Hidden from plain sight: Novel *Chlamydiota* diversity emerging from screening genomic and metagenomic data.

Authors

Helen R. Davison and Gregory D.D. Hurst

Affiliations

Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Crown Street Liverpool L69 7ZB UK

Corresponding author

Helen R. Davison – davstaff@liverpool.ac.uk

Keywords

Chlamydiota, Symbiotic bacteria, Phylogenomics, Rhabdochlamydiaceae, Simkaniaceae

Abbreviations

MAG – Metagenome Assembled Genome

ANI – Average Nucleotide Identity

AAI – Average Amino-acid Identity

Acknowledgements

Conceptualisation, supervision, methodology and funding acquisition – HRD and GDDH

Investigation –HRD

Analysis, Visualisation, and Data Curation – HRD

Original Draft Preparation – HRD and GDDH

Additional thanks to Stefanos Siozios for guidance early on in this project.

Funding

This work was supported by Natural Environment Research Council (NERC) Adapting to the Challenges of a Changing Environment (ACCE) Doctoral Training Studentship NE/L002450/1

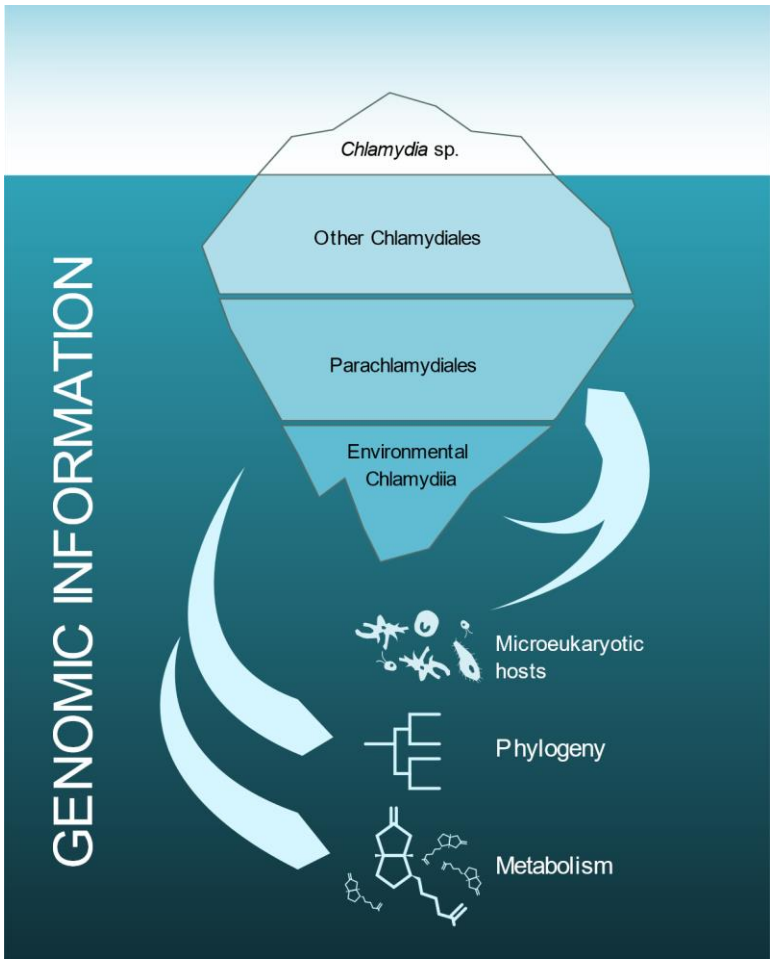
Nucleotide sequence data

Bioproject accession PRJNA901896

Abstract

Chlamydiota are an ancient and hyperdiverse order of obligate intracellular bacteria. The best characterized representatives are pathogens or parasites of mammals, but it is thought that their most common hosts are microeukaryotes like Amoebozoa. The diversity in taxonomy, evolution, and function of non-pathogenic *Chlamydiota* are slowly being described. Here we use data mining techniques and genomic analysis to extend our current knowledge of *Chlamydiota* diversity and its hosts, in particular the family *Parachlamydiales*. We extract one *Rhabdochlamydiaceae* and three *Simkaniaceae* genomes from NCBI Short Read Archive deposits of ciliate and algal genome sequencing projects. We then use these to identify a further 14 and 8 genomes respectively amongst existing, unidentified environmental assemblies. From these data we identify two novel clades with host associated data, for which we propose the names '*Candidatus* Sacchlamydia' (Family *Rhabdochlamydiaceae)* and '*Candidatus* Amphrikania' (Family *Simkaniaceae*), as well as a third new clade of environmental MAGs '*Candidatus* Acheromydia' (Family *Rhabdochlamydiaceae*). The extent of uncharacterized diversity within the *Rhabdochlamydiaceae* and *Simkaniaceae* is indicated by 16 of the 22 MAGs being evolutionarily distant from currently characterised genera. Within our limited data, we observe great predicted diversity in *Parachlamydiales* metabolism and evolution, including the potential for metabolic and defensive symbioses as well as pathogenicity. These data provide an imperative to link genomic diversity in metagenomics data to their associated eukaryotic host, and to develop onward understanding of the functional significance of symbiosis with this hyperdiverse clade.

Graphical Abstract



52 ## Introduction

53 *Chlamydiota* are a phylum of bacteria that are highly diverged from any other microbial clade,
54 such that they were for some time considered a separate kingdom of life (Prowazek, 1912;
55 Horn, 2008). Species within the group uniquely share a biphasic lifestyle, with an inert non-
56 dividing environmental stage and a reproductive intracellular stage (Kahane *et al.*, 1993;
57 Bastidas *et al.*, 2013; König *et al.*, 2017). These bacteria also share proteins that are associated
58 with chloroplast function in plants, which has been speculated to suggest a long-standing
59 relationship with cyanobacteria and algae (Brinkman *et al.*, 2002; Horn *et al.*, 2004). It has
60 also been proposed that ancient members of *Chlamydiota* facilitated the establishment of the
61 first photosynthetic proto-algal eukaryotes (Huang and Gogarten, 2007).

62 *Chlamydiota* are most well known as mammalian pathogens and were first described by the
63 ancient Chinese and Egyptians (Gruber, Lipozenčić and Kehler, 2015). However, the status of
64 the clade has been revised since the advent of DNA taxonomy and it is now recognised that
65 their host range is much broader. They are commonly found in free living amoeba species and
66 environmental samples, though here the original host is often not known (Horn, 2008).
67 Besides amoebae, they can also infect mammals, fish, worms, and a variety of arthropods
68 (Draghi *et al.*, 2004; Horn, 2008; Kjeldsen *et al.*, 2010; Halter *et al.*, 2022). There are potentially
69 many hundreds of unexplored families within *Chlamydiota* (Horn, 2008; Lagkouvardos *et al.*,
70 2014; König *et al.*, 2017). One order of *Chlamydiota*, the *Parachlamydiales*, are widely
71 observed and sequenced in environmental samples, but have little to no functional
72 information and recognised taxonomic affiliations are currently fluid (Horn, 2008; Nylund *et*
73 *al.*, 2018; Halter *et al.*, 2022).

74 Very recently, *Parachlamydia acanthamoebae* has been shown to protect its amoeba host
75 from giant viruses (Arthofer *et al.*, 2022). The mechanism for this defensive symbiosis is not
76 yet known, but *P. acanthamoebae* seems to prevent the virus from co-opting the amoeba
77 machinery, thus preventing the formation of viral factories. Additionally, *Neochlamydia* has
78 previously been implicated in defensive symbioses against *Legionella pneumophila* (Ishida *et*
79 *al.*, 2014; Maita *et al.*, 2018). These data indicate the potential these symbioses have as
80 hidden players in the biology, ecology, and evolution of microeukaryotes beyond a role as
81 pathogens. Moreover, this strong impact on host biology creates an imperative to investigate
82 symbioses in microeukaryotes outside of amoebae.

83 Sequence data provides an opportunity to estimate the level of uncharacterised diversity in
84 microbial symbionts. Two approaches are possible. First, genome sequencing projects for
85 eukaryotes commonly contain the reads from their associated symbionts within them.
86 Second, environmental sequencing projects allow the identification and assembly of draft
87 genomes without host association data. Here, we first screen the NCBI Sequence Read Archive
88 (SRA) databases for ciliates, macro- and micro-algae for the presence of *Chlamydiota* and
89 assemble 1 and 3 new draft genomes for the families, *Rhabdochlamydiaceae* and
90 *Simkaniaceae* respectively (both *Parachlamydiales*). We then use these genomes to identify
91 22 related MAGs from environmental metagenomic studies on online databases. Finally, we
92 conducted a phylogenetic and metabolic analyses of these, and existing genomes, to assess
93 their taxonomic placement and explore the potential impacts on their hosts.

94 # Materials and methods

95 ## Data collection

96 The presence of *Chlamydiota* bacteria was investigated across Algae and Ciliate SRA
97 databases on NCBI as described previously in Davison et al. (2022) using PhyloFlash to identify
98 deposits with evidence of Chlamydiota bacteria 16S rRNA sequence. Those SRA providing
99 evidence of *Chlamydiota* presence were taken forward to metagenomic binning and
100 assembly. We used GTDBtk to screen bins and found that all genomes recovered were related
101 to *Rhabdochlamydiaceae* or *Simkaniaceae*, so analyses from this point on were focused to
102 Parachlamydiales. Genomes assembled here were then used as baits to identify related
103 metagenomically assembled genomes (MAGs) in the NCBI nr database that were previously
104 not fully classified.

105 Existing *Parachalmydiales* genomes, as well as a *Clavichalmydia* and a *Chlamydia* genome
106 were downloaded from NCBI (see Supplementary data 1) alongside the GEM catalogue for
107 genomes identified by Köstlbacher et al. (Köstlbacher *et al.*, 2021; Nayfach *et al.*, 2021).

108 ## Metagenomic assembly and annotation

109 Minimap2, MEGAHIT and MetaBAT2 (Li *et al.*, 2015; Li, 2018; Kang *et al.*, 2019) were used to
110 assemble and bin metagenomic bacterial genomes from SRA data using the pipeline in
111 Davison (2022). Genome quality and identities were checked with CheckM and NCBI Blastn
112 (Camacho *et al.*, 2009; Parks *et al.*, 2015).

113 Four newly assembled genomes, 18 environmental MAGs, 82 previously established
114 *Parachlamydiaceae*, and 2 outgroup *Chlamydiaceae* were then passed to Anvio-7 to be

115 annotated with COG20 and KEGG kofams through their pangenomics pipeline (Aramaki *et al.*,
116 2020; Eren *et al.*, 2021; Galperin *et al.*, 2021; Davison, 2022).

117 Phlyogenomics and metabolic predictions

118 Clusters of protein homologs shared by 90% of the 106 genomes used were identified and
119 extracted through Anvio-7. We extracted 34 single copy core gene clusters that contain a total
120 of 3604 genes. A phylogeny partitioned by gene cluster was constructed with IQTREE, using
121 model finder plus, 1000 ultrafast bootstraps, and 1000 SH-alrt replicates. Models per partition
122 can be found in Supplementary data 2.

123 We used GTDBtk for taxonomic classification of the recovered genomes and to identify
124 taxonomic novelty among them. AAI was calculated for all genomes through kostalabs online
125 enveomics suite (Rodriguez-R and Konstantinidis, 2016). ANIb was calculated with pyANI
126 through anvio-7 (Supplementary data 1) (Pritchard *et al.*, 2016). The network was visualised
127 in Gephi 0.9 (Bastian, Heymann and Jacomy, 2009), with edges filtered for >0.6 AAI and >0.95
128 ANI scores. Annotations were added in Inkscape (Inkscape Project, 2020).

129 Anvio-7 using KEGG kofams was used to assess metabolic completeness for high quality
130 genomes exceeding a CheckM score of 90% completeness to minimize the loss of information
131 caused by incompleteness (Eisenhofer, Odriozola and Alberdi, 2023). Heatmaps were
132 constructed with Seaborn 0.12.1 in Python 3.10 and annotated with Inkscape (Rossum and
133 Drake, 2009; Waskom and Seaborn development team, 2020). Toxin-antitoxin systems were
134 identified with antiSMASH (Blin *et al.*, 2021). CrisprCAS finder (Couvin *et al.*, 2018) was used
135 to establish the presence of cas systems in all extracted genomes.

136 # Results

137 ## Genomes and Phylogeny

138 Four genomes were assembled from SRA data and an additional 22 recovered from the NCBI
139 GenBank non-redundant database (Table 1). Of these 22 genomes, 16 came from
140 environmental metagenome sequencing efforts that did not specifically target *Chlamydiota*,
141 and the six that were targeted came from a single paper. Two were Identified to species and
142 three to family level (Supplementary data 1). One of these, JAGXTH01, is a
143 *Rhabdochlamydiaceae* misidentified as a *Simkania* sp. (see Figure 1 and accession
144 GCA_018333475.1).

145 Several groups, like SM23-39 and most MCF types, are sufficiently different and diverse to
146 warrant their own family name. In literature, SM23-39 has been referred to as both '*Ca*.
147 Limichlamydiaceae' and '*Ca*. Anoxychlamydiaceae' (Pillonel, Bertelli and Greub, 2018;
148 Dharamshi *et al.*, 2020, 2023). Their classification seems to be mixed, with genomes identified
149 as SM23-39 also spanning a second proposed family '*Ca.* Enkichlamydiaceae' (Supplementary
150 data 1 and Köstlbacher *et al.*, (2021). Based on GTDBtk analysis we have opted for using the
151 clade name SM23-39 to refer to this clade to avoid naming confusion, and because all
152 genomes share family level classification by this method, as well as equivalent diversity with
153 *Rhabdochlamydiaceae* and *Simkaniaceae* (Figure 1 and 2). There is no set threshold for
154 classifying family or order level, and although GTDBtk does overclassify higher taxons
155 (Chaumeil *et al.*, 2020), it is currently the only objective way of designating family level
156 classification. Higher classifications can be added later as needed, but at this stage we do not
157 believe it is helpful to establish additional families or orders while the current within family
158 phylogeny is so fluid.

159 AAI and ANI results shown in Figure 2 support most existing genera and species clusters, but
160 also indicates the potential need for some genera to be split (e.g., *Simkania*) and some species
161 to be taxonomically combined ('*Ca*. Rhabdochlamydia oedothora' and '*Ca*. Rhabdochlamydia'
162 sp. W815 RHOW815).

163 Using AAI values with a cut off of 65%, we find objective support for naming the clade
164 containing MGLO01 as its own genus, '*Ca*. Arenachlamydia', as suggested in (Pillonel, Bertelli
165 and Greub, 2018). However, contrary to Pillonel et al. (2018), we do not find that they are
166 unique within *Chlamydiota* in lacking a Menaquinone biosynthesis/futalosine pathway
167 (Supplementary data 1). Amongst the genomes included in this study, all lack the futalosine
168 pathway (KEGG module M00930 and M00931) except for MCF-G and '*Ca.* Sorochlamydiaceae'
169 (clade CC-IV). The Menaquinone biosynthesis pathway (M00116) is partially present in all
170 except 13 genomes, of which span several different clades across *Parachlamydiales*
171 (Supplementary data 1). Other clades form putative genera groups of four or more genomes
172 that are simply unnamed. *Simkaniaceae* and *Rhabdochlamydiaceae* in particular lack solid
173 categorisation, each having only two previously named genera (Figure 1 and 2). Further, the
174 genomic diversity displayed within both families is extensive. Some single genomes possess
175 more than ten times more unique genes than genes shared across all genomes (Figure 3),
176 though this could be an artifact of divergent orthologs not clustering together. It is very likely
177 that *Parachlamydiales* will need to be split into multiple orders as more data becomes
178 available, and revision of family level affiliation, to better describe this ancient clade of
179 bacteria.

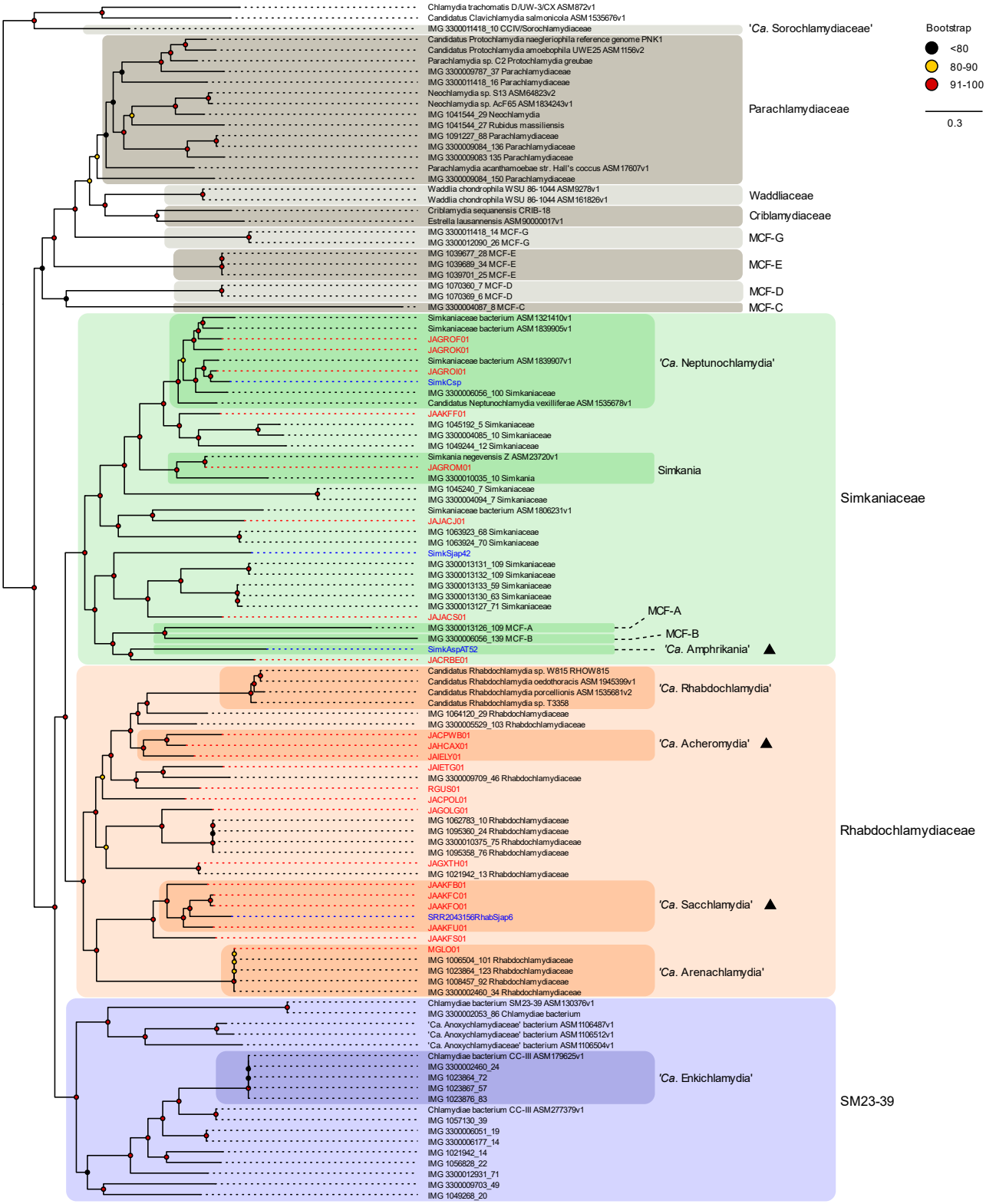


Figure 1. Genome wide phylogeny of *Parachlamydiales*. Maximum likelihood (ML) phylogeny of *Parachlamydiales* constructed from 34 single copy gene clusters that contain a total of 3604 genes. New genomes are indicated by ▲ with blue labels for genomes derived from host SRA bins, and red for MAGS without host assocition data. Bootstrap values based on 1000 replicates are indicated with coloured circles (red = 91-100, yellow = 81-90, black <= 80).

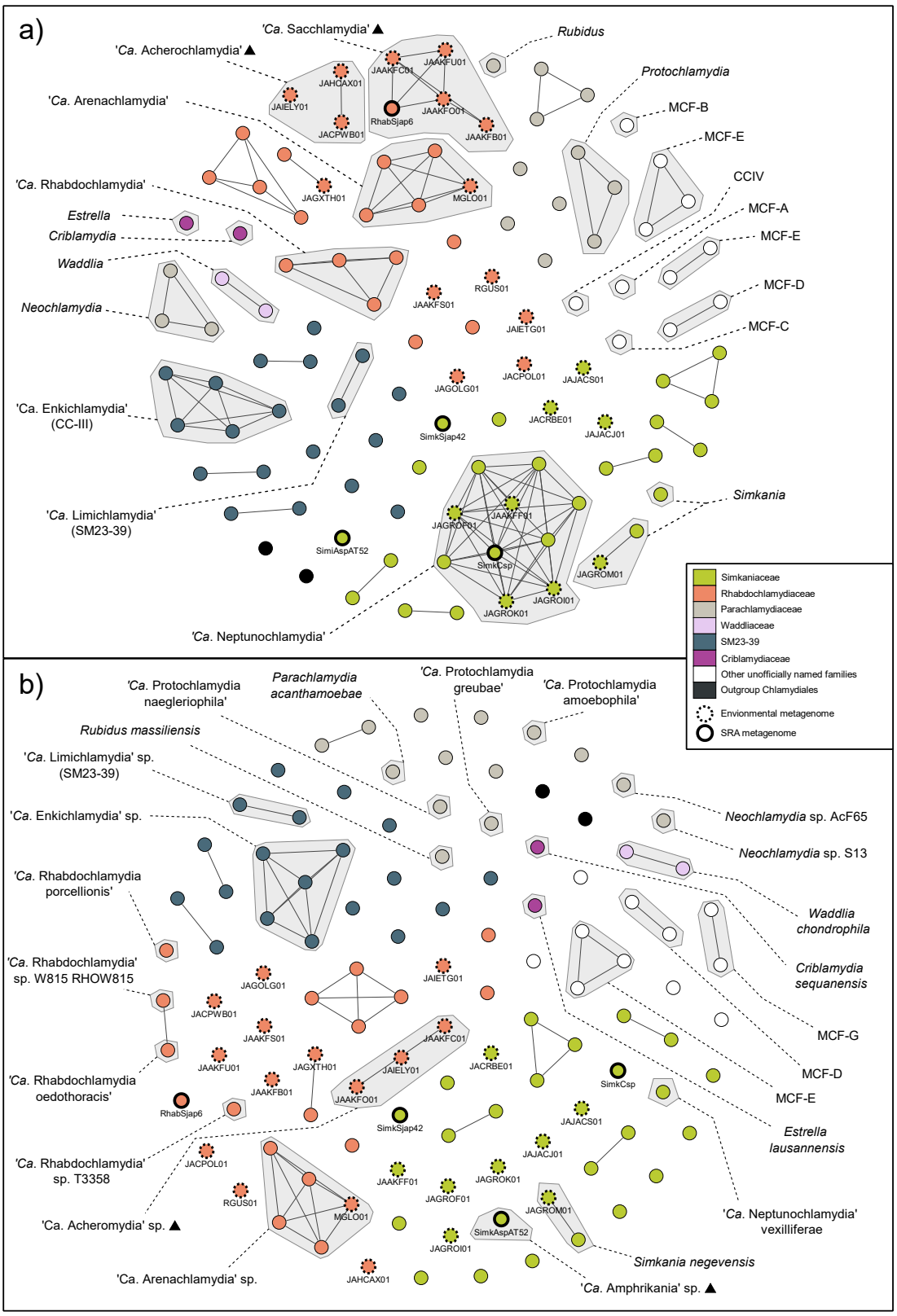


Figure 2. Genus and species level clustering across *Parachlamydiales*. Frutcherman Reingold networks of pairwise a) Average Amino Acid Identity (AAI) with edge weights >65% similarity and b) Average Nucleotide Identity (ANI) with edge weights >95% similarity across all genomes. AAI and ANI illustrate genus and species boundaries, respectively. Proposed genus names are indicated with a ▲.

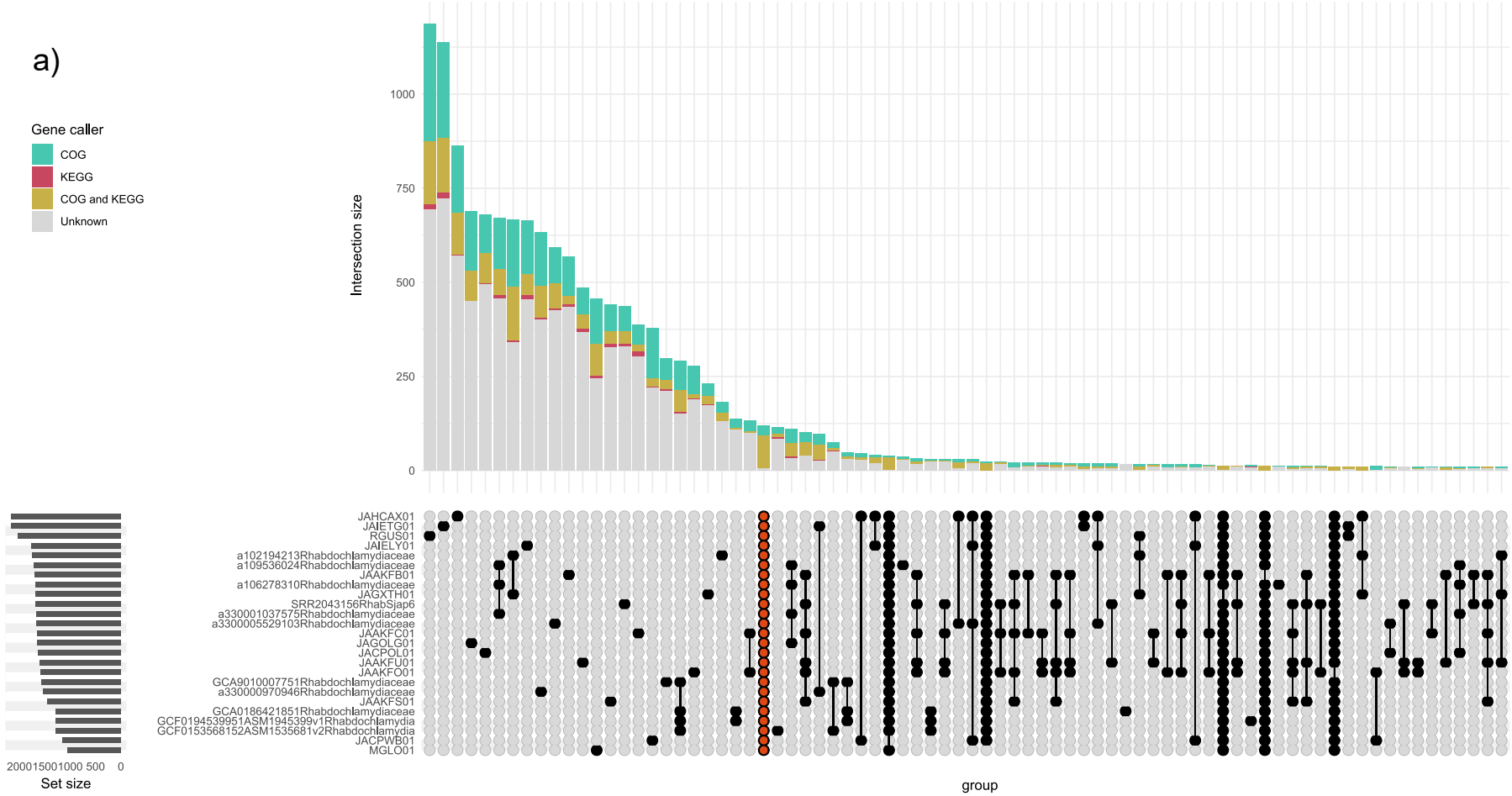


Figure 2a. Gene content comparison for *Rhabdochlamydiaceae* and *Simkaniaceae*. An upsetplot illustrating the disparity of gene presence absence across the families a) *Rhabdochlamydiaceae* and b) *Simkaniaceae*. Vertical bars illustrate the number of genes present shared across compared genomes, indicated by black circles beneath each bar. Orange circles indicate the core genome shared by all genomes in the comparison.

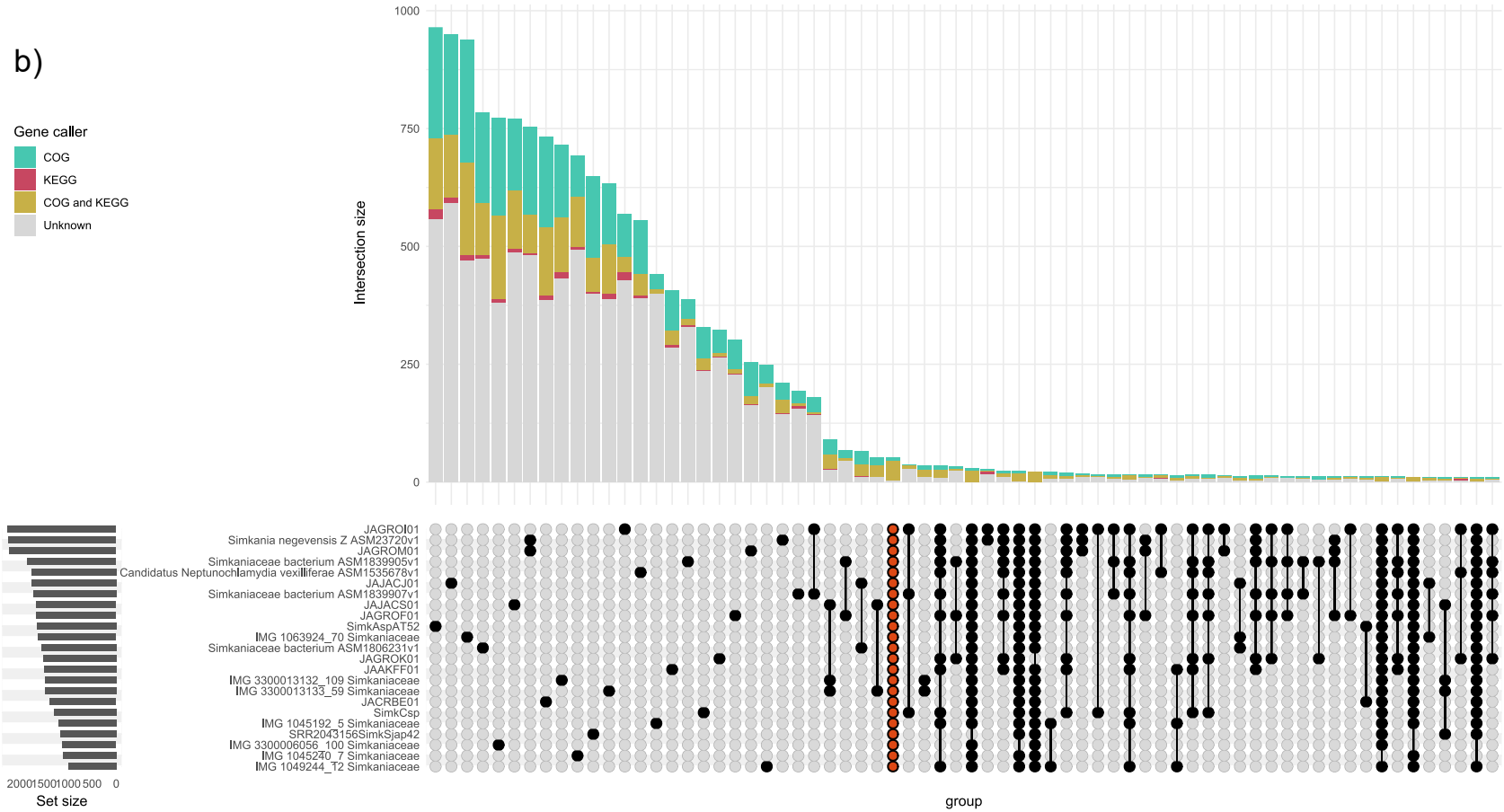


Figure 3b. Gene content comparison for *Rhabdochlamydiaceae* and *Simkaniaceae*. An upset plot illustrating the disparity of gene presence absence across the families a) *Rhabdochlamydiaceae* and b) *Simkaniaceae*. Vertical bars illustrate the number of genes present shared across compared genomes, indicated by black circles beneath each bar. Orange circles indicate the core genome shared by all genomes in the comparison.

202 ## Proposed Taxonomy

203 Based on AAI, ANI, GTDBtk analysis, metabolic results, and phylogeny, we find evidence for at
204 least three new genus-level clusters within Rhabdochlamydiaceae and Simkaniaceae. The
205 genera we propose contain either multiple high-quality environmental genomes found with
206 the methods described in this study, or are high-quality assemblies extracted from potential
207 host SRAs.

208 '*Candidatus* Sacchlamydia' gen. nov. - For the cluster that contains RhabSjap6, a host
209 associated genome, named after the putative host *Saccharina japonica* and the pattern of
210 using "-chlamydia" as the suffix for Chlamydiota genera.

211 '*Candidatus* Amphrikania' gen. nov. - A second new clade with host association data lies
212 within the Simkaniaceae (the genome named SimkAspAT52). Amphri- after Amphritrite
213 Queen of the Sea and echoing its marine host *Amoebophyra*, and -kania after the type genus
214 for the family Simkaniaceae, *Simkania*.

215 Similarly to SimkAspAT52, clade SimkSjap42 is likely a new species group but will not be
216 named because it has genome that is only 65% complete in CheckM analysis (Figure 1., Table
217 1).

218 '*Candidatus* Acheromydia' gen. nov. - For the clade including the MAGs JAHCAX01 and
219 JACPWB01 which share >65% AAI similarity. For convenience, and so as not to over inflate
220 this group with new names, I also include the relatively closely related genome JAIEL01 (>62%
221 AAI similarity). This genus is named after Acheron (Acher-) one of five Greek rivers of the
222 underworld that occasionally surfaces above ground. The MAGs come from drinking water,
223 activated sludge and ground water. -mydia from *Chlamydia*.

224 ## Metabolism

225 Parachlamydiales have a broad array of metabolic pathways (Figure 4 and Supplementary
226 data 1), including nutritional and defensive systems. Of note, several have mostly complete B
227 vitamin synthesis pathways (biotin, riboflavin and thiamine), known to be associated with
228 beneficial contributions to the host in other symbioses with intracellular bacteria.

229 Some Parachlamydiales have uroporphyrinogen decarboxylase and uroporphyrinogen-III
230 synthase which form part of several biosynthesis pathways including heme, chlorophyll,
231 cobalamin, and siroheme. In addition, some have incomplete phylloquinone (vitamin K1)
232 biosynthesis pathways. All Parachlamydiales genomes here also have complete Type III
233 secretion systems and incomplete gene sets associated with flagellar assembly
234 (Supplementary data 1). Several terpenoid synthesis pathways are also present, including
235 dTDP-L-rhamnose biosynthesis.

236 We find that KEGG Kofam identifies CRISPR-associated protein Cas2 in all *Parachlamydiales*
237 studied in this study, including *P. acanthamoeba*. Several have additional types of CRISPR/Cas
238 associated genes (Supplementary data 1). CRISPRcas-finder identifies cas and cas spacers in
239 the four MAGs extracted from the SRA samples (Supplementary data 1) and only RhabSjap6
240 appears to possess both CRISPR elements and a cas cluster. Whether these genes are
241 functional or associated with defensive symbiosis against phages based purely on metabolic
242 prediction is unknown; the class of cas is not predictive of their function.

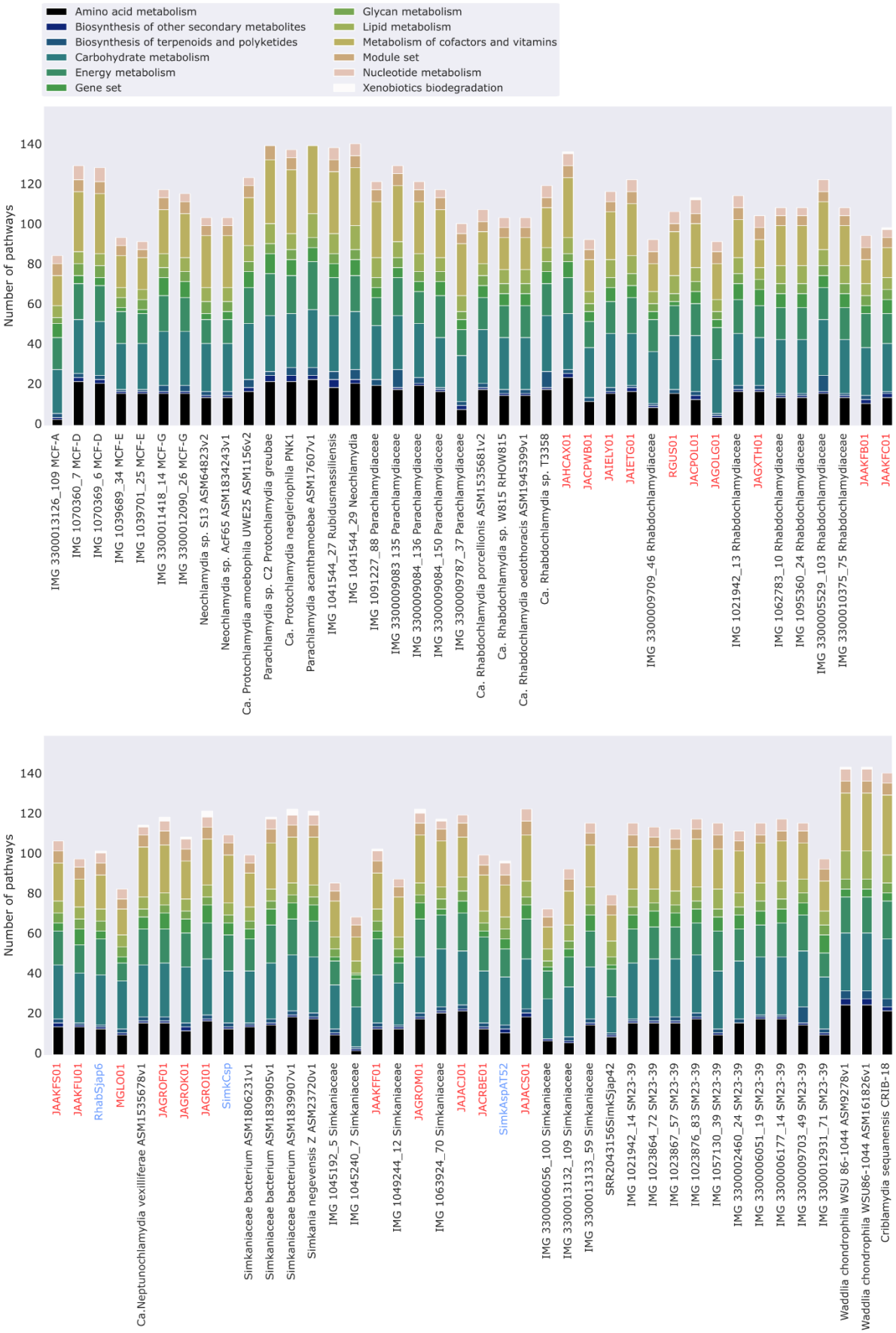


Figure 3. The number of each type of metabolic pathway found in each *Chlamydiota* genome. Red indicates environmental MAGs; blue indicates MAGs assembled in this study.

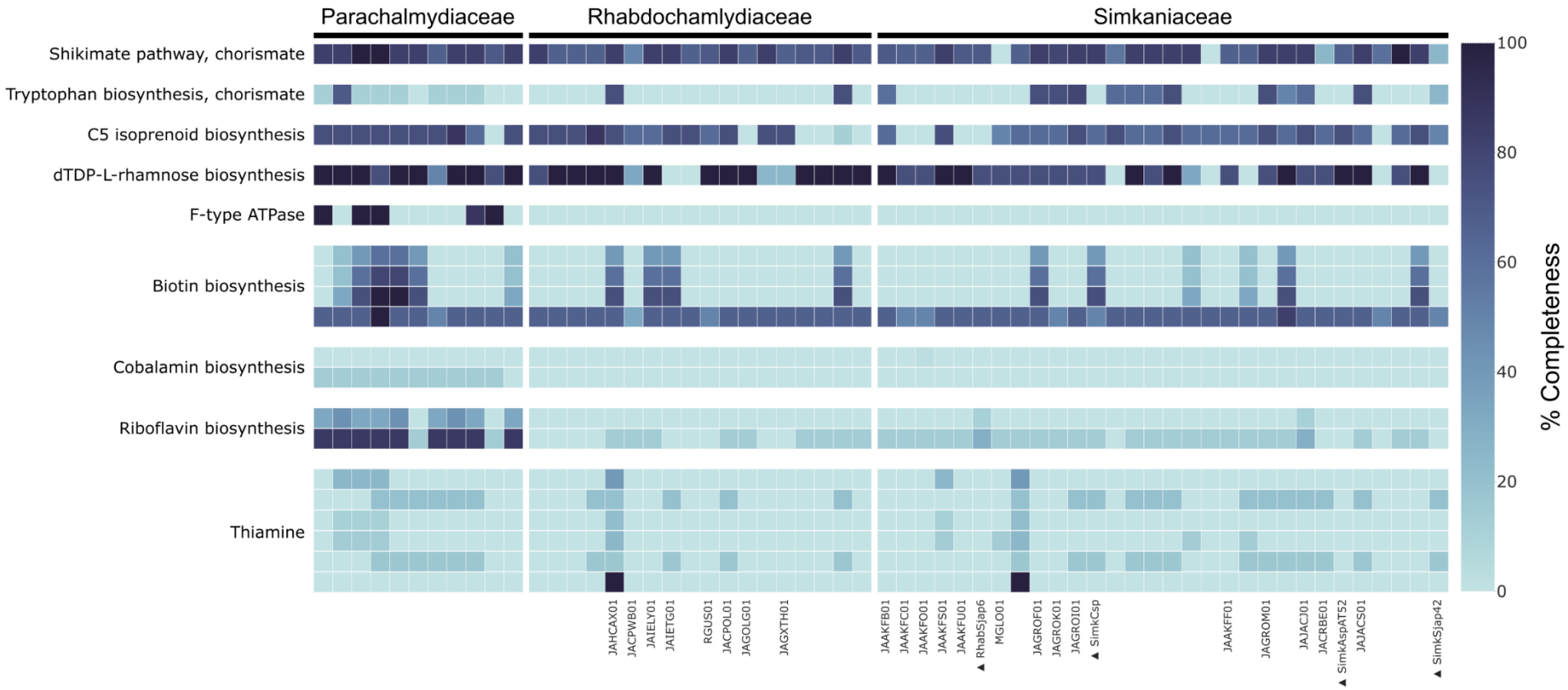


Figure 4. Heatmap of the completeness of KEGG metabolic pathways of interest across the *Parachlamydiaceae, Rhabdochlamydiaceae and Simkaniaceae*. Full metabolic pathway completeness for all genomes and all other pathways are available in Supplementary data 1. New genomes assembled in this study are indicated by ▲.

250 # Discussion

251 ## Genomes and Phylogeny

252 We observe high genomic, phylogenetic, and metabolic diversity in the currently named
253 families, genera, and within species of *Chlamydiota*, in line with other studies (Horn, 2008;
254 Lagkouvardos *et al.*, 2014; König *et al.*, 2017). The current taxonomy is not sufficient and has
255 been left undefined to the detriment of these bacteria, in part due to the nature of the clade
256 being so broad and data being so sparse. This has resulted in some genomes such as *Rubidus*
257 *massiliensis* and JAGXTH01 being miscategorised on NCBI. *Rubidus massiliensis* is categorised
258 as *Chlamydiaceae* rather than *Parachlamydiaceae,* as proposed in the study where it was first
259 described (Bou Khalil *et al.*, 2016) and as shown in Figure 1. MAG JAGXTH01 is identified as a
260 *Simkania* sp., whereas we find it falls within the *Rhabdochlamydiaceae*.

261 Prior to this study, only two Environmental MAGs found had been assigned a taxonomic
262 classification lower than family, and one of those two is misidentified (Supplementary data
263 1). At this stage we do not believe it is helpful to split existing family clades like SM23-39 into
264 multiple families, especially since they are supported by GTDBtk analysis. There is also some
265 confusion in literature for the taxonomy for some clades: '*Ca*. Limichlamydiaceae' (Pillonel,
266 Bertelli and Greub, 2018) is included in '*Ca*. Anoxychlamydiaceae' (Dharamshi *et al.*, 2023),
267 and '*Ca*. Enkichlamydiaceae' (Pillonel, Bertelli and Greub, 2018) == Chlamydiae clade III
268 (Dharamshi *et al.*, 2023).

269 We believe the splitting of *Parachlamydiales* into three new orders are in some respects
270 warranted: 'C*a*. Anoxychlamydiales', 'C*a*. Simkaniales' (which includes *Rhabdochlamydiaceae*
271 and *Simkaniaceae*) and 'C*a*. Amoebachlamydiales' (for *Parachlamydiaceae*) (Dharamshi *et*

272 *al.*, 2023). We could also argue that *Rhabdochlamydiaceae* and *Simkaniaceae* will eventually
273 need their own Order classifications based on the high gene content diversity within the
274 families (figures 3a and 3b). However, we do not believe it is helpful to split *Parachlamydiales*
275 into further orders at this time while most families and genera have not been properly
276 defined. Additionally, Order classification has no robust criteria and the erection of further
277 Orders may contribute to confusion around the taxonomy of *Chlamydiota*.

278 ## Metabolism

279 Vitamin B biosynthesis is known to be associated with beneficial contributions to the host in
280 other symbioses with intracellular bacteria. However, all characterised Parachlamydiales have
281 biphasic lifestyles, and many are known pathogens, so the presence of B vitamin modules
282 does not necessarily indicate potential benefit to the host. For instance, pathogenic
283 *Chlamydia* with bioY genes have previously been observed to uptake biotin from host cells
284 (Fisher *et al.*, 2012).

285 If Parachlamydiales do form beneficial symbioses with their host it is feasibly associated with
286 photosynthesis through vitamin Phylloquinone synthesis or chlorophyll synthesis.
287 Phylloquinone is produced and used by plants, algae, and cyanobacteria in photosynthesis.
288 Most *Parachlamydiales* genomes also have a complete chorismate synthase pathway (Figure
289 5, Supplementary data 1), an important intermediate product for phylloquinone production,
290 as well as intermediates for alkaloids and salicylic acid that are important in plant defence
291 systems (Hamberger *et al.*, 2006; Shanmugabalaji *et al.*, 2022). It should be noted that
292 chorismate can also be used to synthesize a variety of other vital aromatic compounds such
293 as ubiquinone, which is common in bacteria (Dosselaere and Vanderleyden, 2001).

294 There is also a possibility that these bacteria form defensive symbioses, as seen in the case of
295 viral protection of amoebae by *P. acanthamoeba* (Arthofer *et al.*, 2022). Unfortunately, we
296 do not know how *P. acanthamoeba* blocks viral factory formation, so we cannot identify any
297 associated pathways through homology. The presence of CRISPR/cas systems in some
298 genomes could indicate a potential mechanism for this trait, but we do not know that they
299 are functional, and, if they are, they may be associated with the free living stage rather than
300 antiviral activity in the host. CRISPR/Cas Phage defence are common across bacteria but
301 notably rare in microbial symbionts (Deveau, Garneau and Moineau, 2010). They have also
302 previously been found in *Protochlamydia* genomes and other *Chlamydiota* (Deveau, Garneau
303 and Moineau, 2010; Bertelli *et al.*, 2016; Köstlbacher *et al.*, 2021).

304 An alternative route for defensive symbiosis could be terpenoid production. Terpenoids such
305 as dTDP-L-rhamnose are associated with plant-bacteria symbioses as well as pathogenicity
306 (Ma, Pan and McNeil, 2002; Jofré, Lagares and Mori, 2004; French, 2017; Jiang *et al.*, 2021).
307 Terpenoids are also used by algae themselves in defence against threats such as bacteria and
308 heavy metals (French, 2017; Karimi *et al.*, 2019). Some Red algae use bacteria-like pathways
309 to produce their terpenoids (Wei *et al.*, 2019).

310 Conclusions

311 Chlamydiota are hyper diverse and under described. We have added to the known
312 phylogenies of the *Parachlamydiales* families: *Simkaniaceae* and *Rhabdochlamydiaceae*. In
313 addition, we clarify the status of two additional *Rhabdochlamydiaceae* and one *Simkaniaceae*
314 genera, for which we propose the names '*Ca*. Sacchlamydia', '*Ca.* Acherochlamydia', and '*Ca.*
315 Amphrikania'. The metabolic potential of *Chlamydiota* is likewise highly diverse, with several
316 clusters of genes that could be associated with both defensive and nutritional symbiosis or

317 otherwise associated with pathogenicity. Closer study of the metabolic interactions between

318 *Parachlamydiales* and their microeukaryotic hosts are required to elucidate if and how they

319 affect each other.

320 # Tables

321 Table 1. Select genome metadata for newly assembled Rhabdochlamydiaceae and Simkaniaceae genomes and previously assembled NCBI
322 environmental MAGs. Full metadata can be found in Supplementary data 1.

Name	Bacteria accession	Isolate Name	Family	Source accession	Host Type	Source	CheckM completion score	CheckM contamination	Genome size	Number of contigs	GC content %	Completion status
Genomes assembled in this study												
SimiAspAT52	SAMN32091777	-	*Simkaniaceae*	SRR3080743	Algae	*Amoebophrya* sp. AT5.2	91.55	0	1,800,129	71	36.81	contig
SimkCsp1	SAMN32091778	-	*Simkaniaceae*	SRR9841583	Ciliate	*Chilodochona* sp.	90.65	0.23	1,516,029	217	39.52	contig
RhabSjap6	SAMN32091779	-	*Rhabdochlamydiaceae*	SRR2043156	Algae	*Saccharina japonica*	91.05	0.9	1,734,283	4	45.53	contig
SimkSjap42	SAMN32091780	-	*Simkaniaceae*	SRR2043156	Algae	*Saccharina japonica*	65.79	1.73	972,015	316	43.90	contig
Existing environmental MAGs												
JAAKFB01	GCA_011065235.1	*K1000_chlam_2*	*Rhabdochlamydiaceae*	PRJNA504765	Unknown	marine sediment	82.56	2.73	1,738,011	252	44.0	contig
JAAKFC01	GCA_011065225.1	*K1000_chlam_3*	*Rhabdochlamydiaceae*	PRJNA504765	Unknown	marine sediment	78.01	2.03	1,662,393	270	41.5	contig
JAAKFF01	GCA_011065165.1	*K1060_chlam_2*	*Simkaniaceae*	PRJNA504765	Unknown	marine sediment	90.2	0	1,625,346	143	46.5	contig
JAAKFO01	GCA_011064755.1	*K940_chlam_6*	*Rhabdochlamydiaceae*	PRJNA504765	Unknown	marine sediment	82.02	3.6	1,505,120	432	42.5	contig
JAAKFS01	GCA_011064965.1	*KR126_chlam_1*	*Rhabdochlamydiaceae*	PRJNA504765	Unknown	marine sediment	92.91	1.35	1,614,069	25	46.5	contig
JAAKFU01	GCA_011064915.1	*KR126_chlam_3*	*Rhabdochlamydiaceae*	PRJNA504765	Unknown	marine sediment	84.12	2.25	1,677,468	185	42.5	contig
JACPOL01	GCA_016185065.1	*NC_groundwater_600_Pr5_B-0.1um_49_25*	*Rhabdochlamydiaceae*	PRJNA640378	Unknown	groundwater	95.61	0.68	1,865,878	10	48.5	contig

JACPWB01	GCA_016197265.1	*NC_groundwater_800_Pr6_B-0.1um_47_6*	*Rhabdochlamydiaceae*	PRJNA640378	Unknown	groundwater	76.96	5.69	1,071,557	362	46.5	scaffold
JACRBE01	GCA_016213235.1	*NC_groundwater_1612_Pr3_B-0.1um_36_11*	*Simkaniaceae*	PRJNA640378	Unknown	groundwater	82.09	1.04	1,468,519	59	36.0	contig
JAGOLG01	GCA_017994195.1	*Gw_Bodden_bin_68*	*Rhabdochlamydiaceae*	PRJNA524094	Unknown	wastewater	89.02	1.49	1,679,597	130	42.5	contig
JAGROF01	GCA_020440405.1	*HKST-UBA11*	*Simkaniaceae*	PRJNA432264	Unknown	Saline, activated sludge	80.23	0.34	1,731,737	299	44.5	scaffold
JAGROI01	GCA_020440305.1	*HKST-UBA14*	*Simkaniaceae*	PRJNA432264	Unknown	saline, activated sludge	96.28	0.56	2,560,732	197	39.0	scaffold
JAGROK01	GCA_020440285.1	*HKST-UBA16*	*Simkaniaceae*	PRJNA432264	Unknown	Saline, activated sludge	88.85	1.52	1,696,508	180	42.5	contig
JAGROM01	GCA_020440265.1	*HKST-UBA18*	*Simkaniaceae*	PRJNA432264	Unknown	Saline, activated sludge	92.44	5.63	2,391,591	302	41.5	Scaffold
JAGXTH01	GCA_018333475.1	*CROMO_BS_87_3_0*	*Rhabdochlamydiaceae*	PRJNA672823	Unknown	groundwater	84.46	4.05	1,945,191	100	42.0	Contig
JAHCAX01	GCA_019634675.1	*ACE_VER1*	*Rhabdochlamydiaceae*	PRJNA725625	Unknown	activated sludge	96.96	0.68	2,665,416	21	46.5	Contig
JAIELY01	GCA_019752555.1	*co.spades.CONCOCT.1kb_142*	*Rhabdochlamydiaceae*	PRJNA745370	Unknown	drinking water	95.61	0.68	2,067,720	42	44.0	contig
JAIETG01	GCA_019745395.1	*time.spades.CONCOCT.1kb_023*	*Rhabdochlamydiaceae*	PRJNA745370	Unknown	drinking water	93.69	1.35	2,621,400	37	39.5	Contig
JAJACJ01	GCA_022601575.1	*O_S4.1_1*	*Simkaniaceae*	PRJNA742377	Unknown	sponge, *Haliclona oculata*	95.95	0	1,780,284	3	45.0	Scaffold
JAJACS01	GCA_022601395.1	*O_S4.2_37*	*Simkaniaceae*	PRJNA742377	Unknown	sponge, *Haliclona oculata*	94.93	0.68	1,870,253	48	38.5	contig
MGLO01	GCA_001796065.1	-	*Rhabdochlamydiaceae*	PRJNA288027	Unknown	groundwater	84.46	0.17	1,177,368	32	49.0	Scaffold
RGUS01	GCA_010027655.1	*WB9_2_327*	*Rhabdochlamydiaceae*	PRJNA495371	Unknown	Freshwater	91.44	13.91	2,063,700	445	38.0	contig

323

Supplementary material captions

Supplementary data 1 – metadata and metabolic data for genomes used in this study.

Supplementary data 2 – tree partition models for Figure 1.

References

Aramaki, T. *et al.* (2020) 'KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold', *Bioinformatics*. Edited by A. Valencia, 36(7), pp. 2251–2252. Available at: https://doi.org/10.1093/bioinformatics/btz859.

Arthofer, P. *et al.* (2022) 'Defensive symbiosis against giant viruses in amoebae', *Proceedings of the National Academy of Sciences*, 119(36). Available at: https://doi.org/10.1073/pnas.2205856119.

Bastian, M., Heymann, S. and Jacomy, M. (2009) 'Gephi : An Open Source Software for Exploring and Manipulating Networks Visualization and Exploration of Large Graphs', *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1), pp. 361–362.

Bastidas, R.J. *et al.* (2013) 'Chlamydial Intracellular Survival Strategies', *Cold Spring Harbor Perspectives in Medicine*, 3(5), pp. a010256–a010256. Available at: https://doi.org/10.1101/cshperspect.a010256.

Bertelli, C. *et al.* (2016) 'CRISPR System Acquisition and Evolution of an Obligate Intracellular Chlamydia-Related Bacterium', *Genome Biology and Evolution*, 8(8), pp. 2376–2386. Available at: https://doi.org/10.1093/GBE/EVW138.

Blin, K. *et al.* (2021) 'antiSMASH 6.0: improving cluster detection and comparison capabilities', *Nucleic Acids Research*, 49(W1), pp. W29–W35. Available at: https://doi.org/10.1093/nar/gkab335.

Bou Khalil, J.Y. *et al.* (2016) 'Developmental Cycle and Genome Analysis of "Rubidus massiliensis," a New Vermamoeba vermiformis Pathogen', *Frontiers in Cellular and Infection Microbiology*, 6. Available at: https://doi.org/10.3389/fcimb.2016.00031.

Brinkman, F.S.L. *et al.* (2002) 'Evidence That Plant-Like Genes in Chlamydia Species Reflect an Ancestral Relationship between Chlamydiaceae, Cyanobacteria, and the Chloroplast', *Genome Research*, 12(8), pp. 1159–1167. Available at: https://doi.org/10.1101/gr.341802.

Camacho, C. *et al.* (2009) 'BLAST+: Architecture and applications', *BMC Bioinformatics*, 10. Available at: https://doi.org/10.1186/1471-2105-10-421.

Chaumeil, P.-A. *et al.* (2020) 'GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database', *Bioinformatics*, 36(6), pp. 1925–1927. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTZ848.

Couvin, D. *et al.* (2018) 'CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins', *Nucleic Acids Research*, 46(W1), pp. W246–W251. Available at: https://doi.org/10.1093/nar/gky425.

358 Davison, H.R. (2022) 'VibrantStarling/Code-used-to-extract-bacterial-genomes-from-invertebrate-
359 genomes: SRA-dive v1.0.0'. Available at: https://doi.org/10.5281/zenodo.6396821.

360 Deveau, H., Garneau, J.E. and Moineau, S. (2010) 'CRISPR/Cas System and Its Role in Phage-Bacteria
361 Interactions', *Annual Review of Microbiology*, 64(1), pp. 475–493. Available at:
362 https://doi.org/10.1146/annurev.micro.112408.134123.

363 Dharamshi, J.E. *et al.* (2020) 'Marine Sediments Illuminate Chlamydiae Diversity and Evolution',
364 *Current Biology*, 30(6), pp. 1032-1048.e7. Available at: https://doi.org/10.1016/j.cub.2020.02.016.

365 Dharamshi, J.E. *et al.* (2023) 'Gene gain facilitated endosymbiotic evolution of Chlamydiae', *Nature*
366 *Microbiology*, 8(1), pp. 40–54. Available at: https://doi.org/10.1038/s41564-022-01284-9.

367 Dosselaere, F. and Vanderleyden, J. (2001) 'A Metabolic Node in Action: Chorismate-Utilizing
368 Enzymes in Microorganisms', *Critical Reviews in Microbiology*, 27(2), pp. 75–131. Available at:
369 https://doi.org/10.1080/20014091096710.

370 Draghi, A. *et al.* (2004) 'Characterization of " Candidatus Piscichlamydia salmonis" (Order
371 Chlamydiales), a Chlamydia-Like Bacterium Associated With Epitheliocystis in Farmed Atlantic
372 Salmon (Salmo salar)', *Journal of Clinical Microbiology*, 42(11), pp. 5286–5297. Available at:
373 https://doi.org/10.1128/JCM.42.11.5286-5297.2004.

374 Eisenhofer, R., Odriozola, I. and Alberdi, A. (2023) 'Impact of microbial genome completeness on
375 metagenomic functional inference - ISME Communications', *ISME Communications*, 3(1), p. 12.
376 Available at: https://doi.org/10.1038/s43705-023-00221-z.

377 Eren, A.M. *et al.* (2021) 'Community-led, integrated, reproducible multi-omics with anvi'o', *Nature*
378 *Microbiology*, 6(1), pp. 3–6. Available at: https://doi.org/10.1038/s41564-020-00834-3.

379 Fisher, D.J. *et al.* (2012) 'Uptake of Biotin by Chlamydia Spp. through the Use of a Bacterial
380 Transporter (BioY) and a Host-Cell Transporter (SMVT)', *PLoS ONE*, 7(9), p. e46052. Available at:
381 https://doi.org/10.1371/journal.pone.0046052.

382 French, K.E. (2017) 'Engineering mycorrhizal symbioses to alter plant metabolism and improve crop
383 health', *Frontiers in Microbiology*, 8. Available at: https://doi.org/10.3389/fmicb.2017.01403.

384 Galperin, M.Y. *et al.* (2021) 'COG database update: focus on microbial diversity, model organisms,
385 and widespread pathogens', *Nucleic Acids Research*, 49(D1), pp. D274–D281. Available at:
386 https://doi.org/10.1093/nar/gkaa1018.

387 Gruber, F., Lipozenčić, J. and Kehler, T. (2015) 'History of Venereal Diseases from Antiquity to the
388 Renaissance', *Acta Dermatovenerologica Croatica*, 23(1), pp. 1–11.

389 Halter, T. *et al.* (2022) 'Ecology and evolution of chlamydial symbionts of arthropods', *ISME*
390 *Communications*, 2(1), p. 45. Available at: https://doi.org/10.1038/s43705-022-00124-5.

391 Hamberger, B. *et al.* (2006) 'Comparative Genomics of The Shikimate Pathway in Arabidopsis,
392 Populus Trichocarpa and Oryza Sativa: Shikimate Pathway Gene Family Structure and Identification
393 of Candidates for Missing Links in Phenylalanine Biosynthesis', in *Recent Advances in Phytochemistry*.
394 Elsevier, pp. 85–113. Available at: https://doi.org/10.1016/S0079-9920(06)80038-9.

395 Horn, M. *et al.* (2004) 'Illuminating the Evolutionary History of Chlamydiae', *Science*, 304(5671), pp.
396 728–730. Available at: https://doi.org/10.1126/science.1096330.

Horn, M. (2008) ‘Chlamydiae as Symbionts in Eukaryotes’, *Annual Review of Microbiology*, 62(1), pp. 113–131. Available at: https://doi.org/10.1146/annurev.micro.62.081307.162818.

Huang, J. and Gogarten, J. (2007) ‘Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids?’, *Genome Biology*, 8(6), p. R99. Available at: https://doi.org/10.1186/gb-2007-8-6-r99.

Inkscape Project (2020) ‘Inkscape’. Available at: https://inkscape.org.

Ishida, K. *et al.* (2014) ‘Amoebal Endosymbiont Neochlamydia Genome Sequence Illuminates the Bacterial Role in the Defense of the Host Amoebae against Legionella pneumophila’, *PLoS ONE*, 9(4), p. e95166. Available at: https://doi.org/10.1371/journal.pone.0095166.

Jiang, N. *et al.* (2021) ‘Rhamnose in plants - from biosynthesis to diverse functions’, *Plant Science*, 302, p. 110687. Available at: https://doi.org/10.1016/j.plantsci.2020.110687.

Jofré, E., Lagares, A. and Mori, G. (2004) ‘Disruption of dTDP-rhamnose biosynthesis modifies lipopolysaccharide core, exopolysaccharide production, and root colonization in *Azospirillum brasilense*’, *FEMS Microbiology Letters*, 231(2), pp. 267–275. Available at: https://doi.org/10.1016/S0378-1097(04)00003-5.

Kahane, S. *et al.* (1993) ‘Description and partial characterization of a new chlamydia-like microorganism’, *FEMS Microbiology Letters*, 109(2–3), pp. 329–333. Available at: https://doi.org/10.1111/j.1574-6968.1993.tb06189.x.

Kang, D.D. *et al.* (2019) ‘MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies’, *PeerJ*, 2019(7), p. e7359. Available at: https://doi.org/10.7717/peerj.7359.

Karimi, E. *et al.* (2019) ‘Genomic blueprints of sponge-prokaryote symbiosis are shared by low abundant and cultivatable *Alphaproteobacteria*’, *Scientific Reports*, 9(1), p. 1999. Available at: https://doi.org/10.1038/s41598-019-38737-x.

Kjeldsen, K.U. *et al.* (2010) ‘Two Types of Endosymbiotic Bacteria in the Enigmatic Marine Worm Xenoturbella bocki’, *Applied and Environmental Microbiology*, 76(8), pp. 2657–2662. Available at: https://doi.org/10.1128/AEM.01092-09.

König, L. *et al.* (2017) ‘Biphasic Metabolism and Host Interaction of a Chlamydial Symbiont’, *mSystems*. Edited by A.D. Kent, 2(3). Available at: https://doi.org/10.1128/mSystems.00202-16.

Köstlbacher, S. *et al.* (2021) ‘Pangenomics reveals alternative environmental lifestyles among chlamydiae’, *Nature Communications*, 12(1), p. 4021. Available at: https://doi.org/10.1038/s41467-021-24294-3.

Lagkouvardos, I. *et al.* (2014) ‘Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae’, *The ISME Journal*, 8(1), pp. 115–125. Available at: https://doi.org/10.1038/ismej.2013.142.

Li, D. *et al.* (2015) ‘MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph’, *Bioinformatics*, 31(10), pp. 1674–1676. Available at: https://doi.org/10.1093/bioinformatics/btv033.

Li, H. (2018) ‘Minimap2: pairwise alignment for nucleotide sequences’, *Bioinformatics*. Edited by I. Birol, 34(18), pp. 3094–3100. Available at: https://doi.org/10.1093/bioinformatics/bty191.

Ma, Y., Pan, F. and McNeil, M. (2002) ‘Formation of dTDP-rhamnose is essential for growth of *mycobacteria*’, *Journal of Bacteriology*, 184(12), pp. 3392–3395. Available at: https://doi.org/10.1128/JB.184.12.3392-3395.2002.

Maita, C. *et al.* (2018) ‘Amoebal endosymbiont Neochlamydia protects host amoebae against Legionella pneumophila infection by preventing Legionella entry’, *Microbes and Infection*, 20(4), pp. 236–244. Available at: https://doi.org/10.1016/j.micinf.2017.12.012.

Nayfach, S. *et al.* (2021) ‘A genomic catalog of Earth’s microbiomes’, *Nature Biotechnology*, 39(4), pp. 499–509. Available at: https://doi.org/10.1038/s41587-020-0718-6.

Nylund, A. *et al.* (2018) ‘Genotyping of Candidatus Syngnamydia salmonis (chlamydiales; Simkaniaceae) co-cultured in Paramoeba perurans (amoebozoa; Paramoebidae)’, *Archives of Microbiology*, 200(6), pp. 859–867. Available at: https://doi.org/10.1007/s00203-018-1488-0.

Parks, D.H. *et al.* (2015) ‘CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes’, *Genome Research*, 25(7), pp. 1043–1055. Available at: https://doi.org/10.1101/gr.186072.114.

Pillonel, T., Bertelli, C. and Greub, G. (2018) ‘Environmental Metagenomic Assemblies Reveal Seven New Highly Divergent Chlamydial Lineages and Hallmarks of a Conserved Intracellular Lifestyle’, *Frontiers in Microbiology*, 9. Available at: https://doi.org/10.3389/fmicb.2018.00079.

Pritchard, L. *et al.* (2016) ‘Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens’, *Analytical Methods*, 8(1), pp. 12–24. Available at: https://doi.org/10.1039/C5AY02550H.

Prowazek, S. von S. (1912) ‘Chlamydzoan’, in *Handbuch der pathogenen Protozoen Vol. 1*. Leipzig: J. A. Barth, pp. 119–121. Available at: https://doi.org/10.5962/bhl.title.10198.

Rodriguez-R, L.M. and Konstantinidis, K.T. (2016) ‘The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes’. Available at: https://doi.org/10.7287/PEERJ.PREPRINTS.1900V1.

Rossum, G.V. and Drake, F.L. (2009) *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Shanmugabalaji, V. *et al.* (2022) ‘Plastoglobules: A hub of lipid metabolism in the chloroplast’, in *Advances in Botanical Research*. Elsevier, pp. 91–119. Available at: https://doi.org/10.1016/bs.abr.2021.09.002.

Waskom, M. and Seaborn development team (2020) ‘mwaskom/seaborn’. Zenodo. Available at: https://doi.org/10.5281/zenodo.592845.

Wei, G. *et al.* (2019) ‘Terpene biosynthesis in red algae Is catalyzed by microbial type but not typical plant terpene synthases’, *Plant Physiology*, 179(2), pp. 382–390. Available at: https://doi.org/10.1104/pp.18.01413.

472 ## Figure captions

473 **Figure 1. Genome wide phylogeny of Parachlamydiales.** Maximum likelihood (ML) phylogeny
474 of Parachlamydiales constructed from 34 single copy gene clusters that contain a total of 3604
475 genes. New genomes are indicated by ▲ with blue labels for genomes derived from host SRA
476 bins, and red for MAGS without host assocition data. Bootstrap values based on 1000
477 replicates are indicated with coloured circles (red = 91-100, yellow = 81-90, black <= 80).

478 **Figure 2. Genus and species level clustering across *Parachlamydiales*.** Frutcherman Reingold
479 networks of pairwise a) Average Amino Acid Identity (AAI) with edge weights >65% similarity
480 and b) Average Nucleotide Identity (ANI) with edge weights >95% similarity across all
481 genomes. AAI and ANI illustrate genus and species boundaries, respectively. Proposed genus
482 names are indicated with a ▲.

483 **Figure 3a. Gene content comparison for *Rhabdochlamydiaceae* and *Simkaniaceae*.** An
484 upsetplot illustrating the disparity of gene presence absence across the families a)
485 *Rhabdochlamydiaceae* and b) *Simkaniaceae*. Vertical bars illustrate the number of genes
486 present shared across compared genomes, indicated by black circles beneath each bar.
487 Orange circles indicate the core genome shared by all genomes in the comparison.

488 **Figure 3b. Gene content comparison for *Rhabdochlamydiaceae* and *Simkaniaceae*.** An upset
489 plot illustrating the disparity of gene presence absence across the families a)
490 *Rhabdochlamydiaceae* and b) *Simkaniaceae*. Vertical bars illustrate the number of genes
491 present shared across compared genomes, indicated by black circles beneath each bar.
492 Orange circles indicate the core genome shared by all genomes in the comparison.

493 **Figure 4. The number of each type of metabolic pathway found in each *Chlamydiota***
494 **genome.** Red indicates environmental MAGs; blue indicates MAGs assembled in this study.

495 **Figure 5. Heatmap of the completeness of KEGG metabolic pathways of interest across the**
496 ***Parachlamydiaceae*, *Rhabdochlamydiaceae* and *Simkaniaceae*.** Full metabolic pathway
497 completeness for all genomes and all other pathways are available in Supplementary data 1.
498 New genomes assembled in this study are indicated by ▲.