







# Genomic diversity across the *Rickettsia* and 'Candidatus Megaira' genera and proposal of genus status for the Torix group

Helen R. Davison <sup>1</sup>, Jack Pilgrim<sup>1</sup>, Nicky Wybouw <sup>2</sup>, Joseph Parker<sup>3</sup>, Stacy Pirro<sup>4</sup>, Simon Hunter-Barnett <sup>1</sup>, Paul M. Campbell <sup>1,5</sup>, Frances Blow <sup>1,6</sup>, Alistair C. Darby<sup>1</sup>, Gregory D. D. Hurst<sup>1</sup> & Stefanos Siozios <sup>1</sup>✉

Members of the bacterial genus *Rickettsia* were originally identified as causative agents of vector-borne diseases in mammals. However, many *Rickettsia* species are arthropod symbionts and close relatives of 'Candidatus Megaira', which are symbiotic associates of microeukaryotes. Here, we clarify the evolutionary relationships between these organisms by assembling 26 genomes of *Rickettsia* species from understudied groups, including the Torix group, and two genomes of 'Ca. Megaira' from various insects and microeukaryotes. Our analyses of the new genomes, in comparison with previously described ones, indicate that the accessory genome diversity and broad host range of Torix *Rickettsia* are comparable to those of all other *Rickettsia* combined. Therefore, the Torix clade may play unrecognized roles in invertebrate biology and physiology. We argue this clade should be given its own genus status, for which we propose the name 'Candidatus Tisiphia'.

<sup>1</sup>Institute of Infection, Veterinary and Ecological sciences, University of Liverpool, Liverpool L69 7ZB, UK. <sup>2</sup>Terrestrial Ecology Unit, Department of Biology, Faculty of Sciences, Ghent University, Ghent, Belgium. <sup>3</sup>Division of Biology and Biological Engineering, California Institute of Technology, 1200 E California Boulevard, Pasadena, CA 91125, USA. <sup>4</sup>Iridian Genomes, Bethesda, MD, USA. <sup>5</sup>School of Health and Life Sciences, Faculty of Biology Medicine and Health, the University of Manchester, Manchester, UK. <sup>6</sup>Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY, USA. ✉email: [siozios@liverpool.ac.uk](mailto:siozios@liverpool.ac.uk)

Symbiotic bacteria are vital to the function of most living eukaryotes, including microeukaryotes, fungi, plants, and animals<sup>1–4</sup>. The symbioses formed are often functionally important to the host with effects ranging from mutualistic to detrimental. Mutualistic symbionts may provide benefits through the biosynthesis of metabolites, or by protecting their hosts against pathogens and parasitoids<sup>5,6</sup>. Parasitic symbionts can be detrimental to the host due to resource exploitation or through reproductive manipulation that favours its own transmission over the host's<sup>7,8</sup>. Across these different symbiotic relationships, symbionts are often important determinants of host ecology and evolution.

The *Rickettsiales* (Alphaproteobacteria) represent an order of largely obligate intracellular bacteria that form symbioses with a variety of eukaryotes<sup>9</sup>. *Deianiraea*, an extracellular parasite of *Paramecium*, is the one known exception<sup>10</sup>. Within *Rickettsiales*, the family *Rickettsiaceae* represent a diverse collection of bacteria that infect a wide range of eukaryotic hosts and can act as symbionts, parasites, and pathogens. Perhaps the best-known clade of *Rickettsiaceae* is the genus *Rickettsia*, which was initially described as the cause of spotted fever and other rickettsioses in vertebrates that are transmitted by ticks, lice, fleas, and mites<sup>11</sup>.

*Rickettsia* have been increasingly recognised as heritable arthropod symbionts. Since the description of a maternally inherited male-killer in ladybirds<sup>12</sup>, we now know that heritable *Rickettsia* are common in arthropods<sup>13,14</sup>. Further, *Rickettsia*-host symbioses are diverse, with different symbionts being capable of reproductive manipulation, nutritional and protective symbiosis, as well as influencing thermotolerance and pesticide susceptibility<sup>15–21</sup>.

Our understanding of the evolution and diversity of the genus *Rickettsia* and its allies has increased in recent years, with the taxonomy of *Rickettsiaceae* developing as more data becomes available<sup>14,22</sup>. Weinert et al.<sup>14</sup> loosely defined 13 different groups of *Rickettsia* based on 16 S rRNA phylogeny, which showed two early branching clades that appeared genetically distant from other members of the genus. One of these was a symbiont of *Hydra* and designated as Hydra group *Rickettsia*, which has since been assigned its own genus status, '*Candidatus Megaira*'<sup>23</sup>. '*Ca. Megaira*' forms a related clade to *Rickettsia* and is found in ciliates, amoebae, chlorophyte and streptophyte algae, and cnidarians<sup>24</sup>. Members of this clade are found in hosts from aquatic, marine and soil habitats which include model organisms (e.g., *Paramecium*, *Volvox*) and economically important vertebrate parasites (e.g., *Ichthyophthirius multifiliis*, the ciliate that causes white spot disease in fish)<sup>24</sup>. Whilst symbioses between '*Ca. Megaira*' and microeukaryotes are pervasive, there is no publicly available complete genome and the impact of these symbioses on the host are poorly understood.

A second early branching clade was described from *Torix tagoi* leeches and is commonly coined Torix group *Rickettsia*<sup>25</sup>. Symbionts in the Torix clade have since been found in a wide range of invertebrate hosts from midges to freshwater snails to fish-parasitic amoeba<sup>13</sup>. The documented diversity of hosts is wider than other *Rickettsia* groups, which are to date only found in arthropods and their associated vertebrate or plant hosts<sup>14</sup>. Torix clade *Rickettsia* are known to be heritable symbionts, but their impact on host biology is poorly understood, despite the economic and medical importance of several hosts (inc. bed bugs, black flies, and biting midges). Rare studies have described the potential effects on the host, which include larger body size in leeches<sup>25</sup>; a small negative effect on growth rate and reproduction in bed bugs<sup>26</sup>; and an association with parthenogenesis in *Empoasca* leafhoppers<sup>27</sup>.

Current data suggest an emerging macroevolutionary scenario where the members of the *Rickettsia* clade originated as

symbionts of microeukaryotes, before diversifying to infect invertebrates<sup>23,28,29</sup>. Many symbionts belonging to the *Rickettsiaceae* (e.g., '*Ca. Megaira*', '*Candidatus Trichorickettsia*', '*Candidatus Phycorickettsia*', '*Candidatus Sarmatiella*' and '*Candidatus Gigarickettsia*') circulate in a variety of microeukaryotes<sup>23,30–33</sup>. The Torix group *Rickettsia* retained a broad range of hosts from microeukaryotes to arthropods<sup>13</sup>. The remaining members of the genus *Rickettsia* evolved to be arthropod heritable symbionts and vector-borne pathogens<sup>14,34</sup>. However, a lack of genomic and functional information for symbiotic clades limits our understanding of evolutionary transitions within *Rickettsia* and its related groups. No '*Ca. Megaira*' genome sequences are currently publicly available and of the 165 *Rickettsia* genome assemblies available on the NCBI (as of 29/04/21), only two derive from the Torix clade and these are both draft genomes. In addition, dedicated heritable symbiont clades of *Rickettsia*, such as the Rhyzobius group, have no available genomic data, and there is a single representative for the *Adalia* clade. Despite the likelihood that heritable symbiosis with microeukaryotes and invertebrates was the ancestral state for this group of intracellular bacteria, available genomic resources are heavily skewed towards pathogens of vertebrates.

In this study we establish a richer base of genomic information for heritable symbionts *Rickettsia* and '*Ca. Megaira*', then use these resources to clarify the evolution of these groups. We broaden available genomic data through a combination of targeted sequencing of strains without complete genomes, and metagenomic assembly of *Rickettsia* strains from arthropod genome projects. We report the first closed circular genome of a '*Ca. Megaira*' symbiont from a streptophyte alga (*Mesostigma viride*) and provide a draft genome for a second '*Ca. Megaira*' from a chlorophyte (*Carteria cerasiformis*). In addition, we present the complete genomes of two Torix *Rickettsia* from a midge (*Culicoides impunctatus*) and a bed bug (*Cimex lectularius*) as well as a draft genome for *Rickettsia* from a tsetse fly (*Glossina morsitans submorsitans*, an important vector species), and a new strain from a spider mite (*Bryobia graminum*). A metagenomic approach established a further 22 draft genomes for insect symbiotic strains, including previously unsequenced Rhyzobius and Meloidae group draft genomes. We utilize these to conduct pangenomic, phylogenomic, and metabolic analyses of our extracted genome assemblies, with comparisons to existing *Rickettsia*.

## Results and discussion

We have expanded the available genomic data for several *Rickettsia* groups through a combination of draft and complete genome assembly. This includes an eight-fold increase in available Torix-group genomes, and genomes for previously unsequenced Meloidae and Rhyzobius groups. We further report initial reference genomes for '*Ca. Megaira*'.

**Complete and closed reference genomes for Torix *Rickettsia* and '*Ca. Megaira*'.** The use of long-read sequencing technologies produced complete genomes for two subclades of the Torix group limoniae (RiCimp) and leech (RiClec). Sequencing depth of the *Rickettsia* genomes from *C. impunctatus* (RiCimp) and *C. lectularius* (RiClec) were 18X and 52X, respectively. The RiCimp genome provides evidence of plasmids in the Torix group (pRiCimp001 and pRiCimp002) (Table 1). Notably, the two plasmids share more similarities between them than to other *Rickettsia* plasmids. However, both plasmids contain distant homologs of the DnaA\_N domain-containing proteins previously found in other *Rickettsia* plasmids<sup>35</sup>. In addition, only two components of the type IV conjugative transfer system known as RAGEs

**Table 1 Summary of the closed ‘Ca. Megaira’ and Torix *Rickettsia* genomes completed in this project.**

Group	‘Ca. Megaira’	Torix <i>Rickettsia</i>	Torix <i>Rickettsia</i>
Strain Name	MegNEIS296	RiCimp	RiClec
Symbiont genome accession	GCA_020410825.1	GCA_020410785.1	GCA_020410805.1
Host	Mesostigma viride NIES-296	<i>Culicoides impunctatus</i>	Cimex lectularius
Raw reads accession	SRR8439255, SRX5120346	SRR16018514, SRR16018513	SRR16018512, SRR16018511
Total nucleotides	1,532,409	1,566,468	1,611,726
Chromosome size (bp)	1,448,425	1,469,631	1, 611,726
Plasmids	1 (83,984 bp)	2 (77550 bp + 19287 bp)	None
GC content (%)	33.9	32.9	32.8
Number of CDS	1,359	1,397	1,544
Avg. CDS length (bp)	998	900	874
Coding density (%)	88.5	86	84
rRNAs	3	3	3
tRNAs	34	34	35

(*Rickettsiales* Amplified Genetic Elements)<sup>36</sup> were present on the plasmids including homologs of the proteins TrwB/TraD and TraA/MobA. The majority of the RAGE elements including both the F-like (*tra*) and P-like type IV components have been incorporated in the main chromosome. The presence of RAGE elements, alongside the fact conjugation apparatuses have narrow host-ranges<sup>37</sup>, suggest horizontal transfer of these plasmids is likely within the *Rickettsiaceae* and could occur between Torix and the main *Rickettsia* clade, considering co-infections of these genera have been noted previously<sup>38,39</sup>. We additionally assembled a complete closed reference genome of ‘Ca. Megaira’ from *Mesostigma viride* (MegNEIS296) from previously published genome sequencing efforts. Likewise, MegNEIS296 genome contains a plasmid which bears features of other *Rickettsia* plasmids including the presence of a *tra* conjugative element and the presence of two DnaA\_N-like protein paralogs.

General features of both genomes are consistent with previous genomic studies of the Torix group (Table 1). A single full set of rRNAs (16S, 5S and 23S) and a GC content of ~33% was observed. Notably, the two complete Torix group genomes show a distinct lack of synteny (Supplementary Fig. 1), a genomic feature that is compatible with our phylogenetic analyses that placed these two lineages in different subclades (leech/limoniae) (Fig. 1 and Supplementary Fig. 3). Gene order breakdown due to intragenomic recombination has been previously associated with the expansion of mobile genetic elements in both *Rickettsia*<sup>40</sup> and *Wolbachia*<sup>41</sup>, another member of the *Rickettsiales*. Both RiCimp and RiClec genomes predicted to encode for a high number of transposable elements with circa 96 and 119 annotated putative transposases, respectively. This expansion of transposable elements along with their phylogenetic distance is likely responsible for the extreme synteny breakdown between RiCimp and RiClec. Of note within the closed reference genomes MegNEIS296 and RiCimp is the presence of a putative non-ribosomal peptide synthetase (NRPS) and a hybrid non-ribosomal peptide/polyketide synthetase (NRPS/PKS) respectively (Supplementary Fig. 2). Although, the exact products of these putative pathways are uncertain, in silico prediction by Norine suggests some similarity with both cytotoxic and antimicrobial peptides hinting at a potential defensive role (Supplementary Fig. 2). Further homology comparison with other taxa did not provide links with any specific functions or phenotypes. Previously, an unrelated hybrid NRPS/PKS cluster has been reported in *Rickettsia buchneri* on a mobile genetic element, providing potential routes for horizontal transmission<sup>42</sup>. The strongest blastp hits of MegNEIS296 NRPS proteins occur in *Cyanobacteria* (Supplementary Fig. 2)<sup>42</sup>. In addition, putative toxin-antitoxin systems similar to one associated with cytoplasmic incompatibility in *Wolbachia* have recently been

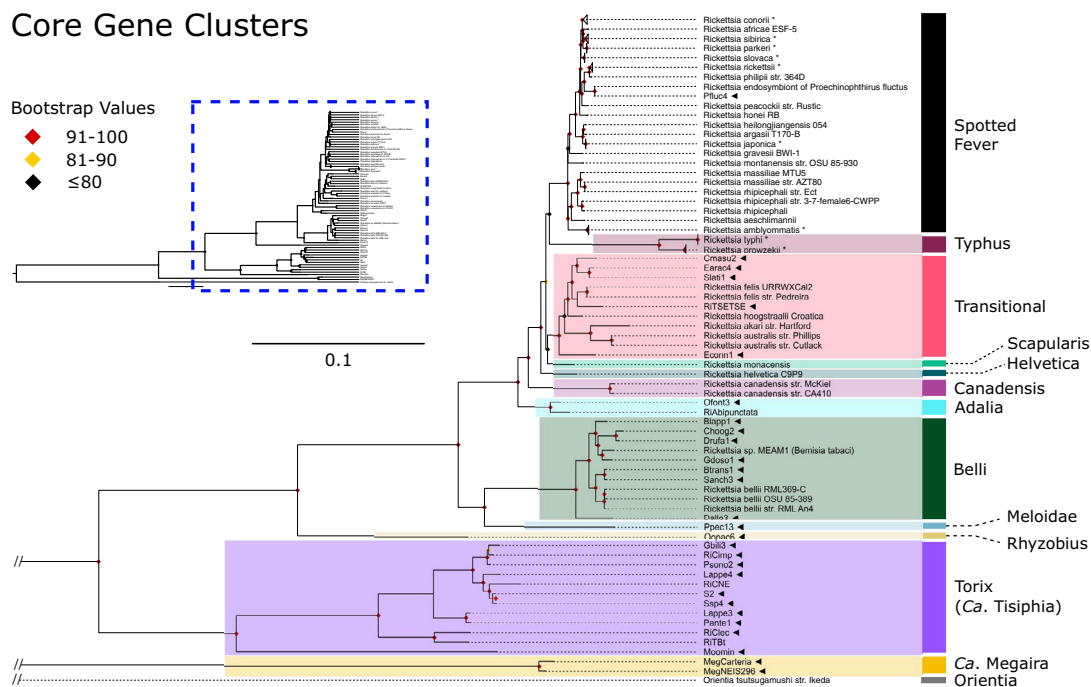
observed on the plasmid of *Rickettsia felis* in a parthenogenetic booklouse<sup>35</sup>. Toxin-antitoxin systems are thought to be part of an extensive bacterial mobilome network associated with reproductive parasitism<sup>43</sup>. A BLAST search found a very similar protein in Oopac6 to the putative large pLbAR toxin found in *R. felis* (88% aa identity), and a more distantly related protein in the *C. impunctatus* plasmid (25% aa identity).

**Sequencing and de novo assembly of other *Rickettsia* and ‘Ca. Megaira’ genomes.** Our direct sequencing efforts enabled assembly of draft genomes for a second ‘Ca. Megaira’ strain from the alga *Carteria cerasiformis*, and for *Rickettsia* associated with tsetse flies and *Bryobia* spider mites. The *Rickettsia* genome retrieved from a wild caught Tsetse fly, RiTSETSE, is a potentially chimeric assembly of closely related Transitional group *Rickettsia*. We identified an excess of 3584 biallelic sites (including 3369 SNPs and 215 indels) when the raw Illumina reads were mapped back to the assembly. High read depth of 104X indicate that this could be a symbiotic association, reflecting previous observations in Tsetse fly cells<sup>44</sup>. However, there is a possibility that RiTSETSE is not a heritable symbiont but comes from transient infection from a recent blood meal.

From the SRA accessions, the metagenomic pipeline extracted 29 full symbiont genomes for *Rickettsiales* across 24 host species. Five of 29 were identified as *Wolbachia* and discarded from further analysis, one was a *Rickettsia* discarded for low quality, and another was a previously assembled Torix *Rickettsia*, RiCNE<sup>45</sup>. Thus, 22 high quality *Rickettsia* metagenomes were obtained from 21 host species. One beetle (SRR6004191) carried coinfecting *Rickettsia* Lappe3 and Lappe4 (Table 2). The high-quality *Rickettsia* genomes covered the Belli, Torix, Transitional, Rhyzobius, Meloidae and Spotted Fever Groups (Table 2 and Supplementary Data 1).

Beetles, particularly rove beetle (*Staphylinidae*) species, appear in this study as a possible hotspot of *Rickettsia* infection. *Rickettsia* has historically been commonly associated with beetles, including ladybird beetles (*Adalia bipunctata*), diving beetles (*Deronectes sp.*) and bark beetles (*Scolytinae*)<sup>14,17,34,46,47</sup>. Though a plausible and likely hotspot, this observation needs to be approached with caution as this could be an artefact of skewed sampling efforts.

**Phylogenomic analyses and taxonomic placement of assembled genomes.** The phylogeny and network illustrate the distance of Torix from ‘Ca. Megaira’ and other *Rickettsia*, along with an extremely high level of within-group diversity in Torix compared



**Fig. 1 Genome wide phylogeny of *Rickettsia* and 'Ca. Megaira'.** Maximum likelihood (ML) phylogeny of *Rickettsia* and 'Ca. Megaira' constructed from 74 core gene clusters extracted from the pangenome. New genomes are indicated by ◀ and bootstrap values based on 1000 replicates are indicated with coloured diamonds (red = 91-100, yellow = 81-90, black <= 80). New complete genomes are: RiCimp, RiClec and MegNEIS296. Asterisks indicate collapsed monophyletic branches and "//" represent breaks in the branch. Accessions used are provided in Supplementary Data 1.

to any other group (Fig. 1 and Supplementary Fig. 3). No significant discordance was detected between the core and ribosomal phylogenies. The phylogenies generated using core genomes are consistent with previously identified *Rickettsia* and host associations using more limited genetic markers<sup>13,14,48,49</sup>. For instance, Pfluc4 from *Proechinophthirus fluctus* lice is grouped on the same branch as a previously sequenced *Rickettsia* from a different individual of *P. fluctus*<sup>48</sup>. The following groups were identified in the 22 genomes assembled from the SRA screening: 4 Transitional, 1 Spotted Fever, 1 Adalia, 8 Belli and 7 Torix limoniae. Targeted sequences were confirmed as: Torix limoniae (RiCimp), Torix leech (RiClec), Transitional (RiTSETSE), 'Ca. Megaira' (MegCarteria and MegNEIS296), and a deeply diverging Torix clade provisionally named Moomin (Moomin) (Table 2, Fig. 1, Supplementary Fig. 3 and 4). The extracted Torix genomes include one double infection giving a total of 10 new genomes across 9 potential host species. The double infection is found within the rove beetle *Labidopullus appendiculatus*, forming two distinct lineages, Lappe3 and Lappe4 (Fig. 1 and Supplementary Fig. 3).

We also report a putative Rhyzobius group *Rickettsia* genomes extracted from the staphylinid beetle *Oxypoda opaca* (Oopac6) and Meloidae group *Rickettsia* from the firefly *Pyrocoelia pectoralis* (Ppec13). They have high completeness, low contamination, and consistently group away from the other draft and completed genomes (Figs. 1, 2, and Supplementary Data 1). MLST analyses demonstrate that these bacteria are most like the Rhyzobius and Meloidae groups described by Weinert et al.<sup>14</sup> (Supplementary Fig. 4). Phylogenies of Oopac6 and Ppec13 suggest that Rhyzobius sits as sister group to all other *Rickettsia* groups, and Meloidae is more closely associated with Belli (Fig. 1, Supplementary Fig. 3–5). Further genome construction will help clarify this taxon and its relationship to the rest of the *Rickettsiaceae*. The sequencing data for the wasp, *Diachasma*

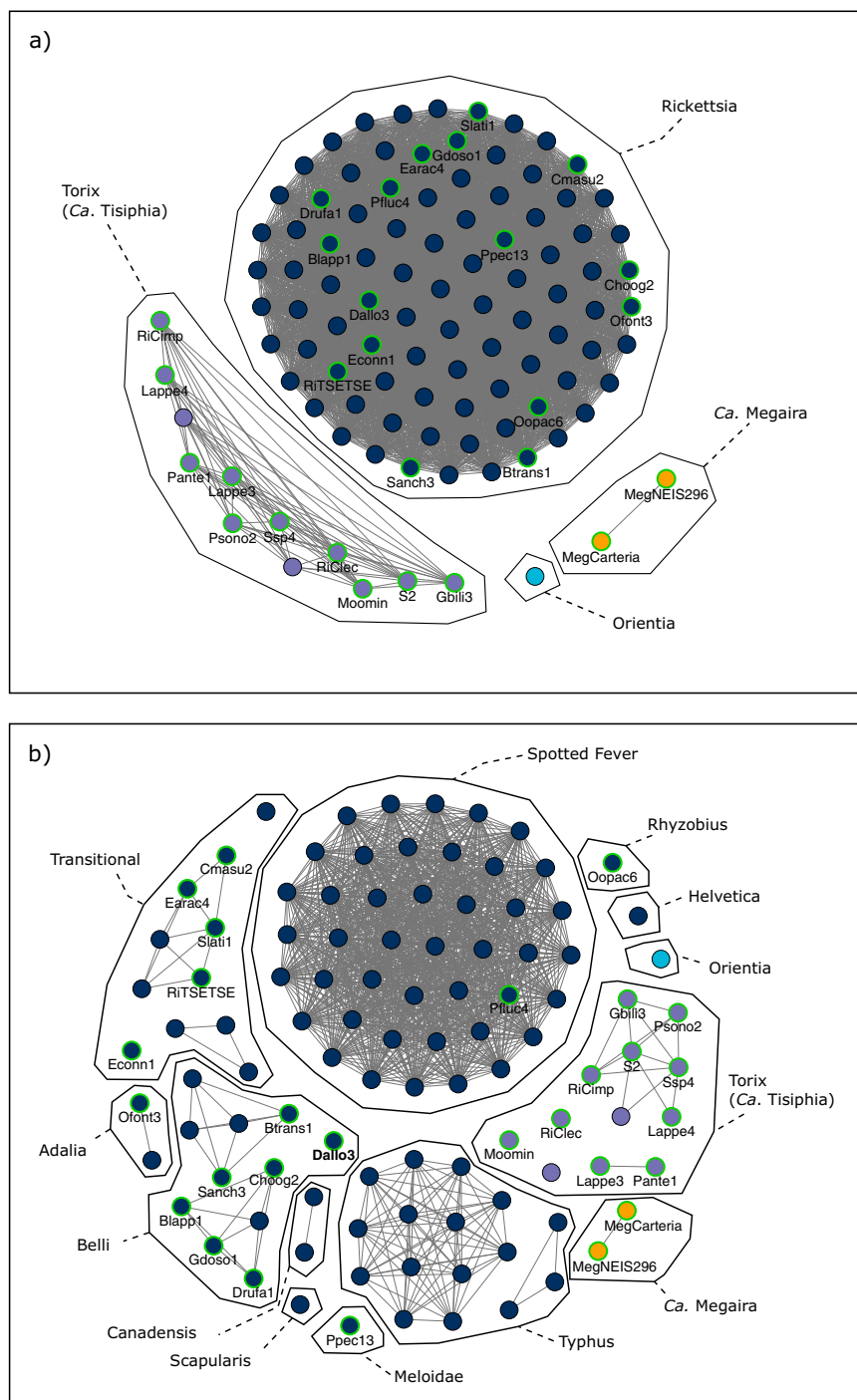
*alloeum*, used here has previously been described to contain a pseudogenised nuclear insert of *Rickettsia* material, but not a complete *Rickettsia* genome<sup>50</sup>. The construction of a full, non-pseudogenised genome with higher read depth than the insect contigs, low contamination (0.95%) and high completion (93.13%) suggests that these reads likely represent a viable *Rickettsia* infection in *D. alloeum*. However, these data do not exclude the presence of an additional nuclear insert. It is possible for a whole symbiont genome to be incorporated into the host's DNA like in the case of *Wolbachia*<sup>51</sup>, or the partial inserts of 'Ca. Megaira' genomes in the *Volvox carteri* genome<sup>52</sup>. The presence of both the insert and symbiont need confirmation through appropriate microscopy methods.

Recombination is low within the core genomes of *Rickettsia* and 'Ca. Megaira' but may occur between closely related clades that are not investigated here. Across all genomes, the PHI score is significant in 6 of the 74 core gene clusters, suggesting putative recombination events. However, it is reasonable to assume that most of these may be a result of systematic error due to the divergent evolutionary processes at work across *Rickettsia* genomes. Patterns of recombination can occur by chance rather than driven by evolution which cannot be differentiated by current phylogenetic methods<sup>53</sup>. The function of each respective cluster can be found in Supplementary Data 1.

**Gene content, pangenome and metabolic analysis.** Across all genomes used in the gene content comparison analysis (Supplementary Fig. 6), Anvi'o identified only 208 core gene clusters of which 74 are represented by single-copy genes. It is particularly evident the large size of the accessory genome across the main *Rickettsia* and the Torix clades. Out of the 2470 predicted ortholog clusters for the Torix clade 1296 (52.5%) are uniquely found among the Torix genomes, while for *Rickettsia* 2460 unique ortholog clusters were predicted from a total of 3811 (64.5%)

**Table 2 Summary of draft genomes generated during the current project and their associated hosts. Full metadata including CheckM completeness scores and levels of contamination can be found in Supplementary Data 1.**

Strain	Symbiotic bacteria assembly accession	Group	Number of contigs	Total length (bp)	Host name	Host Order
<b>Blapp1</b>	GCA_020404495.1	Belli	171	1266633	<i>Bembidion lapponicum</i>	Coleoptera
<b>Btrans1</b>	GCA_020404375.1	Belli	241	1417452	<i>Bembidion nr. transversale</i> OSAC:DRMaddison DNA3205	Coleoptera
<b>Choo2</b>	GCA_020404365.1	Belli	16	1357829	<i>Columbicola hoogastraali</i>	Phthiraptera
<b>Cmasu2</b>	GCA_020404525.1	Transitional	196	1295004	<i>Ceroptres masudai</i>	Hymenoptera
<b>Dallo3</b>	GCA_020404485.1	Belli	196	990679	<i>Diachasma alloenum</i>	Hymenoptera
<b>Drufa1</b>	GCA_020404445.1	Belli	14	1364611	<i>Degeerella rufa</i>	Phthiraptera
<b>Earc4</b>	GCA_020881375.1	Transitional	96	1350066	<i>Ecitomorpha arachnoides</i>	Coleoptera
<b>Econn1</b>	GCA_020881315.1	Transitional	238	1070326	<i>Eriopis connexa</i>	Coleoptera
<b>Gbil3</b>	GCA_020881275.1	Torix limoniae (Ca. Tisiphia)	171	1188102	<i>Gnoriste bilineata</i>	Diptera
<b>Gdoso1</b>	GCA_020881245.1	Belli	34	1420758	<i>Graphium doson</i>	Lepidoptera
<b>Lappe3</b>	GCA_02088125.1	Torix limoniae (Ca. Tisiphia)	122	1368980	<i>Labidopullus appendiculatus</i>	Coleoptera
<b>Lappe4</b>	GCA_020881075.1	Torix limoniae (Ca. Tisiphia)	154	1332357	<i>Labidopullus appendiculatus</i>	Coleoptera
<b>MegCarteria</b>	GCA_020881215.1	'Ca. Megaira'	72	1298707	<i>Carteria cerasiformis</i>	Chlamydomonadales
<b>Ofont3</b>	GCA_020404465.1	Adalia	91	1529137	<i>Omalisus fontisbellaquei</i>	Coleoptera
<b>Oopac6</b>	GCA_020881235.1	Rhyzobius	181	1497231	<i>Oxyopoda opaca</i>	Coleoptera
<b>Pante1</b>	GCA_020881195.1	Torix limoniae (Ca. Tisiphia)	70	1472610	<i>Pseudomimecton antennatum</i>	Coleoptera
<b>Pfluc4</b>	GCA_020404545.1	Spotted Fever	7	1251895	<i>Proechinophthirus fluctus</i>	Phthiraptera
<b>Ppec13</b>	GCA_020404425.1	Belli	90	1426047	<i>Pyrocoelia pectoralis</i>	Coleoptera
<b>Psono2</b>	GCA_020881175.1	Torix limoniae (Ca. Tisiphia)	163	1492063	<i>Platyusa sonomae</i>	Coleoptera
<b>RITSETSE</b>	GCA_020881295.1	Transitional	172	1451997	<i>Glossina moisitans submorsitans</i>	Diptera
<b>S2</b>	GCA_020404555.1	Torix limoniae (Ca. Tisiphia)	103	1251484	<i>Sericostoma</i>	Trichoptera
<b>Sanch3</b>	GCA_020881115.1	Belli	181	1487154	<i>Stiretrus anchorago</i>	Hemiptera
<b>Slati1</b>	GCA_020881155.1	Transitional	109	1301763	<i>Sceptobius lativentris</i>	Coleoptera
<b>Ssp4</b>	GCA_020404565.1	Torix limoniae (Ca. Tisiphia)	87	1231013	<i>Sericostoma</i> sp. HW-2014	Trichoptera
<b>Moomin</b>	GCA_020881085.1	Torix moomin (Ca. Tisiphia)	204	1137559	<i>Bryobia graminum</i>	Trombidiformes

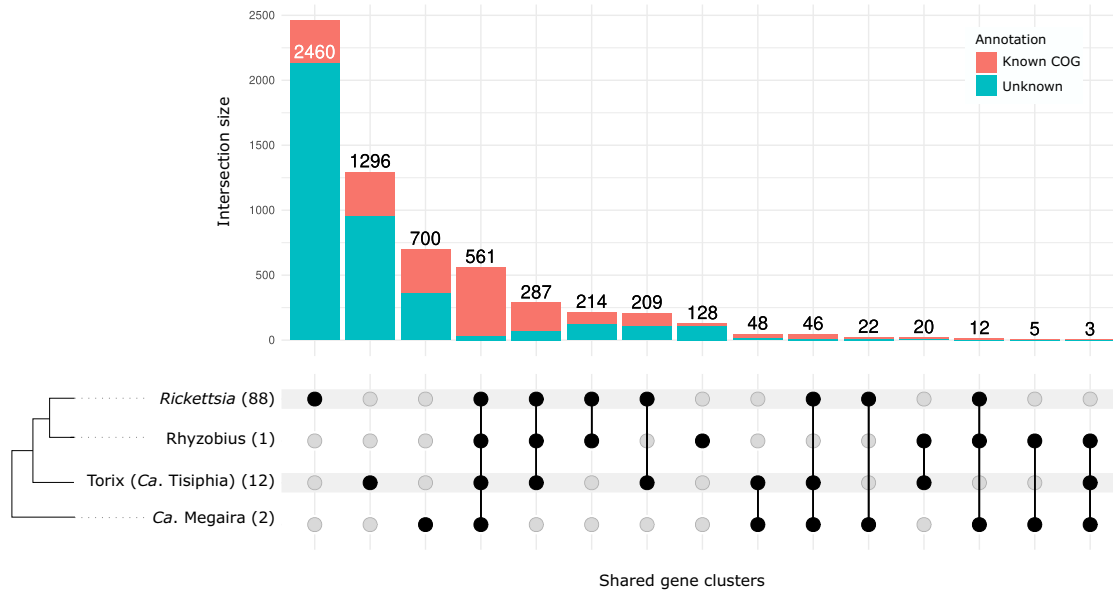


**Fig. 2 Genus and species level clustering across *Rickettsia* and 'Ca. Megaira'.** Fruchterman Reingold networks of pairwise (a) Average Amino Acid Identity (AAI) with edge weights >65% similarity and (b) Average Nucleotide Identity (ANI) with edge weights >95% similarity across all genomes. AAI and ANI illustrate genus and species boundaries, respectively. The 13 current cluster names are annotated over the 23 species clusters found in the ANI network. New genomes are named and have a green outline. Node fill colours indicate *Rickettsia* (Dark blue), 'Ca. Megaira' (orange), Torix/ 'Ca. Tisiphia' (purple), *Orientia* outgroup (light blue). Source data are provided in Source Data.

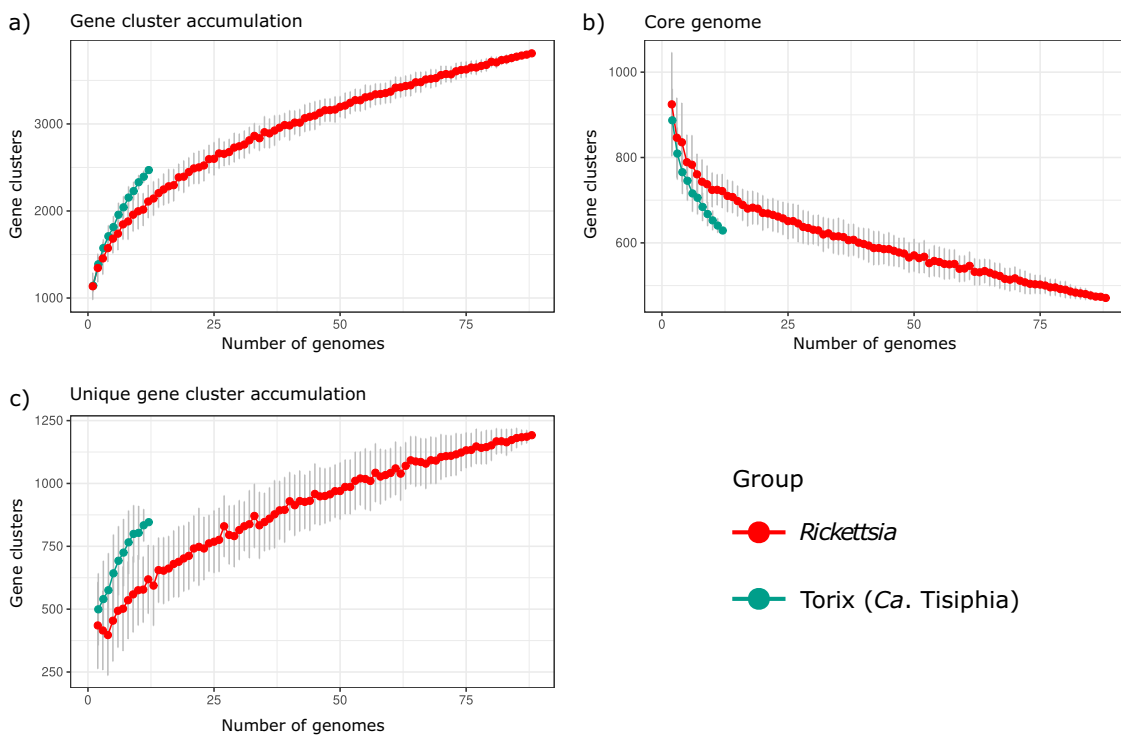
(Fig. 3). However, if we account for the number of genomes available in each clade then Torix shows higher rates of gene cluster and unique gene clusters accumulation with each additional genome (Fig. 4). Our results indicate that the main *Rickettsia* clade and especially the Torix clade, seem to have a high degree of genome diversity, suggesting a wider repertoire of genes and potentially greater rates of gene turnover. As expected, the more genomes that are included in analyses, the smaller the core genome extracted. However, gene content analysis results of

increasingly diverged genomes should be always interpreted with caution as true homology relationship between genes/proteins might get obscured by their sequence divergence.

Torix is a distinctly separate clade sharing less than 65% AAI similarity to any *Rickettsia* or 'Ca. Megaira' genomes (Fig. 2). It contains at least five species-level clusters with >95% ANI similarity that reflect its highly diverse niche in the environment (Fig. 2)<sup>13,54,55</sup>. With only two examples, the true diversity of 'Ca. Megaira' is underestimated here. Overall, our results indicate



**Fig. 3 Gene content comparison.** Shared and unique gene clusters across genus putative genus clusters *Rickettsia*, *Rhyzobius*, *Torix* and ‘*Ca. Megaira*’ as suggested by GTDB-tk. Vertical coloured bars represent the size of intersections (the number of shared gene clusters) between genomes in descending order with known COG functions displayed in coral and unknown in blue. Black dots mean the cluster is present and connected dots represent gene clusters that are present across groups. Numbers in parenthesis represent the number of genomes used in the analysis. Source data are provided in Source Data.

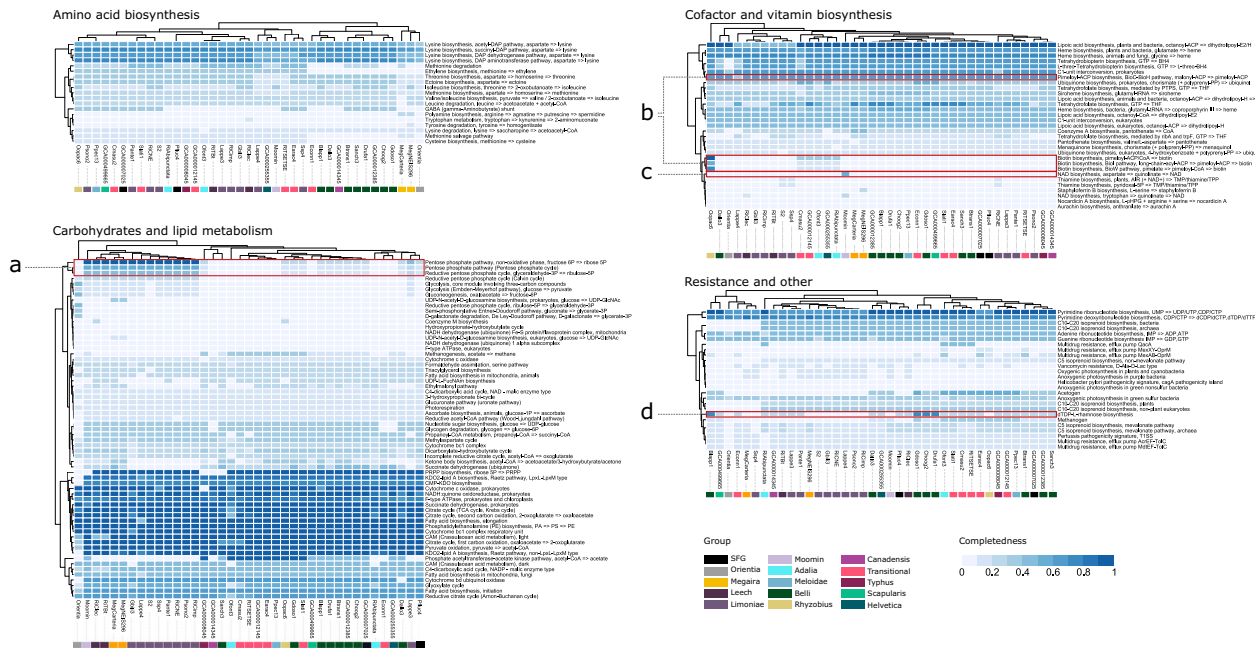


**Fig. 4 Gene cluster accumulation analysis.** **a** Pangenome accumulation curves. **b** Core genome accumulation curves. **c** The unique genome of *Rickettsia* (red) and *Torix* (turquoise) clades as a function of the number of genomes sequenced. Each point represents the mean value while error bars represent  $\pm$  standard deviation based on 100 permutations. Source data are provided in Source Data.

higher genomic plasticity within *Torix* clade in terms of gene content compared to *Rickettsia*.

We also investigated whether *Torix* and *Rickettsia* clades are enriched for particular COGs (Supplementary Data 1). Among the most highly enriched genes in *Torix* clade were genes encoding for invasion associated proteins like the

exopolysaccharide synthesis protein ExoD (COG3932) and the invasion associated protein IaB (COG5342), a carbonic anhydrase (COG0288) and a Chloramphenicol resistance associated protein (COG3896). Both carbonic anhydrase and ExoD homologs has been already reported in *Torix* clade<sup>45</sup> and our results here further support their important role in *Torix* biology. ExoD



**Fig. 5 Comparison of metabolic potential across selected *Rickettsia* and ‘Ca. Megaira’.** Heatmaps of predicted KEGG pathway completion estimated in Anvi’o 7, separated by function, and produced with Pheatmap. High to low completeness is coloured dark to light blue. Species groups are indicated with a unique colour as shown in the legend. Pathways of interest are highlighted in red: **a** The pentose phosphate pathway only present in Torix and ‘Ca. Megaira’, **b** The biotin pathway present only in the Rhyzobius *Rickettsia* Oopac6. **c** NAD biosynthesis only present in Moomin *Rickettsia*. **d** dTDP-L-rhamnose biosynthesis pathway in Gdosol, Choog2, Drufa1, and Blapp1. SFG is Spotted Fever Group. Source data are provided in Source Data.

has been previously reported as essential for successful nodule invasion of the nitrogen-fixing endosymbiont *Rhizobium*<sup>56</sup>. When we consider both Torix and ‘Ca. Megaira’ clades the genes involved in the non-oxidative phase of the PPP pathway were the most highly enriched genes (Supplementary Data 1). It is noteworthy that a large fraction of the enriched genes in both *Rickettsia* and Torix clades are related to cell wall and membrane biogenesis. These are likely associated with differences in the biology of the two clades at the host-microbe interface.

*Rickettsia* lineages group together based on gene presence/absence and produce repeated patterns of accessory genes that reliably occur within each clade (Supplementary Fig. 6). AAI scores separate Torix group, *Rickettsia* and ‘Ca. Megaira’ out into genus groups with no score above 65% similarity outside of each respective clade (Fig. 2)<sup>57</sup>. ANI scores suggest that Torix and the remaining *Rickettsia* clades are multispecies clusters with less than 95% similarity between genomes in the same groups except for the Spotted Fever Group (Fig. 2)<sup>57</sup>.

*Rickettsia* genomes extracted from SRA samples are generally congruent with the metabolic potential of their respective groups (Fig. 5). Torix and ‘Ca. Megaira’ all have complete pentose phosphate pathways (PPP); a unique marker for these groups which seems to have been lost in the other *Rickettsia* clades<sup>45</sup>. The PPP generates NADPH, precursors to amino acids, and is known to protect against oxidative injury in some bacteria<sup>58</sup>, as well as conversion of hexose monosaccharides into pentose used in nucleic acid and exopolysaccharide synthesis. The PPP has also been associated with establishing symbiosis between the *Alpha-proteobacteria* *Sinorhizobium meliloti* and its plant host *Medicago sativa*<sup>59</sup>. This pathway has previously been highlighted in Torix<sup>45</sup> and its presence in all newly assembled Torix and ‘Ca. Megaira’ draft genomes consolidates its importance as an identifying feature for these groups (Fig. 5, Supplementary Data 1). Considering the trend towards gene loss, the PPP is likely an ancestral feature that was lost in the main *Rickettsia* clade<sup>45,60</sup>.

Metabolic pathways for Glycolysis, gluconeogenesis, and cofactor/vitamin synthesis are absent or incomplete across all *Rickettsia* included in these analyses, except in the Rhyzobius group member, Oopac6. Oopac6 has a putatively complete biotin synthesis pathway (Fig. 5, Supplementary Fig. 7) and is likely a separate genus according to GTDBtk analysis (Supplementary Data 1). The Oopac6 biotin synthesis pathway is related to, but distinct from, the *Rickettsia* biotin pathway from *Rickettsia buchneri*<sup>36</sup> with which it shares between 85% to 92% amino acid sequence similarity across genes (Supplementary Fig. 7)<sup>36</sup>. Moreover, there is no sequence similarity outside of the biotin operon. This, along with the presence on a plasmid in *Rickettsia buchneri* makes it likely that Oopac6 operon is a result of horizontal gene transfer. Animals cannot synthesize B-vitamins, so they either acquire them from diet or from microorganisms that can synthesize them. Oopac6 has retained or acquired a complete biotin operon where this operon is absent in other members of the genus. Biotin pathways in insect symbionts can be an indicator of nutritional symbioses<sup>61</sup>, so Rhyzobius *Rickettsia* could contribute to the feeding ecology of the beetle *O. opaca*. However, like other aleocharine rove beetles, *O. opaca* is likely predaceous, omnivorous or fungivorous (analysis of gut contents from a related species, *O. grandipennis*, revealed a high prevalence of yeasts<sup>62</sup>). We posit no obvious reason for how these beetles benefit from harbouring a biotin-producing symbiont. One theory is that this operon could be a hangover from a relatively recent host shift event and may have been functionally important in the original host. Similarly, if the symbiont is undergoing genome degradation, a once useful biotin pathway may be present but not functional<sup>63,64</sup>. Although the pimeloyl-ACP biosynthesis pathway is partially present (Fig. 5), a *bioH* homolog is not found within or outside the biotin operon (Supplementary Fig. 7) suggesting that this pathway may not be functional (as observed in some *Buchnera aphidicola*<sup>64,65</sup>) or that it may be used in a different way. As this is the only member of this group with a



whole genome so far, further research is required to firmly establish the presence and function of this pathway.

A 75% complete dTDP-L-rhamnose biosynthesis pathway was observed in 4 of the draft belli assemblies (Gdosol, Choog2, Drufal, Blapp1) (Fig. 5). Two host species are bird lice (*Columbicola hoogstraali*, *Degeeriella rufa*), one is a butterfly (*Graphium doson*), and one is a ground beetle (*Bembidion lapponicum*). dTDP-L-rhamnose is an essential component of human pathogenic bacteria like *Pseudomonas*, *Streptococcus* and *Enterococcus*, where it is used in cell wall construction<sup>66</sup>. This pathway<sup>67</sup> may be involved in the moulting process of *Caenorhabditis elegans*<sup>68</sup>, and it is a precursor to rhamnolipids that are used in quorum sensing<sup>69</sup>. In the root symbiont *Azospirillum*, disruption of this pathway alters root colonisation, lipopolysaccharide structure and exopolysaccharide production<sup>70</sup>. No *Rickettsia* from typically pathogenic groups assessed in Fig. 5 has this pathway, and the hosts of these four bacteria are not involved with human or mammalian disease. Presence in feather lice provides little opportunity for this *Rickettsia* to be pathogenic to their vertebrate hosts because feather lice are not blood feeders, and Belli group *Rickettsia* are rarely pathogenic. Further, this association does not explain its presence in a butterfly and ground beetle; it is most likely that this pathway, if functional, would be involved in establishing infection in the insect host or host-symbiont recognition.

A partial NAD biosynthesis pathway is present only in the Moomin genome. NAD is used as a coenzyme in numerous reactions as well as a substrate in some synthesis pathways, such as ADP-Ribosyltransferases which are used in bacterial toxin-antitoxin systems<sup>71,72</sup>. NAD pathways have previously found in two other members of *Rickettsiaceae*, '*Ca. Sarmatiella mevalonica*' and '*Occidentia massiliensis*'<sup>31,73</sup>. The most likely explanation for rare occurrence in *Rickettsiaceae* is either a lateral transfer event or remnants from ancestral occurrence.

**Designation of '*Candidatus Tisiphia*'.** In all analyses, Torix group consistently clusters away from the rest of *Rickettsia* as a sister taxon. Despite the relatively small number of Torix

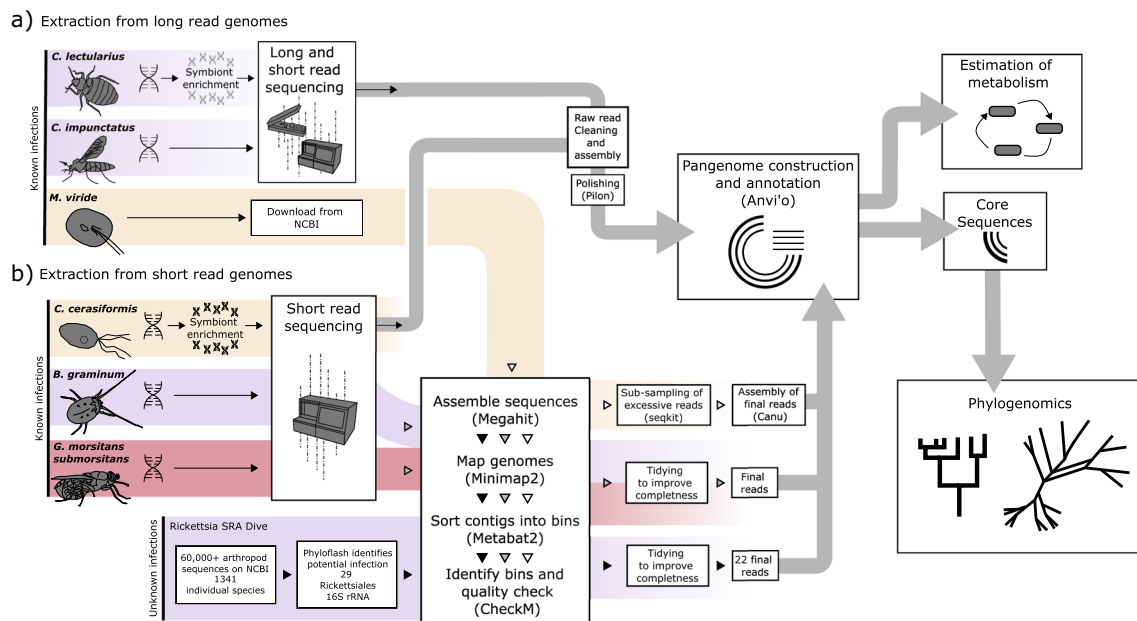
genomes, its within group diversity is greater than any divergence between previously described *Rickettsia* in any other group (Fig. 1, Supplementary Figs. 3 and 4). Additionally, Torix shares characteristics with both '*Ca. Megaira*' and *Rickettsia*, but with many of its own unique features (Figs. 3 and 5). The distance of Torix from other *Rickettsia* and '*Ca. Megaira*' is confirmed in both the phylogenomic and metabolic function analyses to the extent that Torix should be separated from *Rickettsia* and assigned its own genus. This is supported by GTDB-Tk analysis which places all Torix genomes separate from *Rickettsia* (Supplementary Data 1) alongside AAI percentage similarity scores less than 65% in all cases (Fig. 2a). To this end, we propose the name '*Candidatus Tisiphia*'. This name follows the fury Tisiphone, reflecting the genus '*Ca. Megaira*' being named after her sister Megaira.

**Conclusions**

The bioinformatics approach has successfully extracted a substantial number of *Rickettsia* and '*Ca. Megaira*' genomes from existing SRA data, including genomes for putative Rhyzobius *Rickettsia* and several '*Ca. Tisiphia*' (formerly Torix group *Rickettsia*). Successful completion of two '*Ca. Megaira*' and two '*Ca. Tisiphia*' genomes provide solid reference points for the evolution of *Rickettsia* and its related groups. From this, we can confirm the presence of a complete Pentose Phosphate Pathway in '*Ca. Tisiphia*' and '*Ca. Megaira*', suggesting that this pathway was lost during *Rickettsia* evolution. We also describe previously unsequenced Meloidae and Rhyzobius *Rickettsia* and show that Rhyzobius group *Rickettsia* has the potential to be a nutritional symbiont due to the presence of a complete biotin pathway. These genomes provide a much-needed expansion of available data for symbiotic *Rickettsia* clades and clarification on the evolution of *Rickettsia* from '*Ca. Megaira*' and '*Ca. Tisiphia*'.

**Methods**

**Genomic data collection and construction.** We employed two different workflows to assemble genomes for '*Ca. Megaira*' and *Rickettsia* symbionts (Fig. 6). a) Targeted sequencing and assembly of focal '*Ca. Megaira*' and Torix *Rickettsia*.



**Fig. 6 Workflow diagram for extraction, assembly and analyses performed in this study.** Workflows for genome assemble are illustrated for (a) long read host insect sequences and (b) short read host insect sequences. Purple highlights Torix *Rickettsia* and orange highlights '*Ca. Megaira*' and red highlights Transitional *Rickettsia*. Sequencing technologies used vary with source and include Illumina short read sequencing, BGI DNBseq, Oxford Nanopore and PacBio.

b) Assembly from SRA deposits of ‘*Ca. Megaira*’ from *Mesostigma viride* NIES296 and the 29 arthropods identified in Pilgrim et al.<sup>13</sup> that potentially harbour *Rickettsia*. These were analysed alongside previously assembled genomes from the genus *Rickettsia*, and the outgroup taxon *Orientia tsutsugamushi*, a distant relative of *Rickettsia* species<sup>74</sup>.

DNA preparation, sequencing strategies and symbiont assembly methodologies varied between species and are listed below. The pipeline used to assemble genomes from Short Read Archive (SRA) data is deposited on Zenodo<sup>75</sup>.

**Sample collection for targeted genome assembly.** *Cimex lectularius* were acquired from the ‘S1’ isofemale colony maintained at the University of Bayreuth described in Thongprem et al.<sup>26</sup>. *Culicoides impunctatus* females were collected from a wild population in Kinlochleven, Scotland (56° 42′ 50.7″N 4° 57′ 34.9″W) on the evenings of the 2nd and 3rd September 2020 by aspiration. *Carteria cerasiformis* strain NIES 425 was obtained from the Microbial Culture Collection at the National Institute for Environmental Studies, Japan. The *Glossina morsitans submorsitans* specimen Gms8 was collected in Burkina Faso in 2010 and *Rickettsia* infection was present alongside other symbionts as described in Doudoum et al.<sup>76</sup>. The assembly itself is a result of later thesis work<sup>77</sup>.

A *Bryobia* mite community was sampled from herbaceous vegetation in Turku, Finland. The Moomin isofemale line was established by isolating a single adult female and was maintained on detached leaves of *Phaseolus vulgaris* L. cv Speedy at 25 °C, 60% RH, and a 16:8 light:dark photoperiod. The Moomin spider mite line was morphologically identified as *Bryobia graminum* by Prof Eddie A. Ueckermann (North-West University).

**Previously published *Rickettsia* genomes.** A total of 86 published *Rickettsia* genomes, and one genome from *Orientia tsutsugamushi* were retrieved from the European Nucleotide Archive and assessed with CheckM v1.0.13<sup>78</sup>. Inclusion criteria for genomes were high completeness (CheckM > 90%), low contamination (CheckM < 2%) and low strain heterogeneity (Check M < 50%) except in the case of *Adalia* for which there is only one genome (87.6% completeness). Filtering identified 76 high quality *Rickettsia* genomes that were used in all subsequent analyses (Supplementary Data 1).

**High molecular weight DNA extraction, assembly, and annotation of complete genomes for two ‘*Ca. Tisiphia*’ (=*Torix* group *Rickettsia*) from *Culicoides impunctatus* and *Cimex lectularius*.** High molecular weight (HMW) genomic DNA was prepared using four-hundred and eighty whole *C. impunctatus* and 45 *C. lectularius* heads, the latter of which had been symbiont-enriched using a protocol designed to eliminate host nuclei through filtration<sup>79</sup>. *Culicoides impunctatus* individuals were pooled and homogenised in two 1.5 ml Eppendorf tubes containing 0.9 ml of buffer G2 (Qiagen) using a pestle while the filtrate from the enriched *C. lectularius* heads was also split and diluted to the same volumes. Twenty-five µl of proteinase K (50 mg/ml) was added to each Eppendorf before incubation at 56 °C for 90 minutes with gentle inversion every 30 min. The respective lysates were centrifuged at 12,000 x g for 20 minutes before the supernatants were pooled and diluted to 3 ml with buffer G2. After equilibrating a Genomic-tip 20/G (Qiagen) with 1 ml QBT buffer, the lysate was gently inverted before being poured onto the tip membrane. The tip was washed four times with 1 ml of QC buffer (Qiagen) before elution of the DNA using buffer QF (Qiagen). Using wide-bore pipette tips, 667 µl of the eluate was pipetted into three 1.5 ml Eppendorf tubes before the addition of 467 µl isopropanol to each tube and mixing by gentle inversion 10 times. Genomic DNA was pelleted by centrifuging for 20 min at 15,000 x g at 4 °C and washing twice with 70% ethanol before resuspending in buffer EB (Qiagen). Quality control of HMW DNA was then confirmed by running on a gel and assessment by Qubit fluorometric quantitation.

Long-read libraries for Oxford Nanopore sequencing were generated using the SQK-LSK109 Ligation Sequencing Kit and sequenced on Minion R9.4.1 flow cells at the Centre for Genomic Research, University of Liverpool, United Kingdom. Raw Nanopore reads were base called using Guppy version 4.0.15 (Oxford Nanopore) using the high accuracy model option (-c dna\_r9.4.1\_450bps\_hac.cfg). All reads which were over 500 bp in length and had an average phred (Q) score of > 10 were filtered using NanoFilter version 2.7.1<sup>80</sup>. These reads were assembled with Flye version 2.8.1<sup>81</sup> using default options.

Assembled circular contigs of ~1.5 Mb in length were confirmed for *Rickettsia* identity by BLASTing against a *Rickettsia* genomic database<sup>45</sup>. High quality short-read libraries were also generated from the same DNA samples and used to correct the nanopore assemblies. *C. impunctatus* paired-end library (2 x 150 bp) was prepared using a Kapa HyperPrep kit (Roche) and sequenced by BGI Genomics (Hong Kong) on a DNBseq platform, whereas *C. lectularius* sequencing was carried out by BGI Genomics (Hong Kong) on a HiSeq Xten PE150 platform. Data cleaning and filtering was performed by BGI Genomics’ using SOAPnuke version 2.1.4<sup>82</sup> removing adapters and any reads with 50% of bases having phred scores lower than 20.

Remaining reads were assembled with MEGAHIT version 1.2.9<sup>83</sup> using default parameters and contigs were binned using MetaBAT 2 version 2.12.1<sup>84</sup>. The identities of bins were checked with CheckM version 1.1.3<sup>78</sup> and DNBseq reads were mapped to contigs from the *Rickettsia* allocated bin using ‘perfect mode’ in

BBMap version 38.87<sup>85</sup> and filtered using SAMtools version 1.11<sup>86</sup>. Filtered *Rickettsia* reads were then used to polish the Flye assembled *Rickettsia* genomes using two rounds of polishing with Pilon version 1.23<sup>87</sup> and the ‘—bases’ option for correcting SNPs and small indels. Annotation of the polished genomes was accomplished using PROKKA version 1.13<sup>88</sup> and identification of polyketide and non-ribosomal peptide synthases was conducted by antiSMASH version 6.0<sup>89</sup>.

**Extraction and assembly of a complete ‘*Ca. Megaira*’ from *Mesostigma viride*.** ‘*Ca. Megaira*’ genome was extracted from recently published reads of *Mesostigma viride* NIES296 (from accession PRJNA509752). Illumina reads were de novo assembled using MEGAHIT version v1.2.9<sup>83</sup>, reads were mapped back to the assembled contigs. Contigs were clustered and binned based on nucleotide composition and coverage using MetaBAT2 v2.2.15<sup>84</sup> and a minimum contig length of 1.5 kb. The quality of ‘*Ca. Megaira*’ genome bin was inspected using CheckM<sup>78</sup>. The PacBio reads were mapped on the Illumina draft assembly and reads of ‘*Ca. Megaira*’ origin were extracted. Due to the excessive number of obtained reads a sub-sample (reads > 10 k and 1/3 of the total) was taken using seqkit<sup>90</sup> and used for subsequent analysis. This sub-sample of PacBio reads was assembled using Canu version 1.8<sup>91</sup> under default parameters. The final assembly, consisted of two contigs, was manually inspected for circularization and trimmed accordingly. The final and circular assembly was further polished by a combination of PacBio and Illumina reads using Pilon v1.22<sup>87</sup>.

**Extraction of Transitional *Rickettsia*, RiTSETSE, from *Glossina morsitans submorsitans*.** All methods described here for the extraction of *G. morsitans submorsitans* originate from a thesis by Frances Blow<sup>77</sup>.

DNA was extracted immediately using the CTAB (Cetyl trimethylammonium bromide) method and was stored at -20 °C. Whole Genome Shotgun (WGS) libraries were prepared with the Illumina TruSeq Nano DNA kit following the manufacturers’ instructions. Samples were sequenced on two lanes of Illumina HiSeq with 250 bp paired-end reads. Raw sequencing reads were de-multiplexed and converted to FASTQ format with CASAVA version 1.8 (Illumina). Cutadapt version 1.2.1<sup>92</sup> was used to trim Illumina adapter sequences from FASTQ files. Reads were trimmed if 3 bp or more of the 3’ end of a read matched the adapter sequence. Sickle version 1.200<sup>93</sup> was used to trim reads based on quality: any reads with a window quality score of less than 20, or which were less than 10 bp long after trimming, were discarded.

Metagenomic reads were assembled with DISCOVAR<sup>94</sup> and contigs shorter than 500 bp were removed and mapping with Bowtie<sup>95</sup> was used to assess coverage. Taxonomy was assigned to contigs with BLAST and the GC content of contigs assessed with the Blobology package<sup>96</sup>. Contigs were filtered based on GC content, coverage and taxonomy, and reads were extracted using scripts implemented in Blobology. Extracted reads were re-assembled with SPAdes version 3.7.1<sup>97</sup> and mapped to contigs with Bowtie2. Assembly statistics were calculated with custom perl scripts and Qualimap version 2.2<sup>98</sup>.

**DNA extraction of Moomin ‘*Ca. Tisiphia*’ (=*Torix* group *Rickettsia*) from *Bryobia graminum* str. moomin.** Genomic DNA was extracted from ~1000 adult females using the Quick-DNA Universal kit (BaseClear, the Netherlands) and was sequenced by GENEWIZ on an Illumina NovaSeq instrument. *Rickettsia* sequence was extracted from illumina reads as described for other MAGs.

**DNA extraction of ‘*Ca. Megaira*’ from *Carteria cerasiformis*.** Symbiont enriched DNA was extracted from culture using a modified version of the protocol of Stouthamer et al.<sup>79</sup>. Specifically, prior to homogenization the *Carteria cerasiformis* culture was filtered through a 100µm filter/mesh to reduce bacterial contamination. DNA extraction was performed using the QIAGEN DNeasy™ Blood & Tissue Kit. Short read sequencing was carried out by BGI Genomics (Hong Kong) on a HiSeq Xten PE150 platform. *Rickettsia* sequences were assembled from Illumina reads as described for other MAGs.

**Assembly, and annotation of *Rickettsia* genomes from publicly available SRA data.** Pilgrim et al.<sup>13</sup> identified 29 SRA deposits containing *Rickettsia* DNA. We used these datasets to extract and assemble 22 new high quality draft *Rickettsia* genomes. Briefly, short reads from each SRA library were assembled using MEGAHIT v1.2.9<sup>83</sup>, mapped with Minimap 2 v2.17-r941<sup>99</sup> and contigs were binned based on tetranucleotide frequencies using MetaBAT2 v2.2.15<sup>84</sup>. *Rickettsia* like bins were quality inspected with CheckM v1.0.13<sup>78</sup>. Bins with a completeness score of over 50% and contamination below 2% marked as *Rickettsiales* or *Rickettsia* were then retained onward for further refinement, annotation, and scrutiny.

To refine MAGs, insect SRA contigs were compared against a local *Rickettsia* genome database using Blastn<sup>100</sup>. Contigs with significant matches to the database were extracted, non-*Rickettsia* contigs were identified with blastx against the nr database and contigs with atypical coverage were discarded. MetaBAT 2 filtered out reads less than 1.5 kb long for accuracy, but these reads are potentially informative in small symbiont genomes, so contigs with a length of 1–2.5 kb were manually examined and added to MetaBAT 2 assembled genomes. Those with improved CheckM score and no *Wolbachia* in the original host are used as the final draft genome for the *Rickettsia*. The additional genome for the leech *Rickettsia*, RiTBt,

was found to contain *Cardinium* contamination during separate examination. RiTB contigs identified as *Cardinium* using blastx were removed from the genome, reducing contamination from 9.48% to 0.95%. The final pipeline resulted in 22 MAGs each with completeness >90% and contamination <2%.

**Genome content comparison and pangenome construction.** Anvi'o 7<sup>101</sup> was used to construct a pangenome. Included in this were the 22 MAGs retrieved from SRA data, 2 'Ca. Megaira' genomes and 4 targeted Torix *Rickettsia* genomes, and one Transitional group *Rickettsia* genome acquired in this study. To these were added the 76 published and 1 *Orientia* described above, giving a total of 104 genomes. Individual Anvi'o genome databases were additionally annotated with HMMER, KofamKOALA, and NCBI COG profiles<sup>102–104</sup>. For the pangenome itself, orthologs were identified with NCBI blast, mcl inflation was set to 2, and minbit to 0.5. Average nucleotide sequence identity was calculated using pyANI<sup>105</sup> within Anvi'o 7 and Average Amino Acid identity was calculated through the Kostas Lab enveomics online calculator<sup>106</sup>. Networks of ANI and AAI results were produced in Gephi 0.9.2<sup>107</sup> with Frutcherman Reingold layout and annotated in Inkscape 0.92<sup>108</sup>. Exact code and a list of packages used is available on Zenodo<sup>75</sup>.

KofamKOALA annotation<sup>103</sup> in Anvi'o 7 was used to estimate completeness of metabolic pathways and Pheatmap<sup>109</sup> in R 3.4.4<sup>110</sup> was then used to produce heatmaps of metabolic potential. Annotations for function and *Rickettsia* group were added post hoc in Inkscape.

The biotin operon found in the genome *Rhizobium Rickettsia*, Oopac6, was identified from metabolic prediction. To confirm Oopac6 carries a complete biotin pathway that shares ancestry with the existing *Rickettsia* biotin operon, Oopac6 biotin was compared to biotin pathways from five other related symbionts: *Cardinium*, *Lawsonia*, *Buchnera aphidicola*, *Rickettsia buchneri*, and *Wolbachia*. Clinker<sup>111</sup> with default options was used to compare and visualise the similarity of genes within the biotin operon region of all 6 bacteria. Clinker by default displays the highest similarity comparisons based on an all-vs-all similarity matrix.

All generated draft and complete reference genomes were annotated using the NCBI's Prokaryotic Genome Annotation Pipeline (PGAP)<sup>112</sup>. Secondary metabolite biosynthetic gene clusters were identified using AntiSMASH version 6.0<sup>89</sup> along with Norine<sup>113</sup> which searched for similarities to predicted non-ribosomal peptides. BLASTp analysis was additionally used to identify the closest homologues of these biosynthetic gene clusters.

Functional enrichment analyses between the main *Rickettsia* clade and the Torix - 'Ca. Megaira' clades were performed using the Anvi'o program anvi-get-enriched-functions-per-pan-group and the "COG\_FUNCTION" as annotation source. A gene cluster presence - absence table was exported using the command "anvi-export-tables". This was used to create an UpSet plot using the R package ComplexUpset<sup>114</sup> to visualize unique and shared gene clusters between different *Rickettsia* groups. A gene cluster was considered unique to a specified *Rickettsia* group when it was present in at least one genome belonging to that group. Gene cluster accumulation curves were performed for the pan-, core- and unique-genomes based on the same presence-absence matrix using a custom-made R script<sup>115</sup>. In each case the cumulative number of gene clusters were computed based on randomly sampled genomes using 100 permutations. The analysis was performed separately for Torix group and the combined remaining *Rickettsia*. Curves were plotted using the ggplot2 R package<sup>116</sup>.

All information on extra genomes can be found in Supplementary Data 1, and the code pipeline employed can be found on Zenodo<sup>75</sup>.

**Phylogeny, network, and recombination.** The single-copy core of all 104 genomes was identified in Anvi'o 7 and is made up of 74 single-copy gene (SCG) clusters. Protein alignments from SCG were extracted and concatenated using the command "anvi-get-sequences-for-gene-clusters". Maximum likelihood phylogeny was constructed in IQ-TREE v2.1.2<sup>117</sup>. Additionally, 43 ribosomal proteins were identified through Anvi'o 7 to test phylogenomic relationships. These gene clusters were extracted from the pangenome and used for an independent phylogenetic analysis. The best model according to the Bayesian Information Criterion (BIC) was selected with Model Finder Plus (MFP)<sup>118</sup> as implemented in IQ-TREE; this was JTTDCMut+F+R6 for core gene clusters and JTTDCMut+F+R3 for ribosomal proteins. Both models were run with Ultrafast Bootstrapping (1000 UF bootstraps)<sup>119</sup> with *Orientia tsutsugamushi* as the outgroup.

The taxonomic placement of Oopac6, Ppec13 and Dallo3 genomes within the Rhizobium, Meloidae and Belli groups respectively were confirmed in a smaller phylogenetic analysis, performed as detailed in Pilgrim et al.<sup>13</sup> using reference MLST sequences (gltA, 16 S rRNA, 17 kDa OMP, COI) from other previously identified *Rickettsia* profiles (Source Data). The selected models used in the concatenated partition scheme were as follows: 16 S rRNA: TIM3e+I+G4; 17Kda OMP: GTR+F+I+G4; COI: TPM3u+F+I+G4; gltA: K3Pu+F+I+G4a.

A nearest neighbour network was produced for core gene sets with default settings in Splitstree4 to further assess distances and relationships between *Rickettsia*, 'Ca. Megaira' and Torix clades. All annotation was added post hoc in Inkscape. Furthermore, recombination signals were examined by applying the Pairwise Homoplasy Index (PHI) test to the DNA sequence of each core gene cluster extracted with Anvi'o-7. DNA sequences were aligned with MUSCLE<sup>120</sup> and PHI scores calculated for each of the 74 core gene cluster with PhiPack<sup>121</sup>.

The taxonomic identity for genomes was established with GTDB-Tk<sup>122</sup> to support the designation of taxa through phylogenetic comparison of marker genes against an online reference database.

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

## Data availability

The genomes and raw read sets generated in this study have been deposited in the GenBank database under accession code PRJNA763820. The assemblies produced from previously published third party data have been deposited in the GenBank database under accession code PRJNA767332. The genome content data and data for figures generated in this study are provided in the Source Data and Supplementary Data. Accessions and metadata for pre-existing genomic data are listed in the Supplementary Data 1 file.

## Code availability

All code and bioinformatics pipelines used to extract and construct bacterial genomes from SRA data can be found on Zenodo (<https://doi.org/10.5281/zenodo.6396821>), and the R script for generating pangenome accumulation curves can be found on GitHub (<https://github.com/SioStef/panplots> and here [10.5281/zenodo.6408803](https://doi.org/10.5281/zenodo.6408803)). The full pangenome Anvi'o database is available on Figshare (<https://doi.org/10.6084/m9.figshare.14865576.v3>). An interactive html version of Fig. 5 and its associated 'json' file is available on Figshare (<https://doi.org/10.6084/m9.figshare.14865567.v5>). html of bonzai module information for Supplementary Fig. 2 is available on Figshare (<https://doi.org/10.6084/m9.figshare.14865570.v4>).

Received: 18 October 2021; Accepted: 29 April 2022;

Published online: 12 May 2022

## References

- Clay, K., Holah, J. & Rudgers, J. A. Herbivores cause a rapid increase in hereditary symbiosis and alter plant community composition. *Proc. Natl. Acad. Sci.* **102**, 12465–12470 (2005).
- Boettcher, K. J., Ruby, E. G. & McFall-Ngai, M. J. Bioluminescence in the symbiotic squid *Euprymna scolopes* is controlled by a daily biological rhythm. *J. Comp. Physiol. A* **179**, 65–73 (1996).
- Douglas, A. E. Lessons from studying insect symbioses. *Cell Host Microbe* **10**, 359–367 (2011).
- Fujishima, M. & Kodama, Y. Endosymbionts in Paramecium. *Eur. J. Protistol.* **48**, 124–137 (2012).
- Oliver, K. M., Degnan, P. H., Burke, G. R. & Moran, N. A. Facultative symbionts in aphids and the horizontal transfer of ecologically important traits. *Annu. Rev. Entomol.* **55**, 247–266 (2010).
- Hendry, T. A., Hunter, M. S. & Baltrus, D. A. The facultative symbiont *Rickettsia* protects an invasive whitefly against entomopathogenic *Pseudomonas syringae* strains. *Appl. Environ. Microbiol.* **80**, 7161–7168 (2014).
- Leclair, M. et al. Consequences of coinfection with protective symbionts on the host phenotype and symbiont titres in the pea aphid system. *Insect Sci.* **24**, 798–808 (2017).
- Engelstädter, J. & Hurst, G. D. D. The ecology and evolution of microbes that manipulate host reproduction. *Annu. Rev. Ecol. Evol. Syst.* **40**, 127–149 (2009).
- Weinert, L. A., Araujo-Jnr, E. V., Ahmed, M. Z. & Welch, J. J. The incidence of bacterial endosymbionts in terrestrial arthropods. *Proc. R. Soc. B Biol. Sci.* **282**, 20150249 (2015).
- 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).
- Angelakis, E. & Raouf, D. 187 - Rickettsia and Rickettsia-Like Organisms. In *Infectious Diseases* (Fourth Edition) (eds. Cohen, J., Powderly, W. G. & Opal, S. M.) 1666–1675.e1 (Elsevier, 2017). <https://doi.org/10.1016/B978-0-7020-6285-8.00187-8>.
- Werren, J. H. et al. Rickettsial relative associated with male killing in the ladybird beetle (*Adalia bipunctata*). *J. Bacteriol.* **176**, 388–394 (1994).
- Pilgrim, J. et al. Torix *Rickettsia* are widespread in arthropods and reflect a neglected symbiosis. *GigaScience* **10**, giab021 (2021).
- 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).
- Bodnar, J. L., Fitch, S., Rosati, A. & Zhong, J. The folA gene from the Rickettsia endosymbiont of *Ixodes pacificus* encodes a functional dihydrofolate reductase enzyme. *Ticks Tick-Borne Dis.* **9**, 443–449 (2018).

16. Łukasik, P., Guo, H., van Asch, M., Ferrari, J. & Godfray, H. C. J. Protection against a fungal pathogen conferred by the aphid facultative endosymbionts *Rickettsia* and *Spiroplasma* is expressed in multiple host genotypes and species and is not influenced by co-infection with another symbiont. *J. Evol. Biol.* **26**, 2654–2661 (2013).
17. Hurst, G. D. D., Purvis, E. L., Sloggett, J. J. & Majerus, M. E. N. The effect of infection with male-killing *Rickettsia* on the demography of female *Adalia bipunctata* L. (two spot ladybird). *Hereditas* **73**, 309–316 (1994).
18. Giorgini, M., Bernardo, U., Monti, M. M., Nappo, A. G. & Gebiola, M. *Rickettsia* symbionts cause parthenogenetic reproduction in the parasitoid wasp *pnigalio soemius* (Hymenoptera: Eulophidae). *Appl. Environ. Microbiol.* **76**, 2589–2599 (2010).
19. Brumin, M., Kontsedalov, S. & Ghanim, M. *Rickettsia* influences thermotolerance in the whitefly *Bemisia tabaci* B biotype. *Insect Sci.* **18**, 57–66 (2011).
20. Chiel, E. et al. Assessments of fitness effects by the facultative symbiont *rickettsia* in the sweetpotato whitefly (Hemiptera: Aleyrodidae). *Ann. Entomol. Soc. Am.* **102**, 413–418 (2009).
21. Kontsedalov, S. et al. The presence of *Rickettsia* is associated with increased susceptibility of *Bemisia tabaci* (Homoptera: Aleyrodidae) to insecticides. *Pest Manag. Sci.* **64**, 789–792 (2008).
22. Gillespie, J. J. et al. Plasmids and rickettsial evolution: Insight from *Rickettsia felis*. *PLoS One* **2**, e266 (2007).
23. Schrrallhammer, M. et al. ‘*Candidatus Megaira polyxenophila*’ gen. nov., sp. nov.: Considerations on Evolutionary History, Host Range and Shift of Early Divergent *Rickettsia*. *PLoS One* **8**, e27581 (2013).
24. Lanzoni, O. et al. Diversity and environmental distribution of the cosmopolitan endosymbiont “*Candidatus Megaira*”. *Sci. Rep.* **9**, 1179 (2019).
25. Kikuchi, Y. & Fukatsu, T. *Rickettsia* Infection in Natural Leech Populations. *Microb. Ecol.* **49**, 265–271 (2005).
26. Thongprem, P., Evison, S. E. F., Hurst, G. D. D. & Otti, O. Transmission, tropism, and biological impacts of *torix rickettsia* in the common bed bug *Cimex lectularius* (Hemiptera: Cimicidae). *Front. Microbiol.* **11**, (2020).
27. Aguin-Pombo, D., Rodrigues, M. C. P. A., Voetdijk, B. & Breeuwer, J. A. J. Parthenogenesis and Sex-Ratio Distorting Bacteria in *Empoasca* (Hemiptera: Cicadellidae) Leafhoppers. *Ann. Entomol. Soc. Am.* **114**, 738–749 (2021).
28. Kang, Y.-J. et al. Extensive diversity of *Rickettsiales* bacteria in two species of ticks from China and the evolution of the *Rickettsiales*. *BMC Evol. Biol.* **14**, 167 (2014).
29. Driscoll, T., Gillespie, J. J., Nordberg, E. K., Azad, A. F. & Sobral, B. W. Bacterial DNA Sifted from the *Trichoplax adhaerens* (Animalia: Placozoa) Genome Project Reveals a Putative *Rickettsial* Endosymbiont. *Genome Biol. Evol.* **5**, 621–645 (2013).
30. 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).
31. 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).
32. Sabaneyeva, E. et al. Host and symbiont intraspecific variability: The case of *Paramecium calkinsi* and “*Candidatus Trichorickettsia mobilis*”. *Eur. J. Protistol.* **62**, 79–94 (2018).
33. Vannini, C. et al. Flagellar Movement in Two Bacteria of the Family *Rickettsiaceae*: A Re-Evaluation of Motility in an Evolutionary Perspective. *PLoS One* **9**, e87718 (2014).
34. Perlman, S. J., Hunter, M. S. & Zchori-Fein, E. The emerging diversity of *Rickettsia*. *Proc. R. Soc. B Biol. Sci.* **273**, 2097–2106 (2006).
35. Gillespie, J. J. et al. Genomic Diversification in Strains of *Rickettsia felis* Isolated from Different Arthropods. *Genome Biol. Evol.* **7**, 35–56 (2015).
36. Gillespie, J. J. et al. A *Rickettsia* genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. *J. Bacteriol.* **194**, 376–394 (2012).
37. Pukall, R., Tschäpe, H. & Smalla, K. Monitoring the spread of broad host and narrow host range plasmids in soil microcosms. *FEMS Microbiol. Ecol.* **20**, 53–66 (1996).
38. Yan, P. et al. Microbial diversity in the tick *Argas japonicus* (Acari: Argasidae) with a focus on *Rickettsia* pathogens. *Med. Vet. Entomol.* **33**, 327–335 (2019).
39. Dally, M. et al. Cellular localization of two *rickettsia* symbionts in the digestive system and within the ovaries of the mirid bug, macrolophous *pygmaeus*. *Insects* **11**, 530 (2020).
40. Fuxelius, H.-H., Darby, A., Min, C.-K., Cho, N.-H. & Andersson, S. G. E. The genomic and metabolic diversity of *Rickettsia*. *Res. Microbiol.* **158**, 745–753 (2007).
41. Comandatore, F. et al. Supergroup C *Wolbachia*, mutualist symbionts of filarial nematodes, have a distinct genome structure. *Open Biol.* **5**, 150099 (2015).
42. Hagen, R., Verhoeve, V. I., Gillespie, J. J. & Driscoll, T. P. Conjugative transposons and their Cargo genes vary across natural populations of *Rickettsia buchneri* Infecting the Tick *Ixodes scapularis*. *Genome Biol. Evol.* **10**, 3218–3229 (2018).
43. Gillespie, J. J. et al. A tangled web: Origins of reproductive parasitism. *Genome Biol. Evol.* **10**, 2292–2309 (2018).
44. Mediannikov, O., Audoly, G., Diatta, G., Trape, J.-F. & Raoult, D. New *Rickettsia* sp. in tsetse flies from Senegal. *Comp. Immunol. Microbiol. Infect. Dis.* **35**, 145–150 (2012).
45. Pilgrim, J. et al. *Torix* group *Rickettsia* are widespread in *Culicoides* biting midges (Diptera: Ceratopogonidae), reach high frequency and carry unique genomic features. *Environ. Microbiol.* **19**, 4238–4255 (2017).
46. Kuchler, S. M., Kehl, S. & Dettner, K. Characterization and localization of *Rickettsia* sp. in water beetles of genus *Deronectes* (Coleoptera: Dytiscidae). *FEMS Microbiol. Ecol.* **68**, 201–211 (2009).
47. Zchori-Fein, E., Borad, C. & Harari, A. R. Oogenesis in the date stone beetle, *Coccotrypes dactyliperda*, depends on symbiotic bacteria. *Physiol. Entomol.* **31**, 164–169 (2006).
48. Boyd, B. M. et al. Two bacterial genera, *Sodalis* and *Rickettsia*, associated with the seal louse *Proechinophthirus fluctus* (Phthiraptera: Anoplura). *Appl. Environ. Microbiol.* **82**, 3185–3197 (2016).
49. Guillotte, M. L. et al. Lipid A structural divergence in *Rickettsia* pathogens. *mSphere* **6**, e00184–21 (2021).
50. Tvedte, E. S. et al. Genome of the Parasitoid Wasp *Diachasma alloenum*, an emerging model for ecological speciation and transitions to asexual reproduction. *Genome Biol. Evol.* **11**, 2767–2773 (2019).
51. Hotopp, J. C. D. et al. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**, 1753–1756 (2007).
52. Kawafune, K. et al. Two Different *Rickettsial* Bacteria Invading *Volvox carterii*. *PLoS One* **10**, e0116192 (2015).
53. Murray, G. G. R., Weinert, L. A., Rhule, E. L. & Welch, J. J. The phylogeny of *rickettsia* using different evolutionary signatures: How Tree-Like is Bacterial Evolution? *Syst. Biol.* **65**, 265–279 (2016).
54. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
55. Rodriguez-R, L. M., Jain, C., Conrad, R. E., Aluru, S. & Konstantinidis, K. T. Reply to: “Re-evaluating the evidence for a universal genetic boundary among microbial species”. *Nat. Commun.* **12**, 4060 (2021).
56. Reed, J. W. & Walker, G. C. The *exoD* gene of *Rhizobium meliloti* encodes a novel function needed for alfalfa nodule invasion. *J. Bacteriol.* **173**, 664–677 (1991).
57. Konstantinidis, K. T., Rosselló-Móra, R. & Amann, R. Uncultivated microbes in need of their own taxonomy. *ISME J.* **11**, 2399–2406 (2017).
58. Christodoulou, D. et al. Reserve Flux Capacity in the Pentose Phosphate Pathway Enables *Escherichia coli*’s Rapid Response to Oxidative Stress. *Cell Syst.* **6**, 569–578.e7 (2018).
59. Hawkins, J. P., Ordóñez, P. A. & Oresnik, I. J. Characterization of mutations that affect the nonoxidative pentose phosphate pathway in *Sinorhizobium meliloti*. *J. Bacteriol.* **200**, e00436–17 (2018).
60. Driscoll, T. P. et al. Wholly *Rickettsia*! Reconstructed metabolic profile of the quintessential bacterial parasite of eukaryotic cells. *mBio* **8**, e00859–17 (2017).
61. Douglas, A. E. The B vitamin nutrition of insects: the contributions of diet, microbiome and horizontally acquired genes. *Curr. Opin. Insect Sci.* **23**, 65–69 (2017).
62. Klimaszewski, J. et al. Molecular and microscopic analysis of the gut contents of abundant rove beetle species (Coleoptera, Staphylinidae) in the boreal balsam fir forest of Quebec, Canada. *ZooKeys* **353**, 1–24 (2013).
63. Blow, F. et al. B-vitamin nutrition in the pea aphid-*Buchnera* symbiosis. *J. Insect Physiol.* **126**, 104092 (2020).
64. van Ham, R. C. H. J. et al. Reductive genome evolution in *Buchnera aphidicola*. *Proc. Natl Acad. Sci.* **100**, 581–586 (2003).
65. Manzano-Marín, A. et al. Serial horizontal transfer of vitamin-biosynthetic genes enables the establishment of new nutritional symbionts in aphids’ di-symbiotic systems. *ISME J.* **14**, 259–273 (2020).
66. van der Beek, S. L. et al. Streptococcal dTDP-L-rhamnose biosynthesis enzymes: functional characterization and lead compound identification. *Mol. Microbiol.* **111**, 951–964 (2019).
67. Jiang, N., Dillon, F. M., Silva, A., Gomez-Cano, L. & Grotewold, E. Rhamnose in plants - from biosynthesis to diverse functions. *Plant Sci.* **302**, 110687 (2021).
68. Feng, L., Shou, Q. & Butcher, R. A. Identification of a dTDP-rhamnose biosynthetic pathway that oscillates with the molting cycle in *Caenorhabditis elegans*. *Biochem. J.* **473**, 1507–1521 (2016).
69. Daniels, R., Vanderleyden, J. & Michiels, J. Quorum sensing and swarming migration in bacteria. *FEMS Microbiol. Rev.* **28**, 261–289 (2004).
70. Jofré, E., Lagares, A. & Mori, G. Disruption of dTDP-rhamnose biosynthesis modifies lipopolysaccharide core, exopolysaccharide production, and root colonization in *Azospirillum brasilense*. *FEMS Microbiol. Lett.* **231**, 267–275 (2004).
71. Aravind, L., Zhang, D., de Souza, R. F., Anand, S. & Iyer, L. M. The Natural History of ADP-Ribosyltransferases and the ADP-Ribosylation System. In

- Endogenous ADP-Ribosylation* (ed. Koch-Nolte, F.) 3–32 (Springer International Publishing, 2015). [https://doi.org/10.1007/82\\_2014\\_414](https://doi.org/10.1007/82_2014_414).
72. Poltronieri, P. & Ćereković, N. Roles of Nicotinamide Adenine Dinucleotide (NAD<sup>+</sup>) in Biological Systems. *Challenges* **9**, 3 (2018).
  73. Mediannikov, O. et al. High quality draft genome sequence and description of *Occidentia massiliensis* gen. nov., sp. nov., a new member of the family Rickettsiaceae. *Stand. Genom. Sci.* **9**, 9 (2014).
  74. Tamura, A., Ohashi, N., Urakami, H. & Miyamura, S. Classification of *Rickettsia tsutsugamushi* in a New Genus, *Orientia* gen. nov., as *Orientia tsutsugamushi* comb. nov. *Int. J. Syst. Evol. Microbiol.* **45**, 589–591 (1995).
  75. Davison, H. R. *VibrantStarling/Code-used-to-extract-bacterial-genomes-from-invertebrate-genomes: SRA-dive v1.0.0.* (Zenodo, 2022). <https://doi.org/10.5281/zenodo.6396821>.
  76. Doudoumis, V. et al. Challenging the Wigglesworthia, *Sodalis*, *Wolbachia* symbiosis dogma in tsetse flies: *Spiroplasma* is present in both laboratory and natural populations. *Sci. Rep.* **7**, 4699 (2017).
  77. Blow, F. Variation in the structure and function of invertebrate-associated bacterial communities. (University of Liverpool, 2017). <https://doi.org/10.17638/03009325>.
  78. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
  79. Stouthamer, C. M., Kelly, S. & Hunter, M. S. Enrichment of low-density symbiont DNA from minute insects. *J. Microbiol. Methods* **151**, 16–19 (2018).
  80. 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).
  81. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
  82. Chen, Y. et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* **7**, gix120 (2018).
  83. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
  84. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *Peer J.* **7**, e7359 (2019).
  85. Bushnell, B. *BBMap*. <http://sourceforge.net/projects/bbmap/>.
  86. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
  87. Walker, B. J. et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
  88. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
  89. Blin, K. et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
  90. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).
  91. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
  92. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
  93. Joshi, N. A. & Fass, J. N. *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.* (2011). <https://github.com/najoshi/sickle>.
  94. Weisenfeld, N. I. et al. Comprehensive variation discovery in single human genomes. *Nat. Genet.* **46**, 1350–1355 (2014).
  95. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
  96. 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**, (2013).
  97. Nurk, S. et al. Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. In *Research in Computational Molecular Biology* (eds. Deng, M., Jiang, R., Sun, F. & Zhang, X.) 158–170 (Springer, 2013). [https://doi.org/10.1007/978-3-642-37195-0\\_13](https://doi.org/10.1007/978-3-642-37195-0_13).
  98. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
  99. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
  100. Camacho, C. et al. BLAST+: Architecture and applications. *BMC Bioinforma.* **10**, 421 (2009).
  101. Eren, A. M. et al. Community-led, integrated, reproducible multi-omics with anvio. *Nat. Microbiol.* **6**, 3–6 (2021).
  102. Galperin, M. Y. et al. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).
  103. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
  104. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
  105. Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G. & Toth, I. K. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods* **8**, 12–24 (2015).
  106. Rodríguez-R, L. M. & Konstantinidis, K. T. *The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes.* <https://peerj.com/preprints/1900> (2016) <https://doi.org/10.7287/peerj.preprints.1900v1>.
  107. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proc. Int. AAAI Conf. Web Soc. Media* **3**, 361–362 (2009).
  108. Inkscape Project. *Inkscape*. (2020). <https://inkscape.org>.
  109. Kolde R. *heatmap: Pretty Heatmaps.* (2019). <https://cran.r-project.org/web/packages/heatmap/index.html>.
  110. R: *A Language and Environment for Statistical Computing.* (2021). <https://www.R-project.org/>.
  111. Gilchrist, C. L. M. & Chooi, Y.-H. clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).
  112. Tatusova, T. et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624 (2016).
  113. Flissi, A. et al. Norine: Update of the nonribosomal peptide resource. *Nucleic Acids Res.* **48**, D465–D469 (2020).
  114. Krassowski, M. *krassowski/complex-upset: v0.7.4.* (Zenodo, 2020). <https://doi.org/10.5281/zenodo.4308552>.
  115. Siozios, S. *SioStef/panplots:* (Zenodo, 2022). <https://doi.org/10.5281/zenodo.6408803>.
  116. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis.* (Springer-Verlag New York, 2016). <https://doi.org/10.1007/978-0-387-98141-3>.
  117. 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).
  118. 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).
  119. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
  120. Edgar, R. C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinforma.* **5**, 113 (2004).
  121. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2006).
  122. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).

## Acknowledgements

Grant supporting these works: NE/L002450/1 NERC ACCE Doctoral Training Programme, HRD. Ghent University 01P03420 BOF post-doctoral fellowship and 1513719 N Research Foundation - Flanders (FWO) Research Grant, NW. Funding for tsetse fly genomics were to ACD IP BBSRC projects BB/J017698/1 and BB/K501773/1 FB, the materials from which were provided by Philippe Solano (Institut de Recherche pour le Développement, Montpellier, France) and Jean-Baptiste Rayaïsse Centre International de Recherche-Développement sur l’Élevage en zone Subhumide (CIRDES), Bobo Dioulasso, Burkina Faso. Jean-Baptiste died a few years ago but he was a fantastic person to work with and a great field entomologist. We also wish to thank Dr David Montagnes for teaching skills associated with algal culture.

We wish to thank Dr Débora Pires Paula (Embrapa) for granting permission to use SRA data for sample number SRR5651504, Iridian Genomes for allowing use of their SRA data, and the Microbial Culture Collection at the National Institute for Environmental Studies, Japan for use of the sample *Carteria cerasiformis* NIES-425.

## Author contributions

Project concept: H.R.D., S.S., Jack Pilgrim, and G.H. Manuscript written by H.R.D., S.S., J.P., and G.H. All authors commented on the manuscript during development and approved the final version. S.R.A. dive and metagenome assembly carried out by H.R.D. with aid from S.S. Assembly of genome from S.R.A., pangenomics and phylogenomics carried out by H.R.D. with advice from S.S., G.H. Metabolic analysis carried out by H.R.D., Jack Pilgrim and S.S. Sequencing and assembly of bacteria from *Cimex lectularius* and *Culicoides impunctatus* genomes by S.S. and Jack Pilgrim. Sequencing and assembly of symbionts from *Carteria* by S.H.B. and S.S. supervised by P.C. and G.H. Sequencing and construction of RiTSETSE conducted by F.B. as part of thesis work supervised by A.D. J. Parker and S.P. collected and sequenced staphylinid genomes that were released

through NCBI by Iridian Genomes. N.W. collected and sequenced the Bryobia Moomin strain and performed preliminary metagenomic analyses.

### 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/s41467-022-30385-6>.

**Correspondence** and requests for materials should be addressed to Stefanos Siozios.

**Peer review information** *Nature Communications* thanks Joseph Gillespie and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission 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 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