











Host association and intracellularity evolved multiple times independently in the *Rickettsiales*

Received: 14 December 2022

Accepted: 18 January 2024

Published online: 06 February 2024

 Check for updates

Michele Castelli ¹, Tiago Nardi ¹, Leandro Gammuto², Greta Bellinzona ¹, Elena Sabaneyeva ³, Alexey Potekhin ^{4,5}, Valentina Serra ², Giulio Petroni ^{2,7}  & Davide Sassera ^{1,6,7} 

The order *Rickettsiales* (*Alphaproteobacteria*) encompasses multiple diverse lineages of host-associated bacteria, including pathogens, reproductive manipulators, and mutualists. Here, in order to understand how intracellularity and host association originated in this order, and whether they are ancestral or convergently evolved characteristics, we built a large and phylogenetically-balanced dataset that includes de novo sequenced genomes and a selection of published genomic and metagenomic assemblies. We perform detailed functional reconstructions that clearly indicates “late” and parallel evolution of obligate host-association in different *Rickettsiales* lineages. According to the depicted scenario, multiple independent horizontal acquisitions of transporters led to the progressive loss of biosynthesis of nucleotides, amino acids and other metabolites, producing distinct conditions of host-dependence. Each clade experienced a different pattern of evolution of the ancestral arsenal of interaction apparatuses, including development of specialised effectors involved in the lineage-specific mechanisms of host cell adhesion and/or invasion.


Rickettsiales are an early diverging¹ and ancient² alphaproteobacterial order. All the experimentally characterised members of this group engage in obligate associations with eukaryotic host cells³. The most long-term and thoroughly studied *Rickettsiales* include vector-borne pathogens, e.g., *Rickettsia* and *Anaplasma*, causing various diseases in humans and vertebrates^{4–6}, as well as *Wolbachia*, that can establish complex interactions with arthropod and nematode hosts⁷, chiefly reproductive manipulation and mutualism.

In recent years, our knowledge and understanding of the origin, evolution and diversification of *Rickettsiales*, as well as of the diversity of hosts and of interaction modes, have been remarkably improved. We can identify in particular three main advances. The first is the finding of a plethora of novel lineages (over 30 total genera described,

grouped into seven families^{8–11}), living in association with a wide variety of hosts^{12–14}. Most of those hosts are diverse aquatic unicellular eukaryotes (e.g., ciliates, amoebae, algae)^{15–22}, which have been deemed as probable ancestral hosts^{2,23,24}, even though these associations are still poorly investigated.

Second, “*Candidatus* *Deianiraea vastatrix*” (from now on, *Candidatus* will be omitted from taxonomic names, e.g., *Deianiraea vastatrix*), a fully extracellular *Rickettsiales* bacterium equipped with an unexpectedly large biosynthetic repertoire for amino acids, was discovered⁸, opening a new perspective on the evolution of *Rickettsiales*. While previous views implied that obligate intracellular association dated back to the last common ancestor of the order (“intracellularity early” hypothesis), this discovery cast doubt on those.

¹Department of Biology and Biotechnology, University of Pavia, Pavia, Italy. ²Department of Biology, University of Pisa, Pisa, Italy. ³Department of Cytology and Histology, Saint Petersburg State University, Petersburg, Russia. ⁴Department of Microbiology, Saint Petersburg State University, Petersburg, Russia.

⁵Research Department for Limnology, University of Innsbruck, Mondsee, Austria. ⁶IRCCS Policlinico San Matteo, Pavia, Italy. ⁷These authors contributed equally: Giulio Petroni, Davide Sassera.  e-mail: giulio.petroni@unipi.it; davide.sassera@unipv.it

Accordingly, another plausible scenario can be envisioned: obligate intracellularity could have evolved later and multiple times independently in different sublineages (“intracellularity late” hypothesis), together with a stronger dependence on host cells.

Third, metagenome binning recently allowed the discovery of further *Rickettsiales* sublineages, in particular two early-diverging families¹⁰. Their genetic repertoire (including nutrient uptake, detoxification, and multiple biosynthetic pathways) led the authors to hypothesise that these bacteria could be free-living, implying that adaptation to the interaction with host cells might have occurred in more “derived” *Rickettsiales* lineages.

However, many open points still exist on the origin and evolution of the interactions between *Rickettsiales* and their hosts. In particular, major aspects are whether the process(es) of transition towards obligate association and towards obligate intracellularity were one single or distinct phenomena, and whether each of them occurred “early” or “late”. In order to address such salient questions, in this study we collected a large and representative genomic dataset of *Rickettsiales*, thanks to de novo sequencing and selection of metagenomic sequences, thus identifying three novel families and remarkably extending the available diversity within previously known lineages. This allowed us to get a view on the diversity and evolution of host adaptation among *Rickettsiales*, finding multiple and convincing lines of evidence supporting the intracellularity late hypothesis.

Results

Novel genomes

In this work, we obtained nine complete genome sequences of *Rickettsiales* bacteria (belonging to nine species, eight genera, two families). These represent the first sequences for the respective species, with the exception of *Megaira polyxenophila*¹³. Thus, the evolutionary representativeness of available *Rickettsiales* genomes was significantly improved, also considering that all the newly sequenced organisms are hosted by ciliates or other protists. All the assemblies were highly curated, resulting in most cases in a very high quality (four genomes fully closed, five in total having L50 = 1, and one L50 = 2) (Supplementary Data 1), as confirmed by the comparison of BUSCO scores with typical ranges in *Rickettsiales* (Supplementary Data 2). In three cases, the quality of the assembly allowed to clearly determine the presence of plasmids (respectively in two *Rickettsiaceae* and one among *Midichloriaceae*, Supplementary Data 1).

Phylogeny

For the successive analyses, we aimed to capture and analyse the widest available diversity of *Rickettsiales* from published sequences, including under-explored (e.g., *Deianiraeaceae*) and possibly yet uncharacterised lineages. To do so, together with a representative set of 36 available *Rickettsiales* genomes, we selected a total of “high-quality” 314 potential *Rickettsiales* metagenome-assembled-genomes (MAGs) from various sources. Their affiliation to *Rickettsiales* was tested by a multi-step phylogeny-based approach. To this purpose, for each MAG lineage, a customised organism selection was employed, and a site-selection approach was applied to counterbalance compositional heterogeneity¹. After filtering out phylogenetically redundant MAGs, we identified 68 *Rickettsiales* MAGs, ending up with 113 total *Rickettsiales* for the final analysis, including the nine novel high-quality genomes.

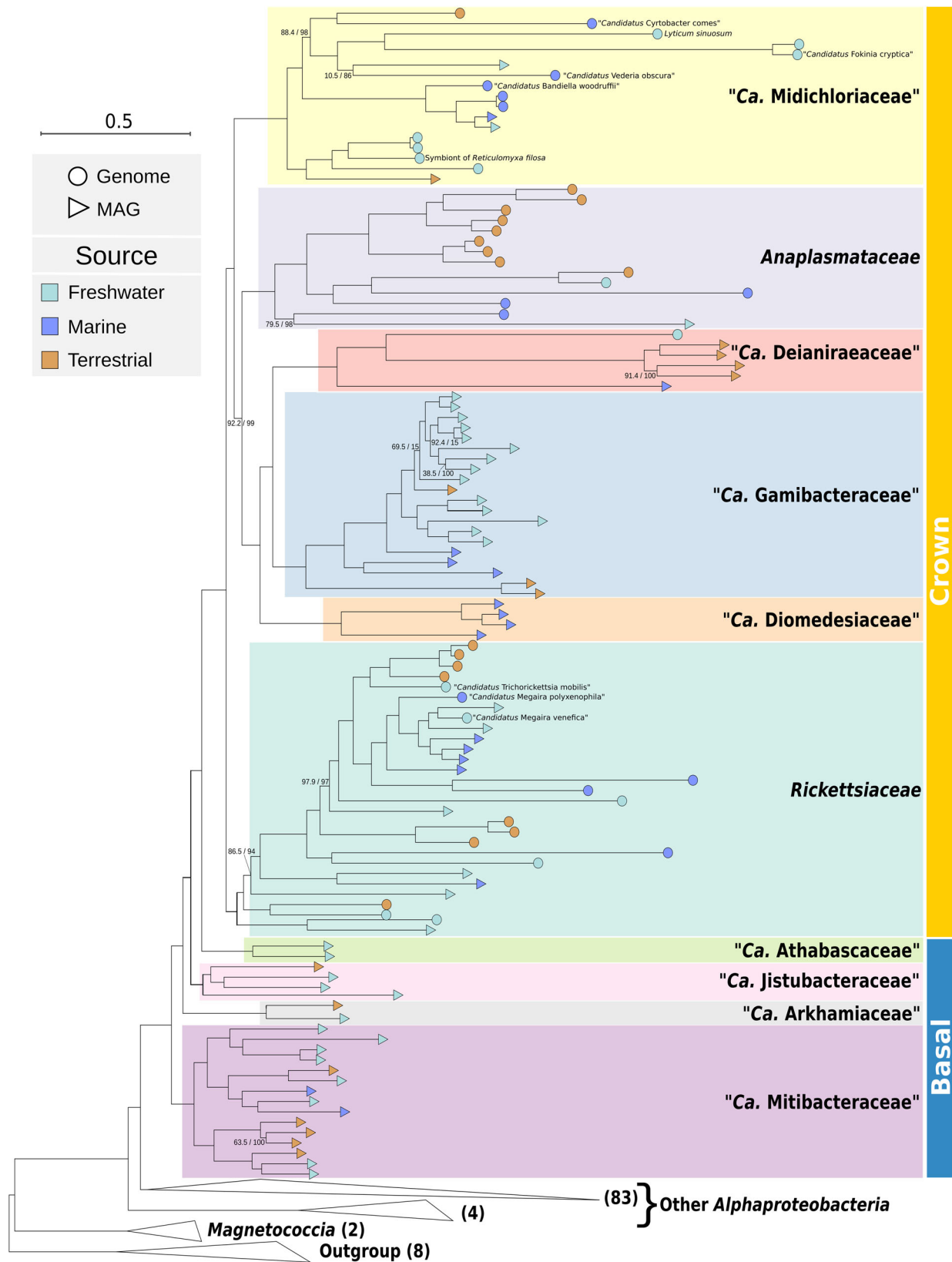
For this final dataset, inferences were performed on the “untreated” concatenated alignment and after applying compositional trimming (Fig. 1; Supplementary Fig. 18). As expected, the most significant differences among those trees pertained the placement of two alphaproteobacterial lineages, namely *Pelagibacterales* and *Holospirales*, which are well-known to be involved in artefacts due to compositional heterogeneity (e.g.,^{1,25,26}).

For what concerns the *Rickettsiales*, in the “untreated” dataset the resulting topology was quite robust, with most nodes finding full support, and only five nodes below the ultrafast bootstrap threshold of 90% (Fig. 1; Supplementary Fig. 18). Specifically, the four families that include at least one characterised organism were highly supported, and mostly with a significantly increased representativeness, thanks to novel genomes sequenced and MAGs identified in this study (novel taxa: 14/29 *Rickettsiaceae*; 10/17 *Midichloriaceae*; 1/14 *Anaplasmataceae*; 5/6 *Deianiraeaceae*). The inner relationships within *Rickettsiaceae* and *Anaplasmataceae* are overall consistent with previous phylogenetic and phylogenomic studies^{8,10,12,15,16,27}. For *Midichloriaceae*, it was possible to infer a novel inner topology with a much higher support with respect to previous studies^{14,19,20,27}. Furthermore, it was possible to determine that the *Midichloriaceae* bacterium associated with *Plagiopyla* represents a novel genus and species (Fig. 1; from now on, *Vederia obscura*, see Supplementary Note 1 for taxonomic description).

The three recently described families composed only by MAGs¹⁰, namely *Gamibacteraceae*, *Athabascaceae*, and *Mitibacteraceae*, were retrieved with a comparatively higher representativeness (Fig. 1). Besides, three further MAG-only families were identified for the first time in this study: (i) *Diomedesiaceae*, (ii) *Jistubacteraceae*, and (iii) *Arkhamiaceae*, the first being sister group of *Deianiraeaceae*+*Gamibacteraceae* (we will name these three families together as the “DDG”-clade), while the latter two forming a sequential branching pattern between *Athabascaceae* and *Mitibacteraceae* (Fig. 1; Supplementary Fig. 18, see Supplementary Note 1 for taxonomic descriptions).

The phyletic relationships among the families of *Rickettsiales* in our “untreated” dataset are in general consistent with the most recent and comprehensive studies^{8,10,18}, with a single partial exception. The “untreated” dataset indicates a closer relationship between the DDG-clade and *Anaplasmataceae* with respect to *Midichloriaceae*, consistent with 16S rRNA gene phylogenies and some phylogenomic analyses^{8,18}, while another study placed *Anaplasmataceae* and *Midichloriaceae* as sister groups¹⁰. Our inferences on the the most pronounced compositionally trimmed datasets (i.e., $\geq 30\%$ most heterogeneous sites removed) also produced a sister group relationship between *Anaplasmataceae* and *Midichloriaceae*, but with low supports (always below 90% ultrafast bootstrap; Supplementary Fig. 18), thus not being themselves fully conclusive. In the study by Schön and co-authors¹⁰, such alternative topology was obtained as well only after compositional trimming, and found maximal support only with Bayesian inference analyses. Therefore, compositional trimming and inference methods could be major reasons behind those different reconstructions, which may also have been potentially significantly influenced by other factors, such as the progressive and large increase of available members of the DDG-clade in successive studies (in particular, this is the first study considering *Diomedesiaceae*). Overall, it seems that the relationships between *Anaplasmataceae*, *Midichloriaceae*, and DDG-clade are not yet fully resolved, and may require further in-depth investigations. Here, we opted for the topology obtained without compositional trimming as “reference” for the following analyses (Fig. 1), considering it preferable to prevent the loss of phylogenetic signal which is an inevitable consequence of trimming. We posit that the scenario we present below for reconstructing the evolution of *Rickettsiales* is not affected by those, relatively minor, differences in topology.

In terms of origin, the vast majority of samples derived from aquatic environments (both freshwater and marine), especially in deeper nodes of the tree (Fig. 1). The same is true within each family, with the significant exceptions of *Anaplasmataceae* and *Deianiraeaceae*, mostly derived from terrestrial environments. Many characterised hosts resulted to be protists (most abundant in all families, except for *Anaplasmataceae*, all hosted by metazoans). For all MAGs, including in particular deeply branching lineages, no conclusive information is available on potential hosts, while in few cases there is loose indication of association, e.g., as originating from rumen^{28,29}.



General genome comparisons

Based on phylogeny, we hereby define as “crown *Rickettsiales*”, the smallest monophylum that comprises all characterised organisms (i.e., the one including the six families *Rickettsiaceae*, *Midichloriaceae*, *Anaplasmataceae*, *Deianiraeaceae*, *Gamibacteraceae*, *Diomedesiaceae*), thus corresponding to the classical *Rickettsiales* as defined

previously¹⁰. Conversely, the four other early diverging families will be defined as “basal *Rickettsiales*”.

Genome sizes are quite variable between and within *Rickettsiales* families (Supplementary Data 3). Crown *Rickettsiales* genomes are mostly in the range 1–1.5 Mb, with some appreciable differences between families, in particular on average *Anaplasmataceae* (1.1 Mb)

Fig. 1 | Phylogenomic tree of *Rickettsiales*. Maximum likelihood phylogenomic tree of 113 *Rickettsiales* inferred on 179 concatenated orthologs. Each *Rickettsiales* family is highlighted by a differently coloured box. At tips, round shapes indicate genome assemblies, while triangular shapes indicate metagenome-assembled-genomes (MAGs). Shape fillings show the sample source, namely light blue for freshwater, dark blue for marine, and orange for terrestrial. Due to space constraints, only the names of the nine newly obtained genome assemblies were reported, and non-*Rickettsiales* lineages (including other *Alphaproteobacteria*,

Magnetococcia, and outgroup) are represented by collapsed triangular shapes, with the respective number of organisms reported (full tree is shown in Supplementary Fig. 18). On each branch, support values by SH-aLRT with 1000 replicates and by 1000 ultrafast bootstraps are reported (full support values were omitted). Bars on the right-hand side indicate the crown and “basal” *Rickettsiales*, respectively. The tree scale stands for estimated sequence divergence. Ca. is an abbreviation for *Candidatus*.

and *Deianiraeaceae* (1.0 Mb), are smaller than others (all averages ≥ 1.3 Mb). Genomes from basal families are much larger (frequently >2 Mb, and on average ≥ 1.8 Mb, except *Arkhamiaceae*). It should be taken into account that the probable incompleteness of some MAGs (Supplementary Data 2) may influence the average size estimates, especially in DDG-clade and basal families. GC content is in general inversely correlated with genome size (32–36% on average in classical families and 40–52% in basal ones, with a maximum of 61%, in *Mitibacteraceae*).

As expected, gene numbers are consistent with respective genome sizes (Supplementary Data 3). We aimed to analyse the gene content and its variation along the *Rickettsiales* phylogeny. For this purpose, we reconstructed “homology groups” from the annotated genes and manually inspected the copy numbers in order to get insight into their evolutionary patterns (see Methods for details on the construction of “homology groups” and the differences with respect to the original eggNOG orthogroups). Our analyses supported the notions from previous studies that *Rickettsiales* have globally experienced genome reduction trends (Supplementary Fig. 19), as a putative consequence of adaptation and specialisation to host-associated lifestyles^{3,8,10,30,31}.

In order to investigate the origin and evolution of such interactions with host cells from a functional and metabolic perspective, we selected a number of relevant traits/functions, by inspecting the global patterns of “homology groups” (Supplementary Fig. 19). The selection included in particular biosynthesis and uptake of metabolic precursors (i.e., amino acids and nucleotides), secretion/adhesion/motility apparatuses and putative effector molecules, which were addressed by further dedicated analyses (see below). A detailed overview of gene content variation and evolution in *Rickettsiales*, with a special focus on general “family-level” trends, as well as on the single newly characterised genomes, is presented in (Supplementary Note 2). Such an approach also allowed to prevent misleading conclusions due to incompleteness of single MAGs (Supplementary Data 2).

Secretion, attachment, and motility

Secretion systems are among the components that may exert a central role in regulating interactions with host cells³², potentially ensuring specific stages of the bacterial life cycle through the delivery of effectors. In *Rickettsiales*, the hallmark apparatus is type IV secretion system (Supplementary Fig. 20), representing a probable ancestral horizontal acquisition³³, and a possible prerequisite for establishing interactions with host cells¹⁰. Only few genomes among *Rickettsiales* are devoid of this apparatus^{16,34}, and few others display an incomplete gene set, possibly indicative of ongoing loss, in particular *Bandiella* and the *Midichloriaceae* symbiont of *Reticulomyxa* (Supplementary Fig. 20). Type VI secretion system is very rare, being found only in two *Rickettsiaceae* (Supplementary Fig. 20), in both cases encoded on plasmids, a probable indication of horizontal acquisition. In *Sneabacter* this system has possibly functionally replaced the type IV secretion system¹⁶, while in *Trichorickettsia* both systems coexist.

Among putative secreted effector proteins (Supplementary Fig. 21), the “repeat-bearing” ones (ankyrin, tetratricopeptide, leucine-rich or pentapeptide repeats) are overall abundant and enriched in crown families, with many lineage-specific patterns, but are also present in basal families. On the other hand, RTX toxins^{35,36} are quite

abundant in basal families, and uncommon in crown ones. Interestingly, proteins involved in the intracellular invasion of eukaryotes, such as hemolysins, patatin-like and other phospholipases³⁷, are common in some crown families such as *Rickettsiaceae* and *Midichloriaceae*, and not uncommon in basal families, while they are rare in DDG clade, especially *Deianiraeaceae* (Supplementary Fig. 21). Several other putative toxins/effectors, characterised in *Rickettsiales*³⁸ and/or in other bacteria^{13,39–45}, were found more rarely, showing patterns of presence/absence that appear to be quite lineage-specific, and without sharp differences between basal and crown *Rickettsiales* (Supplementary Fig. 21).

The flagellum might be important especially during horizontal transmission in *Rickettsiales*^{20,46}. Flagellar genes are common in basal *Rickettsiales* (Supplementary Fig. 22), likely representing ancestral traits⁴⁶. Conversely, they are absent in the DDG-clade, and are very rare in *Anaplasmataceae* (found just in the aquatic *Echinorickettsia* and *Xenolissoclinum*). Within families *Rickettsiaceae* and *Midichloriaceae*, they are extremely rare in terrestrial representatives (the only cases being *Midichloria mitochondrii* and the *Rickettsiaceae* symbiont of *Amblyomma* Ac37b), but quite common in aquatic ones, in particular basal *Rickettsiaceae* and *Midichloriaceae* in general, thereby also confirming the few experimental observations of flagella^{47–49}. The poor correlation of the presence of flagellar genes with *Rickettsiales* phylogeny (including differential cases within the same genus, such as *Megaira* and *Midichloria*³⁴) would be indicative of multiple independent reduction/loss events (Supplementary Fig. 22). Similar considerations may be true for chemotaxis, which possibly works in conjunction with flagella for host targeting⁵⁰, and is present only in basal *Rickettsiales* and in the basal members of the family *Rickettsiaceae* (Supplementary Fig. 22).

Type 4 pili may be involved in the adhesion/attachment to host cells in *Rickettsiales*⁸. Their components are present in basal *Rickettsiales*, in the DDG clade, and only rarely in the other crown families, especially in the respective early-divergent and/or aquatic representatives (Supplementary Fig. 23). Thus, this apparatus was likely ancestral, experiencing multiple independent losses in crown *Rickettsiales*.

Proteins homologous to the FhaBC two-partner secretion system, which were tentatively linked to the attachment and toxicity towards host cells in *Rickettsiales*⁸ and may also participate in bacterial competition⁵¹, were likely ancestral and lost multiple times among *Rickettsiales*, being present in the basal families, *Deianiraeaceae*, *Gamibacteraceae*, and very few representatives of the other crown families (Supplementary Fig. 23).

For what concerns proteins characterised to be involved in adherence/invasion of host cells^{52,53} or even immune evasion⁵⁴ in *Rickettsiales*, they are present (almost) exclusively in subgroups of the respective families (Supplementary Fig. 23).

Nucleotide and amino acid metabolism

Nucleotide biosynthesis is stably present (both purines and pyrimidines) in basal *Rickettsiales* (Fig. 2; Supplementary Fig. 24), while, differently from other biosynthetic pathways (Supplementary Note 2), its presence in crown *Rickettsiales* is “scattered” along the organismal phylogeny. Indeed, it is ubiquitous in the families *Gamibacteraceae*, *Diomedesiaceae*, *Anaplasmataceae*, present in the earliest diverging

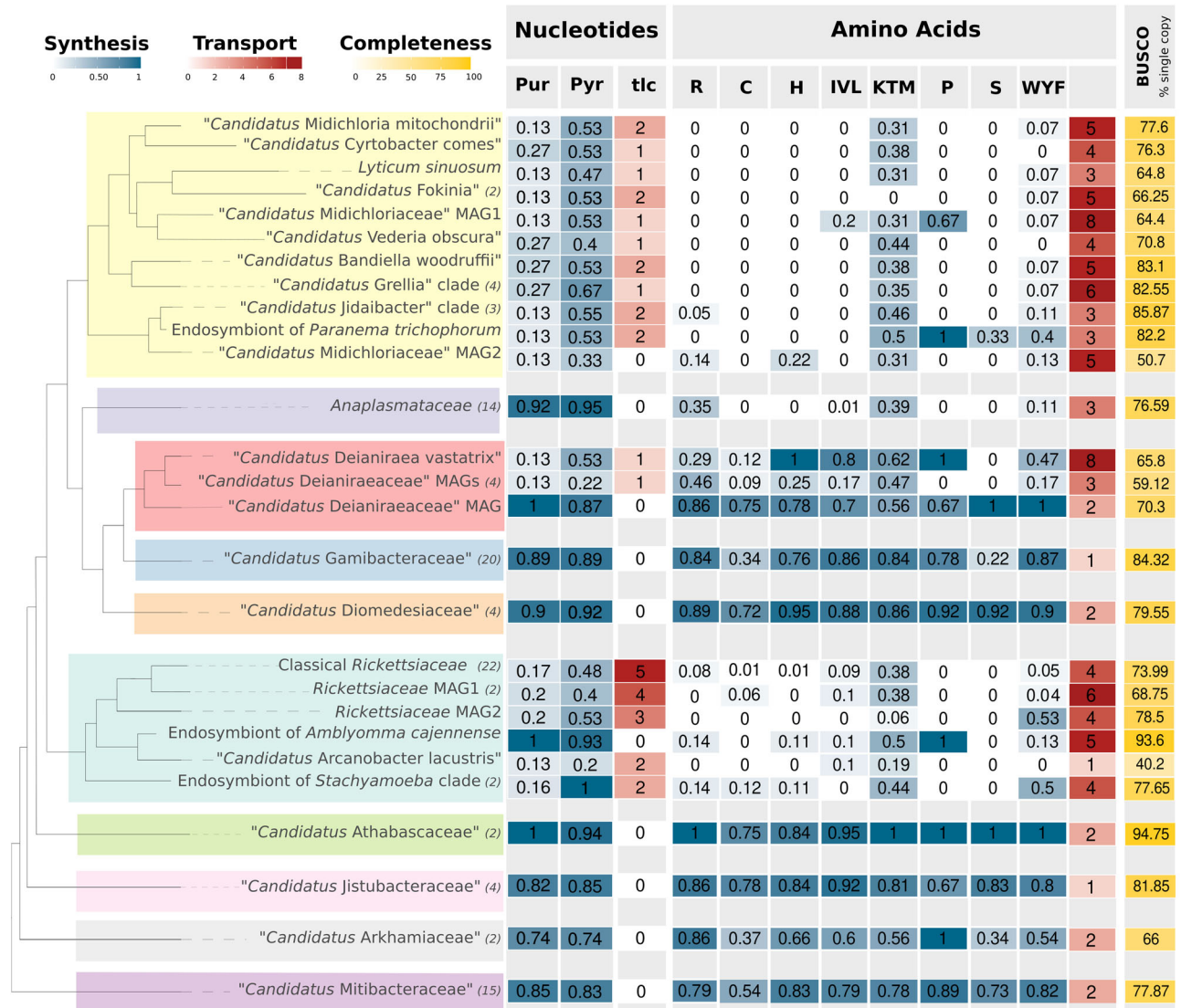


Fig. 2 | Biosynthesis and transport of nucleotides and amino acids. Heat-map showing the presence and abundance of biosynthetic pathways (blue) of nucleotides (purines and pyrimidines) and amino acids (grouped by to their mutually shared enzymatic steps according to BioCyc⁵⁵), as well as their respective transporters (red). For biosynthesis, the proportion of the total genes of the pathway is shown (Supplementary Data 7), while for transporters, the number of genes is reported, in particular for amino acid transporters the sum of the “characterised

hits” (Supplementary Fig. 28). A cladogram of the organisms is shown on the left, with each *Rickettsiales* family highlighted by a differently coloured box. Due to space constraints, selected monophyletic clades with homogeneous gene content were collapsed. For each clade, the number of organisms is shown in brackets (if higher than one), and reported values are averaged, while the complete organism sets are shown in (Supplementary Fig. 24, 28).

Deianiraeaceae MAG and in few basal *Rickettsiaceae* (in particular the endosymbiont of *Amblyomma* Ac37b, able to synthesise both purines and pyrimidines), but, besides few genes, absent in *Midichloriaceae* and in the remaining and most numerous *Rickettsiaceae* and *Deianiraeaceae*. Single-gene phylogenies indicate that the respective pathways are likely ancestral in *Rickettsiales*, with quite good correspondence with organismal phylogeny at various levels (up to and even above the families; Supplementary Fig. 25). Few exceptions represented by potential non-orthologs were observed (Supplementary Data 4) and excluded for the phylogenies on the concatenated pathways, which confirmed the same trends observed in single-gene trees (Supplementary Fig. 26). Some relatively minor exceptions to the correspondence with organismal phylogeny (supportive of ancestry) pertain to the inner relationships among *Rickettsiales* families, namely the placement of the DDG clade (as well as sometimes *Midichloriaceae* or *Anaplasmataceae*) as close relatives of *Rickettsiaceae*, though often not with very high support (Supplementary Fig. 26). This

potentially represents an artefact, due to the fast-evolving rate of the sequences, similarly to the respective organismal phylogeny. In addition, in some cases a few non-*Rickettsiales* sequences (e.g., *Pelagibacteriales* and some representatives of *Holosporales* in several concatenated purine biosynthesis trees) were “nested” within *Rickettsiales*, thus representing possible “residual” artefacts even after applying compositional trimming (potentially resulting from long-branch attraction), or, hypothetically, the results of horizontal transfer from *Rickettsiales* to those bacteria.

In the crown *Rickettsiales* lacking nucleotide biosynthesis, this absence is counterbalanced by the ability to obtain final products (or intermediates) from their hosts. Indeed, the presence/absence pattern of *tlc* nucleotide translocases almost perfectly inversely correlates with nucleotide biosynthesis (Fig. 2; Supplementary Fig. 24), with as noticeable exception a *Midichloriaceae* bacterium MAG (GCA_013288625) that is devoid/depleted in both, but also presents a low BUSCO estimated completeness. This family of transporters

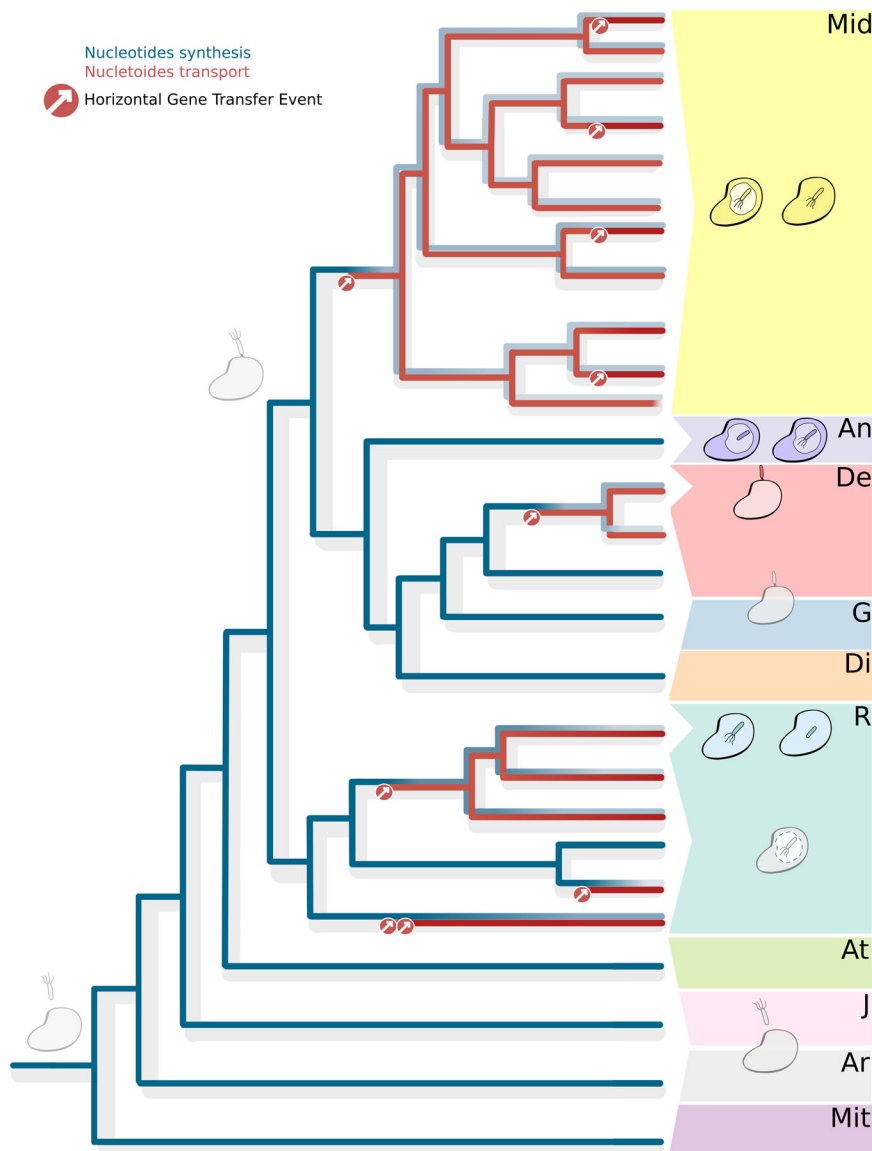


Fig. 3 | Main steps in the evolution of *Rickettsiales*. Reconstruction of the main steps of the evolution of *Rickettsiales* with a specific focus on the interactions with and dependence on eukaryotic cells. A cladogram of the main lineages (as represented in Fig. 2) is shown. The case of the biosynthetic pathways (averaged for purines and pyrimidines, blue) and tlc transporters (red) for nucleotides is represented along the tree by a heat-map-like representation, showing the inferred ancestral conditions and hypothesised steps of variation on each branch. In particular, multiple independent acquisitions of transporters by horizontal gene transfer events (red circles with inward arrows) would have led to the progressive reduction/loss of the biosynthesis. *Rickettsiales* families are highlighted by

coloured boxes and by abbreviated names (R: *Rickettsiaceae*; Mid: *Midichloriaceae*; An: *Anaplasmataceae*; De: *Deianiraceae*; G: *Gamibacteraceae*; Di: *Diomedesiaceae*; At: *Athabascaceae*; J: *Jistubacteraceae*; Ar: *Arkhamiaceae*; Mit: *Mitibacteraceae*). At tips, groups of tips, and at selected nodes, the drawings represent the known (coloured) or hypothesised (grey) features of the bacteria and their interaction with eukaryotic hosts, in particular, intracellular or extracellular associations, or lack of association, as well as presence/absence of a vacuole and of flagella. Multiple side-by-side pictures represent alternative conditions/reconstructions for the same organisms.

include chloroplast ATP/ADP translocases, as well as a vast array of proteins, able to translocate several different nucleotides^{56,57}, and previously reported to have experienced multiple horizontal gene transfer events between phylogenetically unrelated host-associated bacteria^{58,59}.

Comparison of tlc and organismal phylogenies (Supplementary Fig. 27) indicates that these transporters were acquired multiple independent times (up to ten) by different *Rickettsiales* lineages: once among *Deianiraceae*, three-five among *Midichloriaceae*, three-four among *Rickettsiaceae* (see Fig. 3). We also detected multiple independent events of duplication leading to several paralogs, namely five copies in classical *Rickettsiaceae*, two-four in three distinct basal *Rickettsiaceae* sublineages, and two copies in the

Jidaibacter lineage among *Midichloriaceae* (Figs. 2,3; Supplementary Figs. 24, 27).

The presence/absence pattern of biosynthetic pathways for amino acids shows significant analogies with the one we detected for nucleotides (Fig. 2; Supplementary Fig. 28). Indeed, they are quite uniformly present in basal *Rickettsiales*, *Diomedesiaceae*, *Gamibacteraceae*, and partly *Deianiraceae*. Conversely, they are very rare or fully absent (besides few genes) in the non-monophyletic assemblage of *Rickettsiaceae*, *Midichloriaceae* and *Anaplasmataceae*, with very few exceptions, such as arginine in *Ehrlichia* and *Neoehrlichia* (and partly *Anaplasma*), and proline in the endosymbionts of *Amblyomma* Ac37b and *Peranema* (Supplementary Fig. 28). As in the case of nucleotide synthesis, single-gene phylogenies indicate an overall vertical descent

of these pathways (Supplementary Fig. 25), with few exceptions. The most notable case pertains to cysteine biosynthesis genes, for which basal and crown *Rickettsiales* sequences have different phylogenetic positions, which may suggest recent HGT event(s) in crown *Rickettsiales* or an ancestral paralogy. Among the genes of all the other amino acid biosynthetic pathways, only a limited number of potential non-orthologs among *Rickettsiales* were observed (Supplementary Data 4), which were excluded for the phylogenies on the concatenated pathways (Supplementary Fig. 26). Such phylogenies further corroborated the ancestrality of the biosynthetic pathways for amino acids, with relatively minor differences with respect to the organismal phylogenies, such as the relationships among *Rickettsiales* families. As in the case of nucleotide biosynthesis, these may be residual artefacts after compositional trimming, potentially resulting also from fast-evolving rates of *Rickettsiales* sequences, especially when considering the representatives of the DDG clade.

Besides such potential cases of lineage-specific acquisition of amino acid biosynthesis genes by some representatives of *Rickettsiales* presented above or previously proposed⁶⁰, the analyses presented here provide support for “general” vertical inheritance of most amino acids biosynthesis, when present, in most of the *Rickettsiales*, directly from the last common *Rickettsiales* ancestor.

The array of putative amino acids transporters is more complex than that for nucleotides, considering the higher number of amino acids and the multiple independent transporter families with variable substrate specificities, many of which belong to larger gene families of transporters with broader specificity^{61,62}. This impairs precise homology-based prediction of substrate specificity in *Rickettsiales*, making it impossible to define with certainty which amino acid is imported by which transporter. The same reasons make an accurate inference of informative gene phylogenies infeasible. Nevertheless, the total number of transporters is much higher in the three families that are more deprived in biosynthesis (Supplementary Fig. 28). This is especially true for those with a higher similarity with ascertained amino acid transporters (Fig. 2). Moreover, as seen for nucleotide transporters, the presence of homologues of different amino acid transporters is scattered along the *Rickettsiales* phylogeny (Supplementary Fig. 28).

We believe that the data presented above clearly indicate a pattern of multiple independent and successive acquisitions of nucleotide and amino acids transporters in different lineages of *Rickettsiales*, events that would have enabled the recipients to efficiently acquire those compounds from their hosts, thus leading to the reduction and eventual loss of the respective biosynthesis, an evolutionary scenario consistent with the intracellular late hypothesis⁸.

Discussion

The knowledge and understanding of the evolution of the typically host-associated *Rickettsiales*^{3,27}, in particular its earlier steps, has been hampered by the limited and phylogenetically unbalanced set of genomes available. Here we present a dataset of over one hundred phylogenetically-diverse *Rickettsiales* assemblies, thanks to de novo-sequencing of nine high-quality genomes from underexplored protist-associated representatives, and to an accurate selection of published genomes and MAGs. We thus obtained an enriched representativeness of all known families, including the recently described ones¹⁰, and identified three further ones, for a total of ten families in *Rickettsiales*. Leveraging such an extended taxonomic resolution, we investigated whether the obligate association with eukaryotic hosts and specifically the intracellular lifestyle were “early” conditions with a single origin, or “late” achievements that evolved multiple times independently in different *Rickettsiales* sublineages.

Evolution of metabolic dependence

It is a generally accepted notion that metabolic dependence on the hosts is a key feature in obligate associations such as those involving

multiple independent lineages of bacterial and eukaryotic intracellular parasites^{1,63–65}, including the *Rickettsiales*^{8,10,30,31}, that evolved as the consequence of the possibility to efficiently acquire metabolites (including precursors such as amino acids and nucleotides, and, for energy, ATP). Interestingly, such ability is likely due to the acquisition of suitable transporters, making the respective biosynthetic and catabolic pathways dispensable, thus leading to their reduction and eventual loss, concurring in determining a host-dependent condition^{63,66}. The pattern of gain of transporters and loss of synthesis pathways can thus be a strong indicator of the state of host dependence through evolution. Based on the herein produced dataset and analyses, the case of nucleotide synthesis and transport is noteworthy among *Rickettsiales*. Indeed, our analysis indicate that the most likely scenario is one of multiple independent horizontal acquisitions (up to ten) of tlc transporters among crown *Rickettsiales*, likely “triggering” independent losses of the ancestral biosynthetic pathways (Figs. 2 and 3; Supplementary Figs. 24, 25, 26, 27). Similar considerations hold for amino acids, even though the impossibility to predict the precise specificity of all transporters impairs a clear reconstruction of single events leading to the multiple independent losses of the ancestral biosynthetic pathways (Figs. 2 and 3, Supplementary Figs. 25, 26, 28).

Besides these more clear-cut cases, detailed analyses of the presence/absence patterns among the genes involved in multiple other pathways (including glycolysis, gluconeogenesis, pentose-phosphate pathway, Krebs cycle and electron transport chain, synthesis of cofactors, lipids, peptidoglycan, lipopolysaccharide, and polyhydroxyalkanoate granules; see Supplementary Note 2) strongly indicate analogous processes of gradual and independent reduction/losses in different crown *Rickettsiales* sublineages. Sharp differences are present even within single families, such as in the metabolically-rich basal *Rickettsiaceae*, as compared to the classical streamlined ones.

Thus, in contrast with the more traditional views^{10,20,24}, our analyses provide a clear indication that processes of pathway reduction/loss have not taken place just once in *Rickettsiales*, but instead occurred (and are still possibly occurring) multiple times independently in different crown *Rickettsiales* lineages, also in relation with the host features. This hints towards an independent origin of obligate host-associations, and thereby intracellularity, among the *Rickettsiales* (see below section “Evolutionary trajectories of the interactions”).

Evolution of interaction mechanisms

Secretion systems and attachment/invasion molecules are other paramount bacterial components for promoting and actively regulating interactions with host cells^{32,33,67}. Interestingly, we found that in crown *Rickettsiales* the repertoire for such systems, as well as for the flagellar apparatus, is a substantial subset of the one of basal *Rickettsiales*. This result fortifies previous notions that the ancestral *Rickettsiales* already possessed a quite rich set of proteins to interact with (unicellular) eukaryotes^{10,33}. It seems reasonable to hypothesise that such arsenal could have represented a prerequisite for the establishment of associations with eukaryotic cells, possibly through a partial process of repurposing¹⁰ (e.g., delivery of effectors molecules active on eukaryotes, motility and chemotaxis involved in horizontal transmission). In terms of being instrumental for the evolutionary origin and maintenance of the associations, type IV secretion is a good candidate^{10,33}, also considering its almost full conservation among *Rickettsiales* (Supplementary Fig. 20). Other apparatuses, e.g., flagellum and type IV pilus/type II secretion, while widespread in “basal” *Rickettsiales*, are more phylogenetically scattered among crown *Rickettsiales* (Supplementary Figs. 22, 23), which is likely a result of multiple independent losses as well. This indicates unique patterns of specialisations along the evolution of *Rickettsiales*, with the concurrent lineage-specific losses of traits that were dispensable for the

interaction with respective host cells. Interestingly, the correlation with phenotypic traits suggests some functional links, such as flagella and chemotaxis in aquatic environments, likely involved in non host-associated stages such as horizontal transmission²⁰, pili for extracellular attachment to host cells in *Deianiraea*⁸ and possibly in other members of the DDG clade. At the same time, alternative/additional functions for these apparatuses due to their homologies with secretion systems^{68,69} could be possible, and may account for the exceptions to such correlation patterns^{34,46}.

Conversely, specialisation in the interaction with host cells has likely implied the expansion of other gene families, in particular those of putative secreted effectors, such as the “repeat-containing” ones, or acquisition/development of novel ones. In particular, several proteins characterised in pathogenic *Rickettsiales* (e.g., *Rickettsia*, *Anaplasma*) for their direct involvement in the interaction with the host^{38,52–54,70}, resulted to be lineage-specific (at the family level or even below) rather than conserved in *Rickettsiales* as a whole (Supplementary Fig. 23). Thus, it seems likely that many other still uncharacterised lineage-specific proteins could exist in the other much less investigated *Rickettsiales* (e.g., *Midichloriaceae*, DDG clade). Such a scenario of lineage-specific sets of “interactors” suggests that the mechanisms and conditions of host-association have evolved independently among different (crown) *Rickettsiales* lineages along with their molecular players.

The absence of experimental data makes it more difficult to precisely infer the condition of basal *Rickettsiales* in terms of potential interactions with eukaryotes. The study discovering the first two basal families (*Mitibacteraceae* and *Athabascaceae*) found that these bacteria are metabolically rich, which is suggestive of independence from possible host cells¹⁰. Herein, such pattern was fully confirmed in the extended sampling of these two families and in the representatives of the novel ones (*Arkhamiaceae* and *Jistubacteraceae*). At the same time, we found that “basal” *Rickettsiales* bear basically all the putative prerequisite apparatuses for the interaction. Most significantly, many are also equipped with homologues of effectors typical of crown *Rickettsiales*, such as phospholipases³⁷ and the “repeat-containing” effectors, and they are even selectively enriched in potential additional effectors^{35,36}. Conversely, free-living-like traits such as inorganic nutrient transport and detoxification, previously found only among basal *Rickettsiales*¹⁰, were retrieved also in some crown representatives (Supplementary Note 2). This may be indicative that those crown *Rickettsiales* retain some “primitive” traits, and, more in general, as another hint at more complex evolutionary trajectories than a simple transition towards obligate association at the root of crown *Rickettsiales*.

Evolutionary trajectories of the interactions

Taken together, the data presented above clearly indicate that obligate host-association was most likely a “late” condition in *Rickettsiales*. Under such a scenario, we propose that at some point in the early *Rickettsiales* evolutionary history their presumably aquatic free-living ancestors were engaged in some kind of facultative interaction with eukaryotes. The starting point could have been defence from protist predators through the release of active effectors, as previously hypothesised¹⁰. It is possible to envision that such defence mechanisms successively paved the way for the (gradual) development of the (at least occasional) ability to gain further advantages by such interactions. Specifically, they could have become capable of acquiring metabolites from the damaged/killed eukaryotes, somehow reversing and taking control of the predator-prey interactions. The lifestyle of *Deianiraea* could be partly reminiscent of this hypothetical condition⁸. Most likely, such transition towards facultative associations would have occurred prior to the last common ancestor of crown *Rickettsiales*, which are all obligatorily host-associated.

Conversely, it is not straightforward to precisely place an upper bound for such transition, given the complete lack of direct

information on the lifestyles of extant basal *Rickettsiales*. It could have occurred sharply in the common ancestor of crown *Rickettsiales*, or could have been more nuanced, involving also the ancestors of some basal lineages, possibly up to the ancestor of all *Rickettsiales*. Nevertheless, it cannot be excluded that any potential host-associated representative of basal *Rickettsiales* could be the result of a convergent and independent evolution with respect to crown *Rickettsiales*.

In any case, for what concerns crown *Rickettsiales*, we can envision that a single initial facultative association would have differentiated in the descendants, becoming tighter and tighter (and at some point, obligate) through parallel successive steps of acquisition/development of metabolite transporters and interactor molecules. Such further transitions would have occurred separately and independently in different *Rickettsiales* lineages, thus supporting the conclusion of a late origin for the obligate association with hosts. The order and kind of such steps, although somehow similar, would have been unique in each of the different crown *Rickettsiales* phyletic lines, as reflected in the present-day lineages, which exhibit differential features of metabolic dependence (e.g., for most amino acids but not nucleotides in *Anaplasmataceae*, and vice versa in most *Deianiraeaceae*), as well as differential mechanisms and conditions of interaction (chiefly intracellularly *vs* extracellularly).

Obligate intracellularity is a well-documented condition in most of the characterised crown *Rickettsiales*, namely the non-monophyletic assemblage of *Rickettsiaceae*, *Midichloriaceae*, and *Anaplasmataceae*. Conversely, we previously showed that *Deianiraea*⁸, while obligatorily host-associated, is not intracellular, providing the first basis to infer a possible alternative “intracellularity late” hypothesis for *Rickettsiales*. Obligate intracellular bacteria are inherently obligatorily host-associated, and, as such, intracellularity is one among the possible features evolved by bacteria living in obligate host association. Thus, the data and analyses above supporting a “late” obligate host-association similarly represent additional support for the previously proposed⁸ “intracellularity late” hypothesis. Accordingly, it seems most probable that obligate intracellularity would have evolved multiple independent times and with differential features only in some of the obligatorily host-associated crown *Rickettsiales*. The members of the DDG clade other than *Deianiraea*, all represented by MAGs, as based in their genome features (e.g., repertoire of metabolic pathways at least equivalent to *Deianiraea*, and comparable set of putative adhesins such as those of the exoprotein family; Supplementary Figs. 23, 24, 25, 26, 28), could be deemed as extracellularly host-associated as well (or even in some cases potentially not obligatorily host-associated), thus being consistent with such a scenario.

Our reconstruction may also provide a novel perspective on the origin and evolution of other host-associated bacterial lineages that, similarly to *Rickettsiales*, present the prerogative to “hold the control” on the interaction and to switch hosts by horizontal transmission, and were thus termed “professional symbionts”⁷¹ (e.g., *Chlamydiae*⁷², *Legionellales*⁷³ and *Holosporales*³). These lineages could share similarities in the initial establishment and successive stepwise and “late” evolutionary development.

Final remarks

From a more general perspective, it seems worth to compare *Rickettsiales* (and possibly other “professional symbionts”) with the more “canonical” genome evolution models among obligate symbionts, namely nutritional mutualists in insects (with some parallels also in protists⁷⁴). Such symbionts undergo relatively rapid streamlining as a result of an initial host restriction, followed by a more or less prolonged stasis, and are somehow “doomed” to extinction after replacement. Conversely, “professional symbionts” would be the controllers of the interaction from its evolutionary beginning, retaining the ability to horizontally change hosts, and undergoing much more gradual and “flexible” streamlining processes, depending also on

the external environmental conditions and not just on the features of a single host. We can also observe significant differences in the metabolic interchanges with their hosts, with nutritional symbionts providing metabolites such as amino acids to their hosts, while *Rickettsiales* (and other “professional symbionts”) “stealing” the same metabolites from the hosts. Interestingly, the abundance of lineages preferentially associated with marine invertebrates and showing poor co-cladogenesis with their hosts⁷⁴ indicates that there may be more bacteria sharing traits of “professional symbionts” than currently recognised.

Large-scale comparative genomic analyses such as those presented here and elsewhere^{10,72,73} have huge potential to provide major advances in our understanding of functional traits and the underlying evolutionary processes. However, they also face inherent predictive limits. In particular, while they are quite suitable for deriving metabolic dependencies, they are not likewise suited for the inference of more complex and not directly documented traits, such as mechanisms of interaction and subcellular (or extracellular) localisation. For example, it could have been burdensome and highly speculative to infer the so far uniquely observed extracellular condition in *Deianiraea*⁸ only from its genome. Therefore, considering that we still completely lack any experimental data for six out of ten *Rickettsiales* families (including all basal ones), we strongly invoke the need for further experimental investigations. As in other cases⁷⁵, these may provide additional and otherwise unpredictable insights on the lifestyle of present-day organisms, and represent the basis for refining existing hypotheses and inferring novel ones on *Rickettsiales* evolution.

Methods

Sample preparation and sequencing

In this work, the nine novel *Rickettsiales* genome sequences were assembled (for a detailed account on sample preparation, sequencing, and genome assembly, see Supplementary Note 3; Supplementary Data 1, 5), starting from eight protist host samples. Six of these samples were characterised in previous studies^{47–49,76–78}, while two others, namely the ciliates *Plagiopyla frontata* IBS-3 and *Euplotes woodruffi* NDG2, were newly isolated. Each sample was differentially processed. Briefly, for Illumina sequencing most samples were subjected to whole-genome amplification (WGA) with the REPLI-g Single Cell Kit (Qiagen), either directly from few ciliate host cells (four samples: *Paramecium biaurelia* US_BI 11111, *Paramecium nephridiatum* Sr 2-6, *E. woodruffi* NDG2, *P. frontata* IBS-3) or from a previously obtained DNA extract (one sample: *Euplotes harpa* BOD18;⁷⁶ Supplementary Data 1). Two additional samples (*P. biaurelia* USBL-3611 and *Paramecium multimicronucleatum* Kr154-4) were processed for bulk DNA extraction (over 200,000 ciliate host cells each), with CTAB and phenol-chloroform protocols, respectively. Each of these seven DNA samples was processed through a Nextera XT library and sequenced by Admera Health (South Plainfield, NJ, USA) on a Illumina HiSeq X machine, producing 2 × 150 bp paired-end reads. Read quality was assessed with FastQC 0.11.7⁷⁹. NDG2 sample was also subjected to Nanopore sequencing. For this purpose, a bulk (~300,000 *Euplotes* cells) extract with the NucleoSpin™ Tissue Kit (Macherey-Nagel™) was processed through a SQK-LSK109 ligation-sequencing library, and sequenced in a FLO-MINI06 flow cell. Basecalling was then performed with guppy 5.0.11. Then, reads were processed with Porechop 0.2.4⁸⁰ with default options. Quality of the reads was assessed with NanoPlot 1.23.0⁸¹. The eighth sample consisted in a *Rickettsiales* bacterium associated with the foraminiferan *Reticulomyxa filosa*, already sequenced together with its host in a previous study⁷⁸. Sequencing reads were kindly provided by the original authors.

Genome assembly

For each sample, the total Illumina reads were assembled using SPAdes 3.6⁸² with default settings, obtaining a “preliminary assembly”. Then, a

multi-step procedure was applied, in order to select only the contigs belonging to the symbiont of interest and discard those belonging to the host or to additional organisms present in the sample (e.g., residual food, additional associated bacteria), as described previously (e.g.,⁸). For this purpose, the blobology pipeline was applied⁸³, followed by extensive manual curation. Briefly, preliminary contigs were classified according to their length, GC% content, sequencing coverage, and taxonomy. Reads mapping by Bowtie2 2.4.2⁸⁴ on selected contigs were reassembled separately with SPAdes 3.6, or, for NDG2, with Unicycler⁸⁵ in a hybrid assembly with the respective Nanopore reads. Two samples (*P. multimicronucleatum* 12 and *P. biaurelia* US_BI 11111) were subjected to genome finishing, performing PCR reactions with TaKaRa Ex Taq and reagents (Takara Bio, Japan). Successful results were confirmed by bidirectional Sanger sequencing performed by GATC Biotech (Germany). Quality of the novel assemblies was confirmed by their completeness scores on 219 proteobacterial orthologs predicted with BUSCO 5.0.0⁸⁶, as compared with the scores of published *Rickettsiales* (Supplementary Data 2).

Annotation

The newly obtained genomes were all annotated with Prokka 1.10⁸⁷, using the `--rfam` option. Afterwards, annotation of the genomes of ciliate symbionts was manually curated by a detailed inspection of blastp hits on NCBI nr and on *Rickettsiales* proteins as described previously⁸.

Full *Rickettsiales* dataset construction and phylogenomic analyses

Phylogenomic analyses were aimed to collect a representative and comprehensive view on the evolution and diversity of *Rickettsiales*. All sequences were downloaded from NCBI GenBank via ftp (<ftp.ncbi.nlm.nih.gov/genomes/all/GCA>), and are updated to July 2021. We manually selected a representative set of 36 *Rickettsiales* genomes, including at least one representative per genus. For the phylogeny, other 89 representative non-*Rickettsiales* *Alphaproteobacteria*, as well as 8 *Gammaproteobacteria* and *Betaproteobacteria* as outgroup were employed, taking inspiration from the selection by Muñoz-Gómez and co-authors¹. We then aimed to identify all MAGs (metagenome-assembled genomes) which could be assigned to known “core” *Rickettsiales* lineages as of July 2021 (i.e., the four families *Rickettsiaceae*, *Anaplasmataceae*, *Midichloriaceae*, *Deianiraeaceae*), or to any lineage forming a supported monophyletic group with *Rickettsiales* with the exclusion of other (alpha)proteobacterial orders. Identification and representative selection of *Rickettsiales* MAGs was performed by a multi-step procedure (detailed in Supplementary Note 4). Briefly, all MAGs assigned to *Rickettsiales* by NCBI taxonomy, those assigned to deep-branching alphaproteobacterial lineages²⁵, plus all additional monophyletic relatives from the gtdb tree⁸⁸, were downloaded (394 total MAGs). MAGs were first filtered by assembly quality, retaining only 314 MAGs having ≥50% single-copy and <5% duplicated of 219 proteobacterial orthologs according to BUSCO 5.0.0 (Supplementary Data 2).

All phylogenomic analyses were performed on concatenated alignments of 179 orthogroups (Supplementary Data 6), which were manually selected on purpose (i.e., presence, after manual polishing of paralogs and poorly aligned sequences, in at least 85% of *Rickettsiales* -MAGs excluded-, 85% *Alphaproteobacteria*, 50% outgroup) from the eggNOG v5 assignments⁸⁹ predicted with eggNOG-mapper 2.0.6⁹⁰. For each organismal dataset (see below), orthogroups were separately aligned with MAFFT 7.475⁹¹, polished with BMGE 1.12⁹² and concatenated with AMAS⁹³. All phylogenies were performed with IQ-TREE 1.6.12⁹⁴, with 1000 ultrafast bootstraps⁹⁵ and 1000 SH-aLRT replicates, employing the LG + C60 + F + R6 model unless specified. A first phylogeny was performed on the full organismal dataset for an initial classification of MAGs, employing ModelFinder^{96,97} for model

selection. In order to avoid artefacts due to compositional heterogeneity in the dataset (in particular potential “erroneous” phylogenetic proximity of MAG lineages to “core” *Rickettsiales* due GC/AT biases)^{1,25,26}, the approach by Muñoz-Gómez and co-authors¹ was applied, thus removing 10%, 20%, 30%, 40% or 50% of most heterogeneous sites from the alignment, and performing a separate phylogeny on each trimmed alignment (Supplementary Fig. 29). Based on resulting monophyly and Average Amino acid Identity (AAI) > 0.85, phylogenetically-redundant MAGs were discarded (Supplementary Fig. 30). Thirteen clades were identified, grouping all those MAGs which could not be directly assigned to “core” *Rickettsiales* or other orders. In order to minimise potential artefacts (e.g., due to long-branch attraction), the phylogenetic position of the MAGs belonging to each clade was tested separately, with respect to core *Rickettsiales* lineages and other *Alphaproteobacteria* (Supplementary Fig. 31, 32). Phylogenies were performed accounting for compositional heterogeneity as above, and only MAGs in a monophyletic relationship with *Rickettsiales* were retained. Therefore, 68 total *Rickettsiales* MAGs were selected for all the successive analyses, totalling 210 organisms in the final dataset (113 *Rickettsiales*). For the sake of phylogenetic representativeness, this final dataset included sequences from the AT-rich *Holosporales* and *Pelagibacterales* genomes, and lacked, also due to computational limits, sequences from non-*Rickettsiales* MAG-only alphaproteobacterial lineages (e.g., the MarineAlpha by²⁵). Accordingly, the reconstructions of the evolutionary history of *Rickettsiales* genes (see below) could have been potentially affected. Nevertheless, this was taken into account in the interpretation of the results, and potential artefacts due to the AT-rich sequences were directly addressed (see below).

For the final dataset, we applied the same approach as above (IQ-TREE phylogeny accounting for compositional heterogeneity; Supplementary Fig. 18).

Creation of a set of “homology groups” for gene content comparisons

In order to perform gene content comparisons and infer variations along the inferred species tree, we aimed to obtain a set of broad “homology groups” for the 210 total organisms in the final phylogeny rather than analyse directly the “raw” orthogroups from eggNOG. The aim of this step was to get a comprehensive overview of the homologous genes in the dataset, namely regardless of whether the common ancestor of all sequences within the same “homology group” was younger or older than the ancestor of the investigated organisms. This allowed us a comprehensive inspection of homologues, providing the basis for de novo inference/analysis on the possible duplication and/or horizontal transfer events specific to the *Rickettsiales*, including from distantly related organisms falling outside the taxonomic range herein directly investigated. This was realised by merging the orthogroups resulting from eggNOG assignments that, based on eggNOG itself, shared significant homology (see Supplementary Note 5 for details). Briefly, the eggNOG database is hierarchically organised by taxonomy, namely each lineage (at domain, phylum, class, order, family levels) has a dedicated set of orthogroups, linked to those of higher-level taxa. We aimed to maximise the advantages of such a system (chiefly lineage-specific annotations), and at the same time minimise disadvantages for the intended aim. These disadvantages may include incomplete grouping of homologues (including potential orthologs) due to assignment to eggNOGs belonging to distinct taxa⁸. To overcome such limitations, we designed a “telescopic” approach, in order to merge genes into larger meaningful “homology groups”, while still keeping as much as possible lineage-specific refined annotations. Briefly (see Supplementary Note 5 for details), we compared taxonomic paths of all the identified eggNOGs, and grouped into a single “homology group” all the eggNOGs sharing at least a partial taxonomic path, and labelling each “homology group” with the eggNOG at lowest possible

shared rank in the taxonomic path leading to *Rickettsiales* (root; *Bacteria*; *Proteobacteria*; *Alphaproteobacteria*; *Rickettsiales*). Applying such a “telescopic” approach, a total of 444,226 genes were assigned to 20,041 “homology groups”, 4009 of which present in at least one member of *Rickettsiales*, and 2990 of those present in 4 or more organisms of the total dataset, and thus considered for the following analyses. The presence/absence and copy number patterns of each homology group were carefully manually inspected, in order to get a general overview of the functional repertoire of the investigated *Rickettsiales*, and to instruct more specific in-depth analyses (see below).

Phylogenetic analyses on biosynthetic pathways for amino acids and nucleotides

Reference biosynthetic pathways for amino acids and nucleotides were obtained from the Biocyc database⁵⁵ (Supplementary Data 7). The datasets for the respective phylogenies were extensively manually curated (see Supplementary Note 5 for details). Briefly, for each gene the corresponding eggNOG “homology group” was identified by a blastp search, and its composition was refined (e.g., excluding clear paralogs, also with the help of NCBI conserved domain search⁹⁸, and poorly aligned sequences) by inspection of the respective alignment and of the respective single-gene tree. Preliminary and final single-gene trees were obtained with IQ-TREE and ModelFinder, after aligning and trimming as described above (“Full *Rickettsiales* dataset construction and phylogenomic analyses”)⁹⁴. While inspecting the phylogenies, we focused on verifying the support for the monophyly of the whole *Rickettsiales* and of each *Rickettsiales* family, or alternatively on reconstructions suggesting possible HGT events with *Rickettsiales* as recipients. On the other hand, potential events with *Rickettsiales* as donors, such as sequences of other lineages nested within a clade of *Rickettsiales* sequences, were not highlighted, being non-target for the aims of this study.

Based on inspection, whenever appropriate (i.e., all cases except cysteine synthesis, see results), we removed suspected non-orthologs genes, and, in order to get more phylogenetically informative datasets for more robust and reliable inferences, we concatenated together⁹³ the sequences involved in the same pathway, as well as in pathways sharing common reactions. Each concatenated alignment was also processed as described above (“Full *Rickettsiales* dataset construction and phylogenomic analyses”) in order to account for compositional heterogeneity¹. On each resulting alignment, phylogeny was inferred with IQ-TREE and the LG + C60 + F + R6 model, as described above. Such phylogenies were inspected with the same criteria as above for single-genes.

Identification of amino acid transporters

All the proteins of the 113 *Rickettsiales* of our final dataset were employed as queries in a blastp search against the full TCDB database⁶¹. Then, for each *Rickettsiales* genome the number of proteins having a best significant hit (e-value threshold of 1e-5) on each selected entry were counted (see Supplementary Note 5 for details on the selection and refinement of TCDB entries representing putative amino acid transporters and on the rationale for the blastp search).

Identification and phylogenetic analysis of the tlc nucleotide translocases

Analyses were focused on the tlc nucleotide translocase transporters (see Supplementary Note 5 for details), common in *Rickettsiales* and in other host-associated lineages⁵⁹. Briefly, the corresponding eggNOG “homology group” was identified by a blastp search. Then, it was joined together with a selection of the phylogenetic dataset of tlc translocases by Major and co-authors⁵⁹, corresponding to the clade of sequences consisting only of the nucleotide transport protein domain. The sequences were then aligned and trimmed as described above,

and phylogeny was inferred as described above, employing the LG + C60 model as in⁹⁹. Potential HGT events involving *Rickettsiales* were inferred based on the comparison with organismal phylogeny. Namely we inferred potential events for monophyletic groups of genes belonging to (part of) a *Rickettsiales* family with a sister group relationship with non-*Rickettsiales* (or non sister-lineage *Rickettsiales*) sequences. Only events with *Rickettsiales* as recipients were considered, and a range of inferred events is provided, depending on alternative reconstructions involving gene losses.

Identification of genes involved in the interaction with host cells

For getting information on the presence and multiplicity of genes involved in multiple features of the interaction of *Rickettsiales* with host cells, the VFDB core reference database was employed⁹⁹ (see Supplementary Note 5 for details). The VFDB is quite redundant for orthologs identified in different included pathogens. Thus, in order to make it suitable for analyses on our non-model bacteria, for each VFC (Virulence Factor Class) separately, orthologs were identified within the database with OrthoFinder 2.5.4¹⁰⁰, and manually curated. Then, all proteins of our dataset of *Rickettsiales* were employed as queries in a blastp search on the full VFDB core database. For each *Rickettsiales* genome, proteins were counted as “assigned” to each curated orthogroup if displaying a significant (evaluate 1e-5) best blastp hit on any sequence belonging to the orthogroup. Data visualisation was attained with the ComplexHeatmap R package¹⁰¹.

Dataset update

After the full set of analyses had been performed as described above, we aimed to further verify whether any *Rickettsiales* genome recently published in the meantime could provide additional insights on our main conclusions. Thus, we identified and downloaded from NCBI nine novel *Rickettsiales* genomes^{13,19,21,22}, belonging to novel genera or closely related to those herein newly obtained, and performed some general verifications. Specifically, we compared their BUSCO scores (calculated as above), as well as the respective presence of biosynthetic pathways for nucleotides and amino acids and of tlc nucleotide translocases (in terms of predicted eggNOGs, see above). All the results were highly consistent with those obtained with relatives present in the dataset employed for the whole set of analyses of this study (Supplementary Data 2, 8).

Reporting summary

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

Data availability

Sequences obtained in this project were deposited to NCBI under accession number [PRJNA831616](https://doi.org/10.5281/zenodo.10324454). The annotated genome sequences of the symbiont of *Reticulomyxa filosa* is available on Zenodo (<https://doi.org/10.5281/zenodo.10324454>). Accession numbers of published *Rickettsiales* assemblies (including all initial MAGs) analysed in this study are provided herein (Supplementary Data 9). The eggnog (<http://eggnog5.embl.de>), BioCyc (<https://biocyc.org>), TCDB (<https://www.tcdb.org>), and VFDB (<http://www.mgc.ac.cn/VFs>) databases were also employed in this study.

References

- Muñoz-Gómez, S. A. et al. An updated phylogeny of the *Alphaproteobacteria* reveals that the parasitic *Rickettsiales* and *Holospirales* have independent origins. *eLife* **8**, e42535. <https://doi.org/10.7554/eLife.42535> (2019).
- Wang, S. & Luo, H. Dating *Alphaproteobacteria* evolution with eukaryotic fossils. *Nat. Commun.* **12**, 3324 (2021).
- Salje, J. Cells within cells: *Rickettsiales* and the obligate intracellular bacterial lifestyle. *Nat. Rev. Microbiol.* **19**, 375–390 (2021).
- Walker, D. H. & Ismail, N. Emerging and re-emerging rickettsioses: endothelial cell infection and early disease events. *Nat. Rev. Microbiol.* **6**, 375–386 (2008).
- Rikihisa, Y. *Anaplasma phagocytophilum* and *Ehrlichia chaffeensis*: subversive manipulators of host cells. *Nat. Rev. Microbiol.* **8**, 328–339 (2010).
- Renvoisé, A., Merhej, V., Georgiades, K. & Raoult, D. Intracellular *Rickettsiales*: insights into manipulators of eukaryotic cells. *Trends Mol. Med.* **17**, 573–583 (2011).
- Werren, J. H., Baldo, L. & Clark, M. E. *Wolbachia*: master manipulators of invertebrate biology. *Nat. Rev. Microbiol.* **6**, 741–751 (2008).
- Castelli, M. et al. *Deianiraea*, an extracellular bacterium associated with the ciliate *Paramecium*, suggests an alternative scenario for the evolution of *Rickettsiales*. *ISME J.* **13**, 2280–2294 (2019).
- Montagna, M. et al. ‘*Candidatus Midichloriaceae*’ fam. nov. (*Rickettsiales*), an ecologically widespread clade of intracellular alphaproteobacteria. *Appl. Environ. Microbiol.* **79**, 3241–3248 (2013).
- Schön, M.E., Martijn, J., Vosseberg, J., Köstlbacher, S. & Ettema, T.J.G. The evolutionary origin of host association in the *Rickettsiales*. *Nat. Microbiol.* <https://doi.org/10.1038/s41564-022-01169-x> (2022).
- Dumler, J. S., & Walker, D. H. *Rickettsiales*. *Bergey’s Manual of Systematics of Archaea and Bacteria* t, John Wiley & Sons, Inc <https://doi.org/10.1002/9781118960608.obm00074> (2015).
- Carrier, T. J. et al. Microbiome reduction and endosymbiont gain from a switch in sea urchin life history. *Proc. Natl. Acad. Sci. USA* **118**, e2022023118 (2021).
- Davison, H. R. et al. Genomic diversity across the *Rickettsia* and ‘*Candidatus Megaira*’ genera and proposal of genus status for the Torix group. *Nat. Commun.* **13**, 2630 (2022).
- Gruber-Vodicka, H. R. et al. Two intracellular and cell type-specific bacterial symbionts in the placozoan *Trichoplax* H2. *Nat. Microbiol.* **4**, 1465–1474 (2019).
- Yurchenko, T. et al. A gene transfer event suggests a long-term partnership between eustigmatophyte algae and a novel lineage of endosymbiotic bacteria. *ISME J.* **12**, 2163–2175 (2018).
- George, E. E. et al. Highly reduced genomes of protist endosymbionts show evolutionary convergence. *Curr. Biol.* **30**, 925–933.e3 (2020).
- Hess, S. Description of *Hyalodiscus flabellus* sp. nov. (Vampyrellida, Rhizaria) and identification of its bacterial endosymbiont, ‘*Candidatus Megaira polyxenophila*’ (*Rickettsiales*, *Alphaproteobacteria*). *Protist* **168**, 109–133 (2017).
- Castelli, M. et al. ‘*Candidatus Sarmatiella mevalonica*’ endosymbiont of the ciliate *Paramecium* provides insights on evolutionary plasticity among *Rickettsiales*. *Environ. Microbiol.* **23**, 1684–1701 (2021).
- Giannotti, D., Boscaro, V., Husnik, F., Vannini, C. & Keeling, P. J. The ‘Other’ *Rickettsiales*: an Overview of the Family ‘*Candidatus Midichloriaceae*’. *Appl. Environ. Microbiol.* **88**, e0243221 (2022).
- Schulz, F. et al. A *Rickettsiales* symbiont of amoebae with ancient features. *Environ. Microbiol.* **18**, 2326–2342 (2016).
- George, E. E. et al. A single cryptomonad cell harbors a complex community of organelles, bacteria, a phage, and selfish elements. *Curr. Biol.* **33**, 1982–1996.e4 (2023).
- Davison, H. R., Hurst, G. D. D. & Siozios, S. ‘*Candidatus Megaira*’ are diverse symbionts of algae and ciliates with the potential for defensive symbiosis. *Micro. Genom.* **9**, mgen000950 (2023).
- Vannini, C., Petroni, G., Verni, F. & Rosati, G. A bacterium belonging to the *Rickettsiaceae* family inhabits the cytoplasm of the marine ciliate *Diophrys appendiculata* (Ciliophora, Hypotrichia). *Microb. Ecol.* **49**, 434–442 (2005).
- Weinert, L. A., Werren, J. H., Aebi, A., Stone, G. N. & Jiggins, F. M. Evolution and diversity of *Rickettsia* bacteria. *BMC Biol.* **7**, 6 (2009).

25. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
26. Viklund, J., Ettema, T. J. G. & Andersson, S. G. E. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol. Biol. Evolution* **29**, 599–615 (2012).
27. Castelli, M., Sasser, D. & Petroni, G. Biodiversity of ‘non-model’ *Rickettsiales* and their association with aquatic organisms. In: Thomas, S. (ed) *Rickettsiales* pp. 59–91 Springer, Cham https://doi.org/10.1007/978-3-319-46859-4_3 (2016).
28. Xie, F. et al. An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. *Microbiome* **9**, 137 (2021).
29. 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).
30. Driscoll, T. P. et al. Wholly *Rickettsia*! reconstructed metabolic profile of the quintessential bacterial parasite of eukaryotic cells. *MBio* **8**, e00859–17 (2017).
31. Min, C.-K. et al. Genome-based construction of the metabolic pathways of *Orientia tsutsugamushi* and comparative analysis within the *Rickettsiales* order. *Comp. Funct. Genom.* **2008**, 623145 (2008).
32. Green, E. R. & Meccas, J. Bacterial secretion systems: an overview. *Microbiol. Spectr.* **4**, <https://doi.org/10.1128/microbiolspec.VMBF-0012-2015> (2016).
33. Gillespie, J. J. et al. Phylogenomics reveals a diverse *Rickettsiales* type IV secretion system. *Infect. Immun.* **78**, 1809–1823 (2010).
34. Floriano, A. M. et al. The evolution of intramitochondriality in *Midichloria* bacteria. *Environ. Microbiol.* <https://doi.org/10.1111/1462-2920.16446> (2023).
35. Linhartová, I. et al. RTX proteins: a highly diverse family secreted by a common mechanism. *FEMS Microbiol. Rev.* **34**, 1076–1112 (2010).
36. Benz, R. Channel formation by RTX-toxins of pathogenic bacteria: basis of their biological activity. *Biochim. Biophys. Acta* **1858**, 526–537 (2016).
37. Borgo, G. M. et al. A patatin-like phospholipase mediates *Rickettsia parkeri* escape from host membranes. *Nat. Commun.* **13**, 3656 (2022).
38. Niu, H., Kozjak-Pavlovic, V., Rudel, T. & Rikihisa, Y. *Anaplasma phagocytophilum* Ats-1 is imported into host cell mitochondria and interferes with apoptosis induction. *PLoS Pathog.* **6**, e1000774 (2010).
39. Lobato-Márquez, D., Díaz-Orejas, R. & García-Del Portillo, F. Toxin-antitoxins and bacterial virulence. *FEMS Microbiol. Rev.* **40**, 592–609 (2016).
40. Alix, E. et al. The capping domain in RalF regulates effector functions. *PLoS Pathog.* **8**, e1003012 (2012).
41. Nwasike, C., Ewert, S., Jovanovic, S., Haider, S. & Mujtaba, S. SET domain-mediated lysine methylation in lower organisms regulates growth and transcription in hosts. *Ann. N. Y. Acad. Sci.* **1376**, 18–28 (2016).
42. Billington, S. J., Jost, B. H. & Songer, J. G. Thiol-activated cytolytins: structure, function and role in pathogenesis. *FEMS Microbiol. Lett.* **182**, 195–205 (2000).
43. Padmalayam, I., Karem, K., Baumstark, B. & Massung, R. The gene encoding the 17-kDa antigen of *Bartonella henselae* is located within a cluster of genes homologous to the virB virulence operon. *DNA Cell Biol.* **19**, 377–382 (2000).
44. Swart, A. L., Gomez-Valero, L., Buchrieser, C. & Hilbi, H. Evolution and function of bacterial RCC1 repeat effectors. *Cell. Microbiol.* **22**, e13246 (2020).
45. Veyron, S., Peyroche, G. & Cherfils, J. FIC proteins: from bacteria to humans and back again. *Pathog. Dis.* **76**, <https://doi.org/10.1093/femspd/fty012> (2018).
46. Sasser, D. et al. Phylogenomic evidence for the presence of a flagellum and cbb(3) oxidase in the free-living mitochondrial ancestor. *Mol. Biol. Evol.* **28**, 3285–3296 (2011).
47. Boscaro, V. et al. Rediscovering the genus *Lyticum*, multi-flagellated symbionts of the order *Rickettsiales*. *Sci. Rep.* **3**, 3305 (2013).
48. Lanzoni, O. et al. Diversity and environmental distribution of the cosmopolitan endosymbiont ‘*Candidatus* Megaira’. *Sci. Rep.* **9**, 1179 (2019).
49. Mironov, T. & Sabaneyeva, E. A robust symbiotic relationship between the ciliate *Paramecium multimicronucleatum* and the bacterium ‘*Ca. Trichorickettsia mobilis*’. *Front. Microbiol.* **11**, 603335 (2020).
50. Keegstra, J. M., Carrara, F. & Stocker, R. The ecological roles of bacterial chemotaxis. *Nat. Rev. Microbiol.* **20**, 491–504 (2022).
51. Guérin, J., Bigot, S., Schneider, R., Buchanan, S. K. & Jacob-Dubuisson, F. Two-partner secretion: combining efficiency and simplicity in the secretion of large proteins for bacteria-host and bacteria-bacteria interactions. *Front. Cell. Infect. Microbiol.* **7**, 148 (2017).
52. Sears, K. T. et al. Surface proteome analysis and characterization of surface cell antigen (Sca) or autotransporter family of *Rickettsia typhi*. *PLoS Pathog.* **8**, e1002856 (2012).
53. Seidman, D. et al. *Anaplasma phagocytophilum* surface protein AipA mediates invasion of mammalian host cells. *Cell. Microbiol.* **16**, 1133–1145 (2014).
54. Park, J., Kim, K. J., Grab, D. J. & Dumler, J. S. *Anaplasma phagocytophilum* major surface protein-2 (Msp2) forms multimeric complexes in the bacterial membrane. *FEMS Microbiol. Lett.* **227**, 243–247 (2003).
55. Karp, P. D. et al. The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* **20**, 1085–1093 (2019).
56. Audia, J. P. & Winkler, H. H. Study of the five *Rickettsia prowazekii* proteins annotated as ATP/ADP translocases (Tlc): only Tlc1 transports ATP/ADP, while Tlc4 and Tlc5 transport other ribonucleotides. *J. Bacteriol.* **188**, 6261–6268 (2006).
57. Daugherty, R. M. et al. The nucleotide transporter of *Caedibacter caryophilus* exhibits an extended substrate spectrum compared to the analogous ATP/ADP translocase of *Rickettsia prowazekii*. *J. Bacteriol.* **186**, 3262–3265 (2004).
58. Schmitz-Esser, S. et al. ATP/ADP translocases: a common feature of obligate intracellular amoebal symbionts related to *Chlamydiae* and *Rickettsiales*. *J. Bacteriol.* **186**, 683–691 (2004).
59. Major, P., Embley, T. M. & Williams, T. A. Phylogenetic diversity of NTT nucleotide transport proteins in free-living and parasitic bacteria and eukaryotes. *Genome Biol. Evol.* **9**, 480–487 (2017).
60. Weyandt, N., Aghdam, S. A. & Brown, A. M. V. Discovery of early-branching *Wolbachia* reveals functional enrichment on horizontally transferred genes. *Front. Microbiol.* **13**, 867392 (2022).
61. Saier, M. H. et al. The Transporter Classification Database (TCDB): 2021 update. *Nucleic Acids Res.* **49**, D461–D467 (2021).
62. Burkovski, A. & Krämer, R. Bacterial amino acid transport proteins: occurrence, functions, and significance for biotechnological applications. *Appl. Microbiol. Biotechnol.* **58**, 265–274 (2002).
63. Dean, P., Hirt, R. P. & Embley, T. M. Microsporidia: why make nucleotides if you can steal them? *PLoS Pathog.* **12**, e1005870 (2016).
64. Carter, N. S., Yates, P., Arendt, C. S., Boitz, J. M. & Ullman, B. Purine and pyrimidine metabolism in *Leishmania*. *Adv. Exp. Med. Biol.* **625**, 141–154 (2008).
65. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2011).

66. Moran, N. A. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**, 583–586 (2002).
67. Gillespie, J. J. et al. Secretome of obligate intracellular *Rickettsia*. *FEMS Microbiol. Rev.* **39**, 47–80 (2015).
68. Mattick, J. S. Type IV pili and twitching motility. *Annu. Rev. Microbiol.* **56**, 289–314 (2002).
69. Abby, S. S. & Rocha, E. P. C. The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems. *PLoS Genet.* **8**, e1002983 (2012).
70. Kahlon, A. et al. *Anaplasma phagocytophilum* Asp14 is an invasin that interacts with mammalian host cells via its C terminus to facilitate infection. *Infect. Immun.* **81**, 65–79 (2013).
71. Husnik, F. et al. Bacterial and archaeal symbioses with protists. *Curr. Biol.* **31**, R862–R877 (2021).
72. Dharamshi, J. E. et al. Marine sediments illuminate *Chlamydiae* diversity and evolution. *Curr. Biol.* **30**, 1032–1048.e7 (2020).
73. Hugoson, E., Guliaev, A., Ammunét, T. & Guy, L. Host adaptation in *Legionellales* is 1.9 Ga, coincident with eukaryogenesis. *Mol. Biol. Evol.* **39**, msac037 (2022).
74. Boscaro, V. et al. Microbiomes of microscopic marine invertebrates do not reveal signatures of phyllosymbiosis. *Nat. Microbiol.* **7**, 810–819 (2022).
75. Imachi, H. et al. Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).
76. Vannini, C. et al. ‘*Candidatus Anadelfobacter veles*’ and ‘*Candidatus Cyrtobacter comes*’, two new rickettsiales species hosted by the protist ciliate *Euplotes harpa* (Ciliophora, Spirotrichea). *Appl. Environ. Microbiol.* **76**, 4047–4054 (2010).
77. Szokoli, F. et al. Disentangling the taxonomy of *Rickettsiales* and description of two novel symbionts (‘*Candidatus Bealeia paramacronuclearis*’ and ‘*Candidatus Fokinia cryptica*’) sharing the cytoplasm of the ciliate protist *Paramecium biaurelia*. *Appl. Environ. Microbiol.* **82**, 7236–7247 (2016).
78. Glöckner, G. et al. The genome of the foraminiferan *Reticulomyxa filosa*. *Curr. Biol.* **24**, 11–18 (2014).
79. Andrews, S. FastQC: A quality control tool for high throughput sequence data. [<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>]. (2010).
80. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Micro. Genom.* **3**, e000132 (2017).
81. De Coster, W., D’Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
82. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
83. Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. & Blaxter, M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet.* **4**, 237 (2013).
84. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
85. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).
86. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
87. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
88. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
89. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
90. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
91. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
92. 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).
93. Borowiec, M. L. AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ* **4**, e1660 (2016).
94. 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).
95. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
96. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermini, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
97. Fan, L. et al. Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within *Alphaproteobacteria*. *Nat. Ecol. Evol.* **4**, 1213–1219 (2020).
98. Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* **32**, W327–W331 (2004).
99. Liu, B., Zheng, D., Zhou, S., Chen, L. & Yang, J. VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.* **50**, D912–D917 (2022).
100. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
101. Gu, Z. Complex heatmap visualization. *Imeta* **1**, e43 (2022).

Acknowledgements

This project was supported by the Human Frontier Science Program (HFSP) Young Investigator Program grant RGY-0075 to D.S., by the Italian Ministry of Education, University and Research (MIUR): Dipartimenti di Eccellenza Programme (2018–2022) Department of Biology and Biotechnology ‘L. Spallanzani’ University of Pavia to D.S., by the European Community’s H2020 Programme H2020-MSCA-RISE 2019 under grant agreement n° 872767 to G.P., and by EU funding within the NextGenerationEU-MUR PNRR Extended Partnership initiative on Emerging Infectious Diseases (Project no. PE00000007, INF-ACT) to M.C. and D.S. Genomic characterisation of *Trichorickettsia* and *Megaira venefica* was performed partly with support of RSF 20-14-00220 to A.P. The University of Pisa is acknowledged for providing visiting scholarships to E.S. and A.P. The authors would like to thank Venkata Mahesh Nitla for support in culturing *Plagiopyla frontata* IBS-3, Sascha Krenek for DNA preparation of the *Paramecium biaurelia* strain US_BI 11111, Umberto Postiglione for assistance in Nanopore assembly, Laura Quattrini and Marco Fagioli for assistance in genome closing by PCR. Gernot Gloeckner and Marco Groth are gratefully acknowledged for sharing genome sequencing data on *Reticulomyxa filosa* and its associated *Rickettsiales* bacterium.

Author contributions

M.C., G.P and D.S. conceived and designed the study. M.C., E.S., A.P. and V.S carried out the experimental part, which was supervised by G.P. M.C. and T.N. performed data analysis, which was contributed by L.G. and G.B. and supervised by D.S. T.N. and G.B. curated data presentation. M.C and D.S performed conceptualisation and wrote the paper, which were contributed by all authors.

Competing interests

The authors declare no competing interests

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-45351-7>.

Correspondence and requests for materials should be addressed to Giulio Petroni or Davide Sassera.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024