resistance mechanisms are almost ubiquitously found in prokaryotes15. Mitibacteraceae further encode an ammonium transporter which is an important inorganic nutrient uptake system in marine bacteria and is not found in other Rickettsiales16. Despite being aquatic alphaproteobacteria, the Mitibacteraceae and Athabascaceae do not encode proteorhodopsin-related proteins (Supplementary Data 4), indicating that they are unable to harness energy from a light-driven H+ gradient. However, they do encode homologues of hydroperoxidase KatG (not found in other Rickettsiales families) and superoxide dismutase SOD2 (Extended Data Fig. 2, Supplementary Data 4 and 6), suggesting that they experience stress from UV light-induced reactive oxygen species.

Furthermore, we found that Mitibacteraceae and Athabascaceae genomes encode a complete flagellum and associated chemotaxis machinery (Fig. 3a), suggesting a possible motile lifestyle17,18. Their genomes also encode Type 4 pili (T4P; Fig. 3a), which have been implicated in motility, surface attachment and biofilm formation19, but can also function in evasion of protist predators18. Mitibacteraceae genomes additionally encoded a tight adherence pilus (Tad) and an extracellular polysaccharide biosynthesis pathway via pel gene clusters19,20. The latter was also found in the Athabascaceae MAG UBA6187 (Fig. 3a and Extended Data Fig. 3a,b). Both Tad pilus and exopolysaccharide biosynthesis capacity have been shown to play important roles in biofilm formation19,21 and together with the presence of T4P, suggest that Mitibacteraceae, and perhaps Athabascaceae, can form biofilms.

We failed to detect genes encoding an ATP/ADP translocase or any other gene typical of intracellular or parasitic lifestyles in Mitibacteraceae and Athabascaceae, except for those encoding the rvo T4SS, including the duplication of several components31 as well as several distinct modifications compared with the vir T4SS of Agrobacterium32 (Fig. 3b,c and Supplementary Fig. 10). In classical Rickettsiales, this T4SS is typically used for the manipulation of the host cell via translocation of effector proteins10–12.

Besides a few homologues of rickettsial ankyrin repeat protein 1 (RARP-1), a Sec-ToI-C-secreted effector of Rickettsia typhi24, and RARP-2 which may be secreted by the T4SS, we were neither able to identify any of the experimentally verified T4SS effector

---

**Fig. 2** | Phylogeny and comparative analysis of deep-branching Rickettsiales genomes. a, Rickettsiales species tree based on a dataset of 116 marker genes of which the most heterogeneous sites were removed, highlighting the phylogenetic position of Gambobacteraeae (orange shading), Athabascaceae (blue shading) and Mitibacteraceae (green shading). The tree was inferred using PhyloBayes (CAT+LG+F+4, 20,000 generations with a burn-in of 5,000 generations). Support values correspond to posterior probabilities and are only shown for values below 1. See Supplementary Fig. 7 for the complete tree including the outgroup. b–f, Distribution of environmental Gambobacteraeae (b) and Mitibacteraceae sequences (c) obtained from the NCBI nt database. Colours represent the sequences’ source environments. For each genome, the observed genome size in Mbp (d), with estimated genome size in dark grey shading, the GC content (e) and the coding density, calculated as the total length of protein-coding genes divided by the genome size (f), are shown. Source data for graphs in d–f can be found in Supplementary Data 3.
proteins of classical Rickettsiales\textsuperscript{11,12}, nor could we detect similar numbers of putative effector proteins containing eukaryotic-like repeat domains as in other Rickettsiales genomes (ankyrin-, leucine rich- and tetratricopeptide repeats; Supplementary Text, Extended Data Fig. 4 and Supplementary Data 4). In contrast, we detected several proteins containing peptidoglycan-binding (PGB) domains in Mitibacteraceae genomes. These proteins, which are otherwise relatively rare in Rickettsiales (Supplementary Text and Extended Data Fig. 4), have been shown to serve as effectors of a ‘bacteria killing’-type T4SS in other proteobacteria\textsuperscript{45,46}. However, PGB domain proteins are also known to operate in peptidoglycan synthesis and cell shape remodelling\textsuperscript{47}, and may thus allow the Mitibacteraceae to evade predatory protists by providing cell shape plasticity\textsuperscript{48}.

Based on their general genome structure and gene content, Mitibacteraceae and Athabascaceae are atypical compared with the classical obligate intracellular Rickettsiales. Instead, they possibly

\textbf{Fig. 3} Distribution and ancestral inference of phenotypic traits and pathways across Rickettsiales. \textbf{a}, Presence (full circle) and absence (empty circle) of genes or pathways in the seven Rickettsiales families (including Gamibacteraceae, Athabascaceae and Mitibacteraceae), the alphaproteobacterial outgroup as well as the last common ancestors of the classical, obligate host-associated Rickettsiales (LhRCA) and of all Rickettsiales (LRCA). Pathways are grouped into ‘secretion’ (yellow), ‘motility’ (blue), ‘biosynthesis and transport’ (orange) and ‘electron transport chain’ (green). In some cases, pathways are either partial (one or several genes missing) or are absent from some species, but present in others (half circle). Numbers inside the circles represent the number of complete pathways for the biosynthesis of amino acids. The Gln transporter could potentially also transport other polar amino acids such as Ser, Thr and Asn. \textbf{b}, Synteny of the Rickettsiales \textit{vir} homologue (\textit{rvh}) T4SS in classical (black) and environmental (red) Rickettsiales. Taxa are ordered according to the species tree in \textbf{a}. One representative per family except Athabascaceae and Mitibacteraceae with two representatives each. \textbf{c}, Schematic phylogenetic tree showing the common duplication of the \textit{virB}4 gene in LRCA. See Supplementary Fig. 10 for the uncollapsed tree. P-T4SS, type IV secretion system of the P type; T1SS, type I secretion system; CONJ, conjugation system; T4P, type 4 pilus; Tad, tight adherence system.
exhibit a lifestyle similar to aquatic copiotrophic, biofilm-associated and free-living bacteria \(^{49-51}\), or perhaps a facultative host-associated (probably extracellular) lifestyle (Figs. 2d–f and 3a, Extended Data Fig. 2 and Supplementary Data 4). The suggestion that Mitibacteraceae and Athabascaceae are possibly free-living bacteria is reinforced by an analysis with PhenDB \(^{52}\), a machine learning algorithm that has been trained to predict phenotypic traits, including obligate host-association, on the basis of gene content. PhenDB predicted that Mitibacteraceae and Athabascaceae are not obligate intracellular symbionts, in contrast to classical Rickettsiales (Supplementary Table 2).

**Emergence of host association in Rickettsiales.** The deep-branching nature of the Mitibacteraceae and Athabascaceae allowed us to study the origin of the host-associated lifestyle of the Rickettsiales in unprecedented detail. We reconciled \(^{3,53,54}\) 4,240 single-gene trees with the previously obtained species tree (Fig. 2a) to reconstruct the gene family complement as well as gene duplications, transfers, losses and origination events along the entirety of the species tree (Extended Data Fig. 5, Supplementary Fig. 11, and Data 4 and 5). Below, we focus on two ancestors that are central to understanding the emergence of host association: the last Rickettsiales common ancestor (LRCA) and the last obligate host-associated Rickettsiales common ancestor (LhRCA; Fig. 3a).

The LRCA was inferred to feature at least 1,432 protein-coding genes, considerably more than most extant classical Rickettsiales (Supplementary Data 6). This is most probably an underestimate as LRCA gene families could have gone extinct or were not sampled in the analysed genomes. Only a few classical Rickettsiales encode a similar number or more protein-coding genes. For example, *Orientia tsutsugamushi* str. Boryong encodes 2,120 protein-coding genes, but this large number can be explained by a recent massive proliferation of transposons, conjugative T4SS genes and other host-cell interaction genes \(^{25}\). Furthermore, this expanded gene repertoire of *O. tsutsugamushi* also includes around 800 pseudogenes, most of which are duplicates of functional genes \(^{35,56}\). The *rchl* T4SS in LRCA was most probably acquired via horizontal gene transfer from an ancestral proteobacterial donor (Fig. 3c and Supplementary Fig. 10). We were unable to assign any verified Rickettsiales effector proteins \(^{31,112}\) related to eukaryotic host-association to this T4SS, suggesting that it could have served an alternative function. As we inferred LRCA to contain at least four PGB-domain-containing proteins, which are putative effectors of such secretion systems, it could have acted instead as a ‘bacteria killing’ type T4SS \(^{45,46}\). Alternatively, the ancestral T4SS could also have acted in defence against predatory protists. The LRCA was further inferred to encode a very similar gene profile as observed for the Mitibacteraceae and Athabascaceae, hence including many genes affiliated to a free-living lifestyle (for example, full complement of nucleotide and amino acid biosynthesis genes; Fig. 3a, Extended Data Fig. 2 and Supplementary Data 4) and lacking several genes related to the typical symbiotic lifestyle of classical Rickettsiales (for example, an ATP/ADP translocase and several conserved amino acid transporters; Fig. 3a). Besides the capacity to synthesize its own nucleotides and amino acids, LRCA had the genetic potential to take up sulfate and ammonium, carry out assimilatory sulfate reduction and confer arsenic resistance (Fig. 4, Extended Data Fig. 2 and Supplementary Data 4). It had a motile lifestyle, enabled by a flagellum, a T4P and a functioning chemotaxis system. Finally, it most probably was able to initiate biofilm formation using the *pel* and *tad* systems (Fig. 4 and Supplementary Data 4). On the basis of these observations, we propose that LRCA was not an obligate intracellular symbiont, but rather a free-living or perhaps facultative host-associated organism. This would avoid the need for an unprecedented and highly complex intracellular-to-extracellular transition to explain the extracellular lifestyle of *Ca. D. vastatrix*, as originally argued by Castelli et al.\(^{13}\). However, a stage of facultative intracellular host-association cannot be fully excluded either.

The transition from the LRCA to the last common ancestor of all classical Rickettsiales and the Athabascaceae (LhRAtCA) saw little net change to the number of inferred protein-coding genes (from \(n=1,432\) in the LRCA to \(n=1,430\) in LhRAtCA). Besides the loss of several genes encoding components of the *Tad* pilus in LhRAtCA, no genes of note were lost or gained (Fig. 3a). In contrast, the LhRCA had considerably fewer genes (\(n=1,165\)) and had lost genes involved in alanine, cysteine and serine biosynthesis, sulfate assimilation, sulfate transporters and ammonium transporters (Fig. 4). It furthermore lost the *pel* genes, suggesting that the ability to form biosfilms was lost at this stage (Figs. 3a and 4). The characteristic ATP/ADP translocase which enables energy parasitism was gained here, as well as a transporter for polar amino acids, such as glutamine, and a tyrosine or tryptophan transporter. Thus, the LhRCA, as its extant descendants, most probably relied on a host for certain key metabolites but was also still able to synthesize a wide range of metabolites itself (Fig. 4 and Supplementary Data 4). We therefore hypothesize that the LhRCA was the first Rickettsiales to be an obligate symbiont. It did retain all flagellum and T4P genes, many of the chemotaxis-related genes and the ability to synthesize nucleotides de novo. Taken together, we suggest that it exhibited an obligate symbiotic yet extracellular lifestyle, which was conserved in its descendant *Ca. D. vastatrix*\(^{14}\), and possibly also in the Gamibacteraceae. However, based on current data, a facultative intracellular lifestyle cannot be excluded either. Its host was most probably a unicellular eukaryote, as the LhRCA was previously estimated, via molecular dating, to predate the origin of Metazoa by ~700 Myr and predicted, via ancestral trait reconstruction, to be protist-associated \(^{1}\). Irrespective, all known extant descendants of the LhRCA retained this obligate symbiotic lifestyle, but diversified to occupy distinct ecological niches.

**Key transitions in classical Rickettsiales.** From the LhRCA onwards, the evolutionary history of the classical Rickettsiales is generally characterized by reduction of central metabolic capabilities (Extended Data Figs. 2 and 5, and Supplementary Data 6). This trend is typical for obligate host-associated clades \(^{8,13,97}\). Yet, a number of genes linked to host association were gained as well. Below we discuss several observations that are of particular interest with respect to the evolution of host association.

Assuming an extracellular nature of the LhRCA, one of the key transitions that occurred during Rickettsiales evolution was the emergence of an intracellular lifestyle. Castelli et al.\(^{13}\) proposed that intracellularity in Rickettsiales evolved multiple times independently. Given our species tree (Fig. 2a), this would entail independent transitions to intracellularity in the last common ancestor of *Midichloriaceae* and *Anaplasmataceae* (LMIACA), and in the last common ancestor of the Rickettsiaceae (LRiCA). The extensive loss of a largely overlapping set of amino acid biosynthesis genes coincided with these transitions (Extended Data Fig. 5): the LMIACA lost the capacity to synthesize nine amino acids, and the LRiCA eleven amino acids. In contrast, the possibly ectosymbiotic Gamibacteraceae retained its amino acid biosynthesis pathways and the ectosymbiont *Ca. D. vastatrix* only lost three such pathways. This suggests that switching to a predominantly intracellular environment predisposed these lineages to losing many amino acid biosynthesis genes.

Similar to the emergence of an intracellular lifestyle, the association with animal hosts evolved independently in several lineages of classical Rickettsiales. Like the LhRCA, which was inferred to be a protist symbiont \(^{13}\), species from the Rickettsiaceae and *Midichloriaceae* and *Ca. D. vastatrix* are known to be associated with protists such as ciliates and amoebae. Wang and Luo\(^{14}\) inferred that the highly specialized insects and mammal symbionts
and pathogens of the Anaplasmataceae, Midichloriaceae and Rickettsiaceae evolved by frequent and independent transitions from protist-associated ancestors\(^1\)\(^2\). A similar host switch could have occurred in Gamibacteraceae as well (see above).

The hallmark feature of energy parasitism, the ATP/ADP translocase\(^9\), was lost twice independently as well: once in the last common ancestor of the Gamibacteraceae (LGCA) and once in the last common ancestor of the Anaplasmataceae (LACA). The Anaplasmataceae can produce their own ATP via oxidative phosphorylation using the electron transport chain\(^1\) and the Gamibacteraceae have the genetic potential for this as well (Extended Data Fig. 2). The lack of an ATP/ADP translocase in Gamibacteraceae suggests that this group may be less dependent on a host for energy and nutrients (Fig. 3a and Extended Data Fig. 2).

The genetic potential for storing carbon as polyhydroxybutyrate (PHB) was lost twice independently as well: once in the last common ancestor of the Gamibacteraceae and ‘Ca. D. vastatrix’ (LGDCA), and once in the LACA (Extended Data Figs. 2 and 5, and Supplementary Data 4). Rickettsial species may store carbon as PHB when host energy sources are unavailable\(^1\), for instance, when they are horizontally transmitted to a new host. PHB carbon storage is otherwise commonly found in marine heterotrophic bacteria\(^1\). Intriguingly, losing PHB carbon storage coincided in both cases with the loss of the flagellum (Extended Data Fig. 5) and may indicate a diminished capability for both ancestors to move and survive independent of a host.

DsbB, a protein previously proposed to be involved in oxidative folding of secreted Wolbachia effectors\(^5\), was gained by the LMiACA, as was a DnaJ-like chaperone. Genomes of the endosymbiotic Midichloriaceae and Anaplasmataceae experience an increased deleterious mutational load due to their intracellular nature\(^6\). This chaperone could potentially help stabilize enzymes to retain their function despite an accumulation of destabilizing mutations in their respective genes\(^6\). The LRiCA additionally gained several genes involved in host entry and manipulation, including the known family of autotransporters called surface cell antigens (Sca)\(^4\), the known Sec-TolC-secreted effector RARP-1\(^44\) and several copies of related ankyrin repeat-containing proteins (Extended Data Fig. 5 and Supplementary Data 4).

PhaB and FhaC, proteins involved in adhesin production and export\(^1\), were gained by the LGDCA (Extended Data Fig. 5 and Supplementary Data 4). The homologues in the ectosymbiont ‘Ca. D. vastatrix’ (DeiVas_00243/WP_146820351.1 and DeiVas_00245/ WP_161982782.1) were previously proposed to be involved in adhesion to cells of the host ‘Paramoecium primaurelia’\(^1\), as they are essential for host-cell adhesion in Bordetella species as well\(^6\). Their conservation in ‘Ca. D. vastatrix’ and Gamibacteraceae MAGs underpins a probably ancestral ectosymbiotic lifestyle of this group.

**Discussion**

The origin and emergence of host association during the early evolution of Rickettsiales remains poorly understood. Here we identify and reconstruct eleven MAGs distributed over the proposed deep-branching Mitibacteraceae and Athabascaceae and the more nested Gamibacteraceae. We propose that the Mitibacteraceae and Athabascaceae exhibit free-living or possibly facultative host-associated lifestyles, and are capable of forming biofilms. Their genomes lack the signatures of reductive evolution and contain the genetic repertoire for chemotactic motility, biofilm formation and, most importantly, a rich metabolism with complete biosynthetic pathways for all amino acids and nucleotides. Subsequent ancestral genome reconstruction analyses indicate a similar, free-living lifestyle for the LRCA, although a facultative host-association cannot be fully excluded. The rvh-type T4SS, characteristic for Rickettsiales, was acquired by the LRCA through horizontal gene transfer (HGT) from a proteobacterial donor. This ancestral T4SS may have functioned in so-called ‘bacteria
selected samples (125_MIX_0.22-3 and 067_SRF_0.22-0.45) were re-assembled. The preprocessed metagenomic reads from the two assemblies were cut every 10 kb, unless the remaining fragment was shorter than 20 kb. Then the preprocessed reads of a set of sequencing runs 125_MIX_0.22-3 (Supplementary Data 2) were mapped onto the fragmented contigs with KALLISTO v0.42.5 and yielded differential coverage profiles per fragmented contig. This was then used together with tetranucleotide frequency information by CONCOCT v0.4.0 to group the fragmented contigs into bins. Bins containing the Rickettsiales ribocontigs (Fig. 1) were then assessed and cleaned with MMGENOME (accessed June 2016) using differential coverage, GC composition, and read-pair linkage to determine which fragment contigs were associated with the two re-assembled metagenomes and from additional samples (Supplementary Data 1). The ribosomal proteins encoded on these contigs were then incorporated in a second RP15 dataset (alignment: MAFFT L-INS-i v7.471, alignment trimming: trimAl v1.4.rev15 -gt 0.5). To reduce compositional bias—a phylogenetic artefact to which alphaproteobacteria are particularly sensitive—20% of the sites that contributed most to compositional heterogeneity were removed 20% of the sites that contributed most to compositional heterogeneity. A phylogenetic tree was inferred from the compositionally trimmed supermatrix alignment with IQTREE v1.6.9 (-l LG+C60+F4+G, selected by ModelFinder, -b 1000 -nm 250; Supplementary Fig. 1). Binning of metagenomic contigs. For each metagenome assembly, contigs larger than 2 kb were grouped into bins on the basis of differential coverage across samples, tetranucleotide frequency profiles, GC composition and read-pair linkage as previously described. The contigs were cut every 10 kb, unless the remaining fragment was shorter than 20 kb. Then the preprocessed reads of a set of sequencing runs 125_MIX_0.22-3 and 067_SRF_0.22-0.45 were mapped onto the fragmented contigs with KALLISTO v0.42.5 and yielded differential coverage profiles per fragmented contig. This was then used together with tetranucleotide frequency information by CONCOCT v0.4.0 to group the fragmented contigs into bins. Bins containing the Rickettsiales ribocontigs (Fig. 1) were then assessed and cleaned with MMGENOME (accessed June 2016) using differential coverage, GC composition, and read-pair linkage to determine which fragment contigs were associated with the two re-assembled metagenomes and from additional samples (Supplementary Data 1). The ribosomal proteins encoded on these contigs were then incorporated in a second RP15 dataset (alignment: MAFFT L-INS-i v7.471, alignment trimming: trimAl v1.4.rev15 -gt 0.5). To reduce compositional bias—a phylogenetic artefact to which alphaproteobacteria are particularly sensitive—20% of the sites that contributed most to compositional heterogeneity were removed 20% of the sites that contributed most to compositional heterogeneity. A phylogenetic tree was inferred from the compositionally trimmed supermatrix alignment with IQTREE v1.6.9 (-l LG+C60+F4+G, selected by ModelFinder, -b 1000 -nm 250; Supplementary Fig. 1). Binning of metagenomic contigs. For each metagenome assembly, contigs larger than 2 kb were grouped into bins on the basis of differential coverage across samples, tetranucleotide frequency profiles, GC composition and read-pair linkage as previously described. The contigs were cut every 10 kb, unless the remaining fragment was shorter than 20 kb. Then the preprocessed reads of a set of sequencing runs 125_MIX_0.22-3 and 067_SRF_0.22-0.45 were mapped onto the fragmented contigs with KALLISTO v0.42.5 and yielded differential coverage profiles per fragmented contig. This was then used together with tetranucleotide frequency information by CONCOCT v0.4.0 to group the fragmented contigs into bins. Bins containing the Rickettsiales ribocontigs (Fig. 1) were then assessed and cleaned with MMGENOME (accessed June 2016) using differential coverage, GC composition, and read-pair linkage to determine which fragment contigs were associated with the two re-assembled metagenomes and from additional samples (Supplementary Data 1). The ribosomal proteins encoded on these contigs were then incorporated in a second RP15 dataset (alignment: MAFFT L-INS-i v7.471, alignment trimming: trimAl v1.4.rev15 -gt 0.5). To reduce compositional bias—a phylogenetic artefact to which alphaproteobacteria are particularly sensitive—20% of the sites that contributed most to compositional heterogeneity were removed 20% of the sites that contributed most to compositional heterogeneity. A phylogenetic tree was inferred from the compositionally trimmed supermatrix alignment with IQTREE v1.6.9 (-l LG+C60+F4+G, selected by ModelFinder, -b 1000 -nm 250; Supplementary Fig. 1). Binning of metagenomic contigs. For each metagenome assembly, contigs larger than 2 kb were grouped into bins on the basis of differential coverage across samples, tetranucleotide frequency profiles, GC composition and read-pair linkage as previously described. The contigs were cut every 10 kb, unless the remaining fragment was shorter than 20 kb. Then the preprocessed reads of a set of sequencing runs 125_MIX_0.22-3 and 067_SRF_0.22-0.45 were mapped onto the fragmented contigs with KALLISTO v0.42.5 and yielded differential coverage profiles per fragmented contig. This was then used together with tetranucleotide frequency information by CONCOCT v0.4.0 to group the fragmented contigs into bins. Bins containing the Rickettsiales ribocontigs (Fig. 1) were then assessed and cleaned with MMGENOME (accessed June 2016) using differential coverage, GC composition, and read-pair linkage to determine which fragment contigs were associated with the two re-assembled metagenomes and from additional samples (Supplementary Data 1). The ribosomal proteins encoded on these contigs were then incorporated in a second RP15 dataset (alignment: MAFFT L-INS-i v7.471, alignment trimming: trimAl v1.4.rev15 -gt 0.5). To reduce compositional bias—a phylogenetic artefact to which alphaproteobacteria are particularly sensitive—20% of the sites that contributed most to compositional heterogeneity were removed 20% of the sites that contributed most to compositional heterogeneity. A phylogenetic tree was inferred from the compositionally trimmed supermatrix alignment with IQTREE v1.6.9 (-l LG+C60+F4+G, selected by ModelFinder, -b 1000 -nm 250; Supplementary Fig. 1).
were performed using HMMER (-e 1e-05) for eukaryotic domains commonly found in Rickettsiales T4SS effectors (Ankyrin repeats: CL0465; leucin rich repeats: CL0022; Tetratricopeptide repeats: CL0020, Pentapeptide: CL0505) as well as in Xanthomonas bacterial-killing T4SS effectors (PGS domains: CL0244: Peptidase_M23; PF01551). All annotations are available from the Figshare repository (https://doi.org/10.6084/m9.figshare.c.5494977, see Data availability). We finally searched a selection of genomes for candidate sequences of vRf7 homologues using tblastn and references (rickettsiales+AND+vr7f) from Uniprot.

Alphaproteobacteria and mitochondria species tree. A phylogenomics dataset was constructed by updating the '24 alphamitoCOGs with more diverse mitochondria' dataset from Martijn et al.3 (henceforth ‘alphamito24’) with the 3 Rickettsiales MAGs reconstructed in this study, eight Rickettsiales MAGs identified from public MAG datasets4,5,6,7, ‘Ca. Deinariaea vastatrix’8, endosymbiont of Panenaea trichophorum, and endosymbiont of Acanthamoeba lipophora9,10, Phycrichococcus halmicosci11,12. Orthologues were identified by PSI-BLAST (E-value cut-off: 1e-10) using the gene alignments of the alphamito24 dataset as queries. Non-orthologues were detected and removed via single-gene tree inspections (alignment: MAFFT L-INS-i99; alignment trimming: trimAl v1.4.rev15 -gappyout100; phylogenetic inference: IQTREE 73 with 100 non-parametric bootstraps).

Orthologues were identified through PSI-BLAST (E-value cut-off: 1e-10) searches using the 129 gene alignments as a query, and non-orthologues were detected and removed via single-gene trees inspections as described above. A discordance filter13 was applied to remove the most discordant genes as follows: (i) single-gene alignments were prepared with MAFFT E-INS-i14 and trimAl v1.4.rev15 -gappyout100, (ii) single-gene trees were inferred with IQTREE v1.6.9 (with -bbn)15,16, (iii) bipartition count profiles were constructed from the bootstraps (tre_make_splits.pl) and compared between all possible gene pairs to calculate discordance scores (tre_discordance_two.pl). The two most discordant genes were removed from the dataset (Supplementary Fig. 2). After preliminary phylogenomics analyses with the remaining 116 genes, we decided to omit extremely long-branched taxa ‘Ca. Xenioslocculinum pacificiscenis’ and ‘Ca. Fokinia solitaria’17,18. An alignment was prepared using the corresponding eggNOG v.5.019 clusters for all Bacteria. First, we removed all Rickettsiales sequences from these clusters to avoid self-alignment. We then used the sequences from our taxon selection as queries to search the COGs using DIAMOND (–top 50–ultra-sensitive) and clustered the obtained sequences using cd-hit19 at 80% identity. We then aligned these reference sequences together with the sequences from our taxon sampling using MAFFT E-INS-i20 and applied a light gap trimming (trimAl v1.4.rev15 -gt 0.01)21. Finally, trees were inferred in IQTREE22 with automatic model selection23 and support values were estimated by 1,000 ultrafast bootstraps (see Data availability).

Ancestral gene content reconstruction. We reconstructed ancestral gene family repositories with the results from ALE by selecting all gene families predicted to be present at a given node with a frequency ≥0.3. Consensus annotations for gene families were inferred on the basis of the abovementioned gene annotations. Metabolic pathways or enzyme complexes were inferred as present if at least half of the corresponding gene families were present. We assessed the metabolic capabilities of ancestral genomes using the KEGG24 Module tool or MetaCyc pathways25.

Data availability. In addition to data available in the supplementary materials, files containing sequence datasets, alignments and phylogenetic trees in Newick format are archived at the digital repository Figshare: https://doi.org/10.6084/m9.figshare.c.5494977. MAGs generated in this study are linked to BioProject PRJNA746308. Accessions for genomes analysed in this study can be found in Supplementary Data 3. Publicly available datasets include eggNOG v.4.5 (eggnoavg5.embl.de), KEGG (kegg.jp), CAzy (cazy.org), TCDR (cdbr.org), Pfam (pfam.xfam.org), TIGRFAM and InterPro (ebi.ac.uk/interpro), NCBI nucleotide (ncl.nim.nih.gov/nuccore) and MetaCyc v26.0 (biocyc.org/MetaC).
the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) at Uppsala University and the Swedish National Infrastructure for Computing (SNIC) at the PDC Center for High-Performance Computing for providing computational resources. This work was funded by European Research Council Consolidator (817834), Dutch Research Council (VIC.192.016), Swedish Research Council (2015-04959), Volkswagen Foundation (96725) and European Union (H2020-MSCA-ITN-2015-675752) grants to T.J.G.E., and by Swedish Research Council (2018-06727) grant to J.M. We acknowledge J. E. Dharamshi for the original idea of the iterative $\chi^2$-trimming as well as for discussions regarding ALE.

Author contributions
T.J.G.E. conceived and supervised the study. J.M., J.V. and M.E.S. screened, assembled and binned metagenomic datasets. J.M. and M.E.S. performed phylogenomic analyses and ancestral genome reconstruction analyses. M.E.S., J.M. and S.K. analysed ancestral genome content. M.E.S., J.M., S.K. and T.J.G.E. interpreted the obtained results and wrote the first manuscript draft. All authors edited and approved the final version of the manuscript.

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/s41564-022-01169-x.
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-022-01169-x.
Correspondence and requests for materials should be addressed to Thijs J. G. Ettema.
Peer review information Nature Microbiology thanks Joseph Gillespie 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.

Open Access
This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.
© The Author(s) 2022
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Maximum likelihood phylogeny of Alphaproteobacteria and mitochondria. Based on the \( \chi^2 \)-trimmed (20% most heterogeneous sites removed) concatenated alignment of 24 highly conserved mitochondrially encoded proteins. ML tree was inferred under the PMSF approximation of LG+C60+F+Γ4 with 100 non-parametric bootstraps as implemented by IQ-TREE. Distinct alphaproteobacterial clades are given their own unique color, including the Rickettsiales (orange) and mitochondria (salmon). Rickettsiales MAGs identified and/or reconstructed in this study are given in black.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Overview of conservation of central metabolism, transporters, and electron transport chain (ETC) in Rickettsiales families and key ancestors. All enzymatic steps are depicted with enzyme commission (EC) numbers. The corresponding enzyme name and distribution in Rickettsiales can be found in Supplementary Data 4. Arrows between molecules and arrow heads indicate catalyzed enzymatic reactions and directionality, respectively. Presence/absence is indicated for extant genomes of Anaplasmataceae (light blue), Midichloriaceae (dark blue), Deianiraeaceae (light yellow), Gamibacteraceae (dark yellow), Rickettsiaceae (light violet), Athabascaceae (light orange), Mitibacteraceae (dark orange) as well as the reconstructed genomes of the ancestor of all obligate host-associated (classical) Rickettsiales (LHRCA, dark violet) and the ancestor of all Rickettsiales including Mitibacteraceae and Athabascaceae (LRCA, green). For the seven families, full squares represent absence in at most one member, half squares absence in more then one member but presence in some, and empty squares absence in all members. For reconstructed ancestral genomes, genes are inferred as present if the ALE program inferred an ancestral relative frequency of at least 0.3.
Extended Data Fig. 3 | Gene synteny of and domain architecture of derived pel exopolysaccharide gene cluster in free-living Rickettsiales. (a) Gene synteny plot of pelABCDEF gene cluster in the model organism Pseudomonas aeruginosa PAO1 and related clusters in selected free-living bacteria, and free-living Rickettsiales. Arrows indicate the genomic orientation and size of genes. Scale bar indicates a length of 2.5 kb. (b) Domain architecture of likely functional equivalents of the P. aeruginosa PelB and PelE proteins in the Rickettsiales bacterium Candidatus Mitibacter marchionensis.
Extended Data Fig. 4 | Frequency of proteins containing selected PFAM domain families. Domains typical of Rickettsiales effectors (Ank, LRR, TPR) and domains often found in Xanthomonas bacterial-killing effectors (Peptidoglycan binding domain PGBD, Peptidase M23) in the selected taxa used for phylogenomic and ancestral reconstruction. Domains were identified using HMMER hmmsearch (-E 1e-05) of the corresponding Pfam domain profiles.
Extended Data Fig. 5 | Key gains, losses and characteristics of host-associated Rickettsiales ancestors. Schematic phylogenetic tree depicting the relationships of the last common ancestor of host-associated Rickettsiales (LhRCA) and Rickettsiales families ancestors. Last common ancestor of Anaplasmataceae, Midichloriaceae, Deiniraea and Gamibacteraceae: LMIGDACA; Last common ancestor of Anaplasmataceae and Midichloriaceae: LMiACA; Last common ancestor of Anaplasmataceae: LACA, Midichloriaceae: LMiCA, Gamibacteraceae: LGCA; Rickettsiaceae: LRiCA. Gains (violet) and losses (green) of genes encoding for key characteristics are written above and below the corresponding branch, respectively. The ancestors are depicted as cells with inferred features such as the presence of a flagellum, Type 4 pilus, the capacity to synthesize amino acids and the number of inferred ancestral genes. Amino acid biosynthesis pathways are represented by the three-letter code of the produced amino acid. Proteins related to disulfide bond formation (DsbB); filamentous hemagglutinin production (FhaB), secretion and activation (FhaC); rickettsial ankyrin repeat protein (RARP-1); surface cell antigen (Sca); polyhydroxybutyrate (PHB) synthesis.
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.

- 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 | No software was used for data collection |
|-----------------|-----------------------------------------|

Data analysis

SEQPREP v1.3.2, TRIMMOMATIC v0.35, FASTQC v0.11.4, metaSPAdes (SPAdes 3.7.0), Prodigal v2.60, MAFFT v7.471, trimAl v1.4.rev15, IQTREE v1.6.9, v16.12, v2.0, KALLISTO v0.42.5, CONCOCT v0.4.0, MMGENOME (as on GitHub june 2016), Bowtie2 v2.3.3, CLARK-S v1.2.3, mComplete v.1.1.1, prokka v1.12, eggNOG-mapper v1.0.3, GhostKOALA (KEGG tools) v2.2, HMMER v3.3, InterProScan v5.32-78.0, MacSyFinder v2.0rc1, DIAMOND v2.0.6.144, PSI-BLAST v2.8.1+, BMGE v1.12, PREQUAL v1.01, Divvier v1.01, PhyloBayes MPI v1.8, ALE v0.4, BLASTN v2.8.1+, BLASTP v2.8.1+

Custom code:
- https://github.com/maxemil/rickettsiales-evolution
- https://github.com/maxemil/ALE-pipeline
- https://github.com/novigit/broCode

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

In addition to data available in the supplementary materials, files containing sequence datasets, alignments, and phylogenetic trees in Newick format are archived at the digital repository Figshare: 10.6084/m9.figshare.c.5494977. MAGs generated in this study are linked to BioProject PRJNA746308. Accessions for genomes analyzed in this study can be found in Supplementary Data 3. Publicly available datasets include eggNOG v4.5.1 (eggnoG45.embl.de/), KEGG (kegg.jp), CAZY (cazy.org), TCDB (tcdb.org), PFAM (pfam.xfam.org), TIGRFAM and InterPro (ebi.ac.uk/interpro/), NCBI nucleotide (ncbi.nlm.nih.gov/nucleotide/), MetaCyc v.26.0 (biocyc.org/META/)

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
- [X] 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 | Reconstruction of novel alphaproteobacterial genomes (MAGs). Phylogenomic analyses of Rickettsiales and related MAGs. Gene-tree species tree reconciliation and ancestral reconstruction of the last common ancestor of Rickettsiales. |
|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Research sample   | Available metagenomic datasets from the Tara Oceans consortium, published MAGs and reference genomes of Rickettsiales and alphaproteobacteria                                                                                                                                 |
| Sampling strategy | We selected particular metagenomic datasets of the Tara Oceans consortium and the other MAGs based on a phylogenetic screen of contigs containing ribosomal protein genes. Those datasets and MAGs that contained contigs related to Rickettsiales were selected       |
| Data collection   | N/A because the primary data collection was done by other parties (i.e. the Tara Oceans expedition and other research groups that collected the raw sequence data underlying the MAGs we selected)         |
| Timing and spatial scale | N/A because we did not do the primary data collection                                                                                                                                            |
| Data exclusions   | Certain Rickettsiales taxa were excluded from phylogenetic analyses as they were found to have extremely long branches and their inclusion would lead to untrustworthy results due to long branch attraction artefacts. Criteria for excluding were not pre-established |
| 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                                                                                                                                 |
| Blinding          | N/A because blinding was not required for the purposes of this study                                                                                                                                 |

Did the study involve field work? [ ] Yes [X] 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 | Methods |
|---------------------------------|---------|
| Involved in the study           | Involved in the study |
| Antibodies                      | ChIP-seq |
| Eukaryotic cell lines           | Flow cytometry |
| Palaeontology and archaeology   | MRI-based neuroimaging |
| Animals and other organisms     |         |
| Human research participants     |         |
| Clinical data                   |         |
| Dual use research of concern    |         |