Host-adaptation in *Legionellales* is 2.4 Gya, coincident with eukaryogenesis

Eric Hugoson[1,2], Tea Ammunét[1 †], and Lionel Guy[1]*

[1] Department of Medical Biochemistry and Microbiology, Science for Life Laboratories, Uppsala University, Box 582, 75123 Uppsala, Sweden

[2] Department of Microbial Population Biology, Max Planck Institute for Evolutionary Biology, D-24306 Plön, Germany

[†] current address: Medical Bioinformatics Centre, Turku Bioscience, University of Turku, Tykistökatu 6A, 20520 Turku, Finland

* corresponding author

Abstract

Events of host-adaptation involving bacteria is at the origin of the most salient events in the evolution of Eukaryotes: the seminal fusion with an archaeon [1,2], and the emergence of both the mitochondrion and the chloroplast [3]. Understanding how bacteria contributed to these events is difficult, due to the scarcity of ancient clades containing host-adapted bacteria only. One such clade is the deep-branching gammaproteobacterial order *Legionellales* – containing among others *Coxiella* and *Legionella* – of which all known members grow inside eukaryotic cells [4]. Here, by analyzing 35 novel *Legionellales* genomes mainly acquired through metagenomics, we show that this order is much more diverse than previously thought, and that key host-adaptation events took place very early in its evolution. Crucial virulence factors like the Type IVB secretion (Dot/Icm) system and two shared effector proteins were gained in the last *Legionellales* common ancestor (LLCA), while many metabolic gene families where lost in LLCA and its immediate descendants. We estimate that the LLCA is circa 2.4 billion years old, predating the last eukaryotic common ancestor by at least 0.5 Ga [5]. This strongly indicates that host-adaptation occurred only once, and early, in the whole order and suggests that *Legionellales* started exploiting and manipulating eukaryotic cells with advanced molecular machines during early eukaryogenesis.

33 # Main text

34 The recent discovery of Asgard archaea and their placement on the tree of life [1,2] shed a bright
35 light on early eukaryogenesis, confirming that Eukaryotes evolved from a fusion between an
36 archaeon and a bacterium. However, many questions pertaining to eukaryogenesis remain
37 unanswered. In particular, the timing and the nature of the fusion is vigorously debated [6–9].
38 Mito-late scenarios [8,10,11] imply that endosymbiosis of the mitochondrion occurred in a proto-
39 eukaryote, likely capable of phagocytosis, while mito-early scenarios [12] imply that
40 mitochondrial endosymbiosis triggered eukaryogenesis. Phagocytosis is a key component of
41 eukaryogenesis, because it is regarded by many to be a pre-requisite for mitochondrion
42 endosymbiosis [7]. Despite uncertainties as to exactly when eukaryotes became able to
43 phagocytose bacteria, it most certainly happened prior to the last eukaryotic common ancestor
44 (LECA) [6,7].

45 The intracellular space of eukaryotic cells is a very rich environment, and many
46 bacteria have adapted to exploit this niche, by resisting being digested by their hosts.
47 Although host-adaptation occurred many times in many different taxonomic groups [13], only a
48 few very large groups encompass only host-adapted members, including *Legionellales*,
49 *Chlamydiales*, *Rickettsiales* and *Mycobacteriaceae*. Hallmarks of adaptation to the
50 intracellular space include smaller population sizes, genome reduction and degradation, and
51 pseudogenization [13]. *Legionellales* is a large and diverse group of the *Gammaproteobacteria*
52 [4,14], encompassing so far only intracellular organisms, including well-studied accidental
53 human pathogens, *Coxiella burnetii* and *Legionella* spp. *Legionellales* show great variation in
54 lifestyle [4], ranging from facultative intracellular (*Legionella* spp.) to the highly reduced,
55 vertically inherited *Coxiella* symbionts of ticks [15,16]. Due to their fastidious nature and their
56 rarity, they are underrepresented in genomic databases, with only 7 genera with sequenced

57 representatives, out of an estimated >450 genera [14]. The Type IVB secretion system (T4BSS)
58 has a key role in interacting with hosts in *Coxiella* [17] and *Legionella* [18,19], injecting a wide
59 diversity of protein effectors into the host cell. Effectors alter the hosts behavior, preventing
60 the bacterium to be digested and helping benefiting from the host resources. In the *Legionella*
61 genus, with over 50 species, the number of effectors reaches several thousand, up to 18 000,
62 but with only eight conserved throughout the genus [20,21].

63 To better understand the evolutionary history of *Legionellales* and their relationships
64 with their early hosts, we gathered 35 genomic sequences from novel *Legionellales*, through
65 whole-genome shotgun, binning of metagenome-assembled genomes (MAGs) and database
66 mining. Bayesian and maximum-likelihood phylogenies on two different datasets confirm that
67 both families in the order, *Legionellaceae* and *Coxiellaceae* (hereafter jointly referred as
68 *Legionellales* sensu stricto) are monophyletic, sister clades, and diverged rapidly after the
69 LLCA (**Fig. 1**). Two other groups, *Berkiella* spp. and *Piscirickettsia* spp. could not be placed
70 with equally high confidence (**Supp. Table 1**). However, Bayesian phylogenies
71 (**Supplementary Figure 1 and 2**) inferred with the CAT model consistently placed *Berkiella*
72 as sister-clade to *Legionellales* sensu stricto (ss). The two latter groups are hereafter
73 collectively referred to as *Legionellales* sensu lato (sl). The placement of *Piscirickettsia* is not
74 as consistent, and future studies will firmly establish whether it groups with *Legionellales* or
75 with *Francisella*.

76 The phylogeny shows a comprehensive, previously unknown diversity among
77 *Legionellales* (**Fig. 1**). The *Coxiellaceae* is divided in two large clades, one including the
78 known genera *Coxiella*, *Rickettsiella* and *Diplorickettsia*, and another ("Aquicella"), about
79 which very little is known, includes the amoebal pathogen *Aquicella* and environmental
80 MAGs. Strikingly, both MAGs branching as sister groups to the *Legionella* genus have small

genomes (2.2 and 1.9 Mb, respectively, see **Supplementary Figure 3**), whereas genomes in the *Legionella* genus (except for *Ca.* L. polyplacis, an obligate louse endosymbiont [22]) range from 2.3 to 4.9 Mb [21]. This suggests that *Legionellaceae* last common ancestor that had a smaller genome than the extant *Legionella* species, although it cannot be excluded that genome reduction occurred independently in these two MAGs.

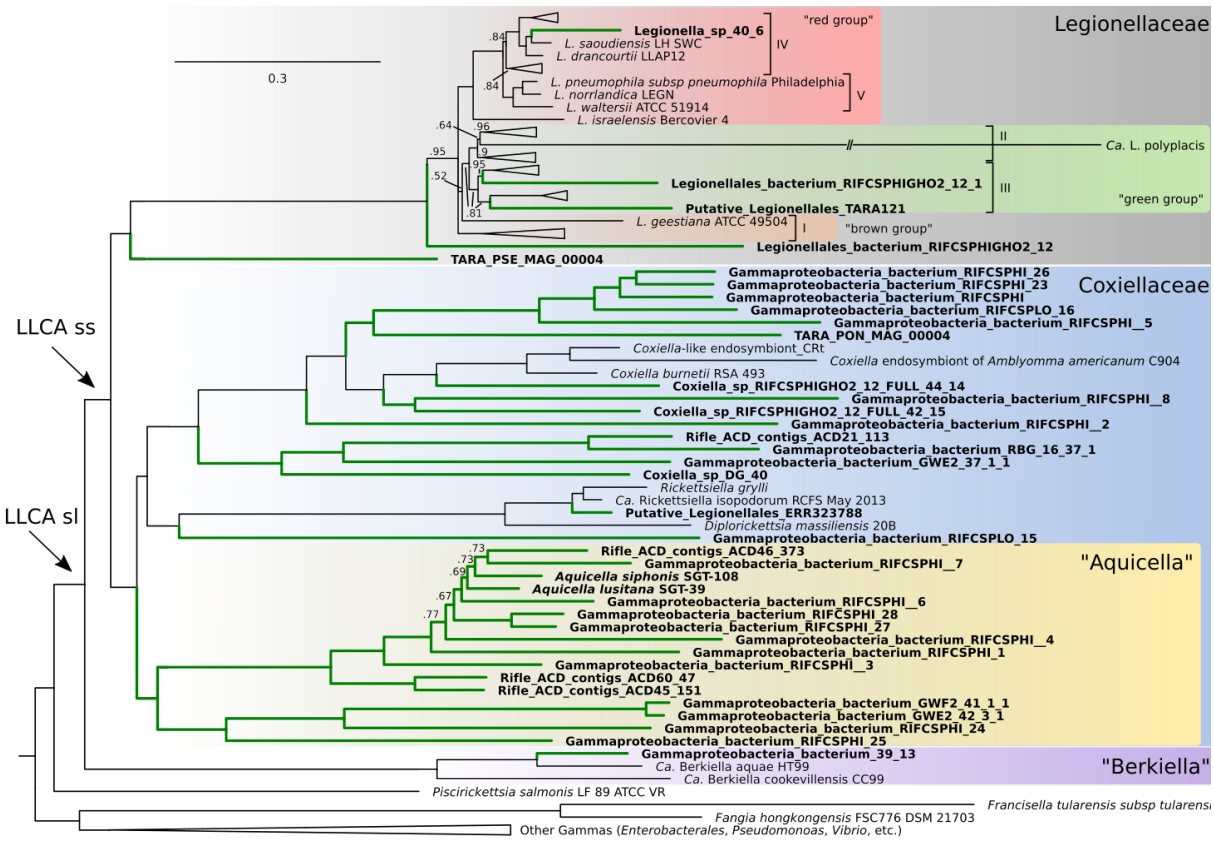


Figure 1 | Bayesian phylogeny of the order *Legionellales*. Based on a concatenated amino-acid alignment of 109 single-copy orthologs (Bact109), the tree was inferred with PhyloBayes using a GTR+CAT model. The tree encompasses 93 *Legionellales* genomes (Legio93) and is rooted with an outgroup of 4 genomes from *Proteobacteria* other than *Gammaproteobacteria*. Numbers on branches show posterior probabilities (pp). Branches without number indicate pp = 1. Terminal nodes in bold indicate taxa recovered in this study. Green branches indicate clades where no genomic data was available prior to this study. The scale shows the average number of substitutions per site. The colored groups and Roman numerals in the

95 *Legionellaceae* correspond to the groupings in [20] and [21], respectively. The full tree is available

96 in **Supplementary Figure 1**.

97 To better understand the evolution of host-adaptation in the order, we reconstructed

98 the gene flow within the *Legionellales*. The phylogenetic birth-and-death model implemented

99 in Count [23] was used on the same set of genomes (Legio93) as the tree in **Fig. 1**. The

100 reconstructed gene flow reveals a total of 553 gene family losses, from the last free-living

101 ancestor to LLCA sl, (**Fig. 2, Supplementary Figure 3, Supp. Table 2**). These 553 losses

102 can be compared to the 781 on the branch leading to *Fangia*/*Francisella*, another intracellular

103 group. A subsequent large gain in gene families in the last common ancestors of

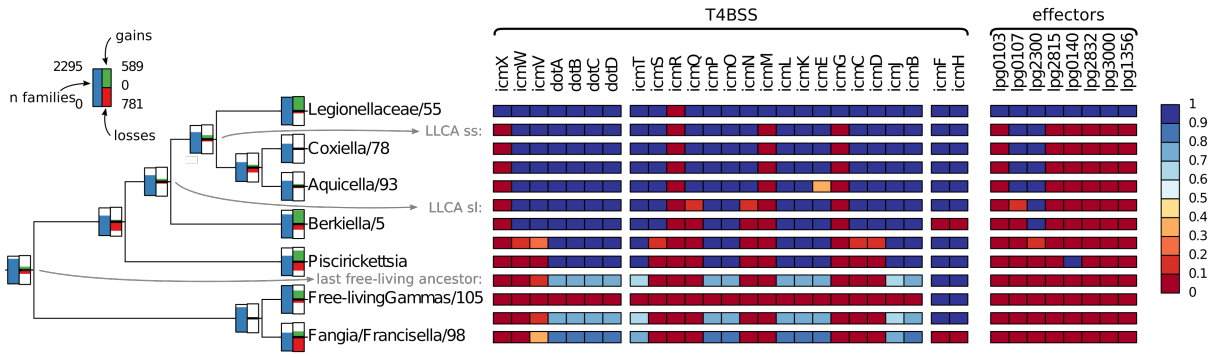
104 *Legionellaceae* and *Berkiella* spp. (588 and 556, respectively) but not in the ancestors of

105 *Coxiellaceae* (*Coxiella/Aquicella*) suggests that LLCA had a genome size comparable to that

106 of the latter group (average: 1.8 Mb), compatible with genome sizes of gammaproteobacterial

107 intracellular bacteria [13].

108



109 **Figure 2 | Ancestral reconstruction of the *Legionellales* genomes, particularly on the**

110 **T4BSS system and the *Legionella*-conserved effectors.** The ancestral reconstruction is

111 based on the Bayesian tree in **Fig. 1**. At the left of each node, barplots depict the number of

112 gene families (blue, ranging from 0 to 2295) inferred in that ancestor, as well as the number

113 of family gains (green, from 0 to 589) and losses (red, from 0 to 781) occurring on the branch

114 leading to that ancestor. On the right panel, squares represent proteins included in the T4BSS

115 (separated according to the three operons present in R64 [19]) and the eight effectors conserved
116 in *Legionella* [21]. Rows in-between terminal nodes correspond to the last common ancestor of
117 the two nearest terminal nodes, e.g. the row between *Aquicella* and *Berkiella* corresponds to
118 ancestor 96, the last common ancestor of *Legionellales* sensu stricto. The color represents the
119 posterior probability that each protein was present in the ancestor of each group. A complete
120 tree is shown on **Supplementary Figure 3** and the underlying numbers are available in **Supp.**
121 **Tables 2 and 3**.

122 A crucial feature for the success of many host-adapted bacteria is their ability to
123 secrete proteins into their hosts cytoplasm, typically through secretion systems. Among the
124 proteobacterial VirB4-family of ssDNA conjugation systems (Type IV Secretion Systems,
125 T4SS), the T4BSS, also referred to as MPF_I or I-type T4SS, is probably the earliest diverging
126 [24]. It is present in almost all extant *Legionellales*, with the exception of the extremely reduced
127 endosymbionts of the group (**Supplementary Figure 3**), where it is either missing completely
128 or pseudogenized. It is partly missing in several of the small, novel *Coxiellaceae* MAGs,
129 indicative of an increased dependence on their host, although metagenomics binning artefacts
130 cannot be excluded. T4BSS in *Legionellales* appear largely collinear (**Fig. 3, Supplementary**
131 **Figure 4**). Ancestral reconstruction of the order also revealed that the T4BSS, as it is found in
132 extant *Legionellales*, was shaped between the last free-living ancestor and the LLCA, with the
133 additions of several proteins (**Fig. 2, Supplementary discussion**). A homologous T4BSS
134 system is present in *Fangia hongkongensis*, but mostly absent in the free-living
135 *Gammaproteobacteria* and in *Francisella*, a sister clade to *Fangia*. It is likely that the T4BSS
136 in *Fangia* has been acquired by horizontal gene transfer, possibly from *Piscirickettsia*
137 (**Supplementary Figure 5** and **6**).

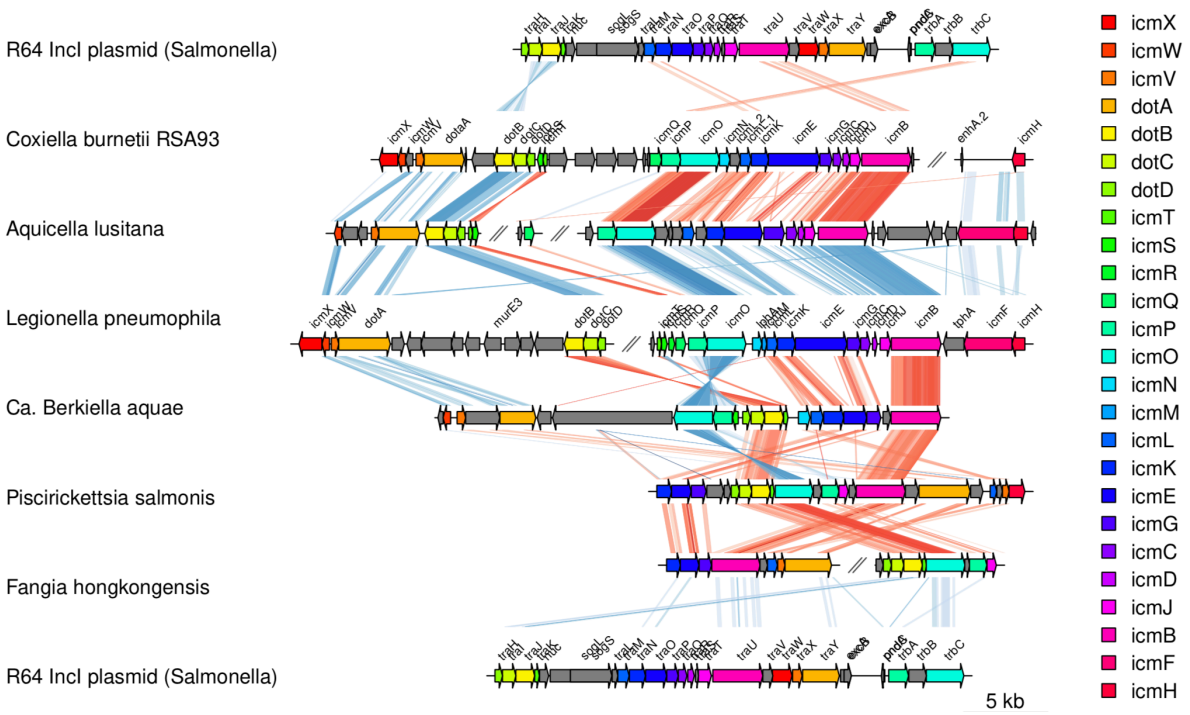


Figure 3 | Collinearity of the T4BSS in *Legionellales*. Genes are colored as they appear in *Legionella pneumophila* Philadelphia. Areas between the rows display similarities between the sequences, as identified by tblastx.

Contrarily to the T4BSS itself, which is highly conserved throughout the order, effectors are much more versatile. Of the eight effectors conserved in the *Legionella* genus [21], only two are found outside of the *Legionellaceae* family (**Fig. 2**): RavC (lpg0107) is found in most *Legionellales* sensu stricto, LegA3/AnkH/AnkW (lpg2300) in the latter as well as in *Berkiella*. MavN (lpg2815), which is conserved in *Legionella* spp. and was found in *Rickettsiella* [20], is found only in a few *Coxiellaceae* species (**Fig. 2**; **Supplementary Figure 3**), and is unlikely to have been acquired in the LLCA. Although the function of these effectors is not precisely known, LegA3/AnkH contains ankyrin repeats, and has been suggested to play a role in process involved in modulation of phagosome biogenesis [25].

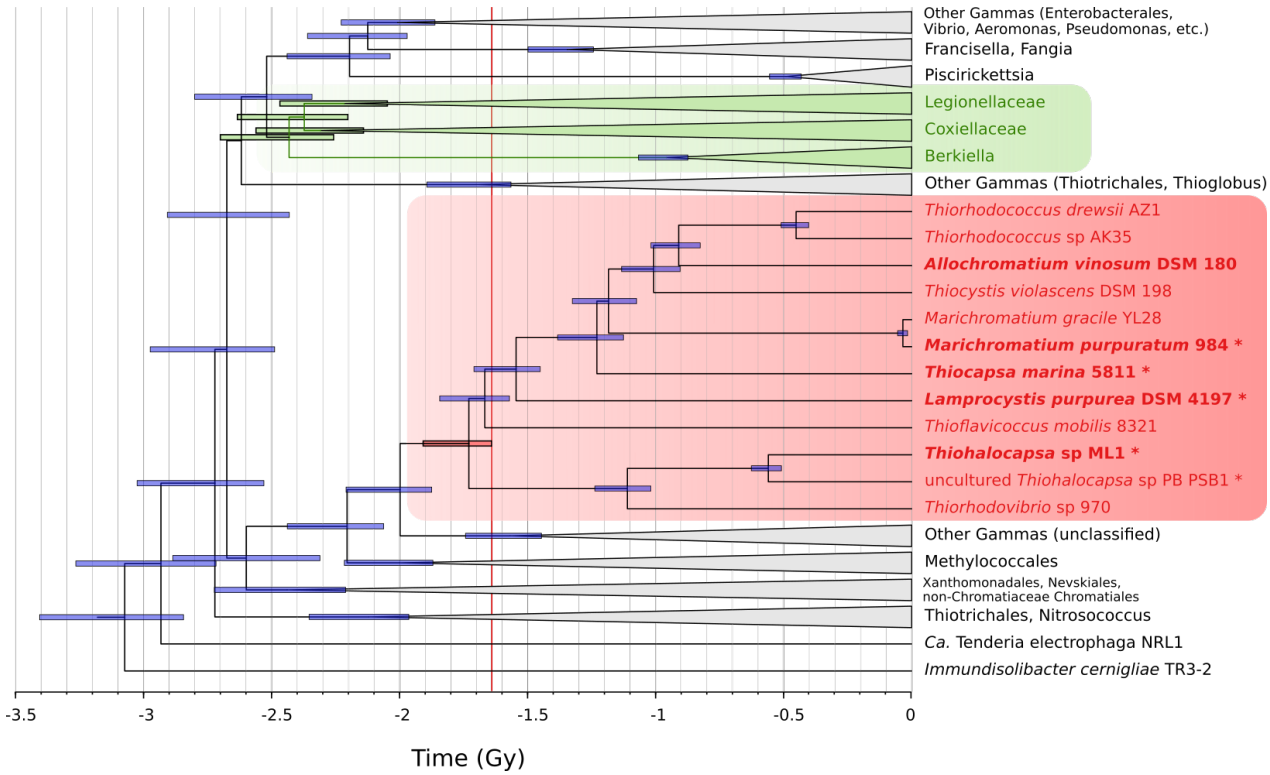


Figure 4 | Bayesian estimation of the appearance of the LLCA. *Chromatiaceae* are highlighted in red, with okenone-producing species in bold. *Legionellales* are highlighted in green. A red bar is drawn at 1.64 Gy, the age of the Barney Creek Formation which contains okenane, and gives the lower estimate for the divergence of *Chromatiaceae*. Colored areas at the nodes represent the highest posterior density interval (95%). The red area is located at the last common ancestor of all okenone-producing *Chromatiaceae*, the green ones are ancestors of the *Legionellales*, with, from top to bottom, the last common ancestor of *Legionellaceae*, *Legionellales* ss, *Coxiellaceae*, and *Legionellales* ls (LLCA). The underlying tree is shown on **Supplementary Figure 2**.

Absolute dating of bacterial trees is often inconclusive, because very few reliable biomarkers can be unambiguously attributed to a bacterial clade and thus be used as calibration points [26]. One of these is okenone, a pigment whose degradation product okenane has been found in the 1.64-Ga-old Barney Creek Formation [27], and is, to the best of our knowledge, exclusively produced by a subset of *Chromatiaceae*. *Chromatiaceae* (purple

166 sulfur bacteria) and *Legionellales* are phylogenetically close to each other, both belonging to
167 the basal *Gammaproteobacteria* [28]. We reasoned that the last common ancestor of all
168 okenone-producing *Chromatiaceae* had to be at least as old as the earliest known trace of
169 okenone, i.e. ≥1.64 Ga (**Fig. 4**). Using this constraint and a relaxed molecular clock model,
170 we were able to calibrate a Bayesian tree, based on a dataset including genomes from 105
171 *Gammaproteobacteria* (of which 22 belong *Legionellales* and 19 to *Chromatiaceae*) and 5
172 outgroups (Gamma105 dataset). We estimate that LLCA existed ca. 2.43 Ga ago (**Fig. 4**),
173 with a 95% high posterior density ranging from 2.70 to 2.26 Ga (see **Supp. Table 4**).

174 *Legionellales'* common ability to invade eukaryotic cells combined with the presence
175 and high conservation of the crucial host-adaptation genes, namely a complete T4BSS and
176 two effectors, strongly suggest that LLCA was already infecting other cells. It is also likely
177 that, like extant *Legionellales*, LLCA depended on phagocytosis, a mechanism exclusively
178 found in Eukaryotes. This implies that LLCA came after the division of Archaea and the first
179 eukaryotic common ancestor, FECA, thus placing an upper bound for the existence of
180 phagocytosis and for FECA at 2.26 Ga. Given an estimate of the age of LECA of 1.0-1.6 Ga
181 [5], the time for eukaryogenesis, defined as the time elapsed between FECA and LECA [6],
182 would be at least of 660 Ma, and up to over 1.5 Ga.

183 Four of the most prominent hypotheses describing eukaryogenesis (reviewed e.g. in [9])
184 make different assumptions about the timing of the mitochondrial endosymbiosis. In the
185 hydrogen hypothesis [12], the mitochondria arrived early ('mito-early'), with the mitochondrial
186 endosymbiosis event itself triggering eukaryogenesis. In the phagocytosing archaeon model
187 (PhAT) [29], the syntrophy hypothesis [10] and the serial endosymbiosis model [8], the
188 mitochondrion arrive late ('mito-late'). In the PhAT model, the existence of the phagocytosis
189 machinery is a pre-requisite for the fusion of the mitochondrial progenitor and the Asgard

190 archaeon. The very early emergence of phagocytosis thus favors a mito-late scenario. The
191 recently proposed ability of Heimdallarchaeota, the closest relatives of the first eukaryotic
192 common ancestor (FECA), to metabolize organic substrates [11] is consistent with early
193 eukaryotes being able to profit from phagocytosed prokaryotes.

194 In conclusion, we show that *Legionellales* are a very diverse group, which evolved an
195 intracellular lifestyle only once, mainly owing to a T4BSS resembling the one found in extant
196 *Legionellales*. This single event of host-adaptation took place circa 2.43 Ga ago, and is
197 consistent with a scenario where very early eukaryotes developed phagocytic properties,
198 possibly preying on prokaryotes. Some of these prokaryotes, among them the LLCA, acquired
199 the ability of resisting digestion by their hosts, and became able to exploit this novel, rich
200 ecological niche.

201 Material and Methods

202 Bacterial strains

203 Two species of *Aquicella* (*A. lusitana*, DSM 16500 and *A. siphonis*, DSM 17428) [30] were
204 ordered from the DSMZ strain culture. They were grown at 37°C for 5 days on Buffered
205 Charcoal Yeast Extract (BCYE) agar plates with Legionella BCYE supplement (OXOID).

206 DNA Extraction and sequencing, assembly and annotation

207 DNA was extracted from pure culture isolates (*Aquicella* spp.) using the Genomic Tip 100
208 DNA extraction kit (QIAGEN), following the manufacturer's instructions. DNA extracted
209 from pure cultures was sequenced by PacBio, using one SMRT cell for each isolate. Reads
210 were assembled using the HGAP3 pipeline, resulting in one unitig per replicon. The assembly
211 was confirmed by performing cumulative GC skews on each replicon [31]. All replicons had
212 repetitive sequences at the beginning and the end of the sequences, suggesting that they were

213 circular. One unitig in *A. lusitana* had multiple repeats of the same sequences and was edited
214 manually. All chromosomes were rotated to have the origin of replication, as determined by
215 cumulative GC skews, at the beginning of the sequence. The final sequences were confirmed
216 by mapping ~500k shorter reads (2x300 bp, obtained from an Illumina MiSeq run) to the
217 sequences obtained from PacBio. No SNP could be found in any of the replicons. All
218 sequencing was performed at NGI, SciLifeLab, Uppsala and Stockholm, Sweden.

219 The sequences were then annotated with prokka 1.12-beta [32], using prodigal 2.6.3 [33] to
220 call protein-coding genes, Aragorn 1.2 [34] to predict tRNAs and barrnap 0.7 [35] to predict
221 rRNAs. Default functional annotation was performed by prokka. Both genomes are submitted
222 at the European Nucleotide Archive (ENA) under study number PRJEB29684.

223 Protein marker sets

224 A set of 139 PFAM protein domains was used as a starting point both to perform quality
225 control of MAGs and for phylogenomic reconstructions. The original set was used by Rinke
226 et al (Supplementary Table 13 in [36]) to estimate the completeness of their MAGs. We used the
227 same set (referred to as Bact139) to estimate MAG completeness. A large subset of the 139
228 protein domains are very frequently located in the same protein in a vast majority of
229 Proteobacteria. This was evidenced by investigating 1083 genomes, using phyloSkeleton [37] to
230 select one representative per proteobacterial genus and to identify proteins containing the
231 Bact139 domain set. Of these domains which often co-localized in the same protein, only the
232 most widespread one was retained. In total, only 109 domains were used to identify proteins
233 suitable for phylogenomics analysis. This set is referred to as Bact109 and is available in
234 phyloSkeleton [37].

235 ## Metagenome assembly

236 The 86 metagenomes with the highest fraction of reads belonging to *Legionellales* were
237 selected (**Supp. Table 5**). Metagenomes sequenced only with 454 (Roche), or that were too
238 complex to be assembled on the computational cluster at our disposition (512 Gb RAM) were
239 not included. Raw reads were downloaded from the European Nucleotide Archive (ENA) and
240 trimmed using Trim Galore (v0.4.2) [38] to remove standard adapter sequences. Additionally,
241 reads were trimmed using Trimmomatic (v0.36) [39] with a sliding window (4:15) to only
242 discard low-quality bases without losing whole reads. Trimmed reads were then assembled
243 using SPAdes (St. Petersburg genome assembler) (v3.7.1) [40], using the --meta option and
244 kmer sizes 21, 33, and 55 for metagenome sizes that did not require running on a HPCC
245 (High Performance Computer Cluster). More complex metagenomes requiring a HPCC were
246 assembled on UPPMAX with a single kmer, 31. Contigs smaller than 500 bp were then
247 discarded.

248 ## Identification of MAGs belonging to *Legionellales*

249 To identify MAGs belonging to *Legionellales*, the ggkBase (https://ggkbase.berkeley.edu/)
250 and NCBI Genome databases were screened for MAGs attributed to Gammaproteobacteria.
251 MAGs obtained through metagenomic binning from a previous published study [41] and from a
252 prepublication [42] that branched close to *Legionella pneumophila* were also included.

253 The selected MAGs and the assembled metagenomes were screened using the same
254 ribosomal protein (r-protein) pipeline as described in [2], hereafter referred to as RP15. Briefly,
255 this pipeline uses first PSI-BLAST to search the given metagenomes and MAGs for 15
256 ribosomal proteins, which are universally located in a single chromosomal locus. Contigs
257 containing at least 8 of the 15 r-proteins are then retained, and referred to as r-contigs.

258 A backbone of known organisms (a set of organisms with complete genome and
259 known taxonomy) was used to help identifying to which known clades the r-contigs belong.
260 The package phyloSkeleton [37] was used to select one representative genome in each family
261 within Gammaproteobacteria, one in each order of the Beta- and Alphaproteobacteria, as well
262 as *Mariprofundus* and *Acidithiobacillus*. PhyloSkeleton also identified orthologs of 15
263 ribosomal proteins in the gathered genomes.

264 The RP15 orthologs gathered from both the metagenomic sources and the backbone
265 were aligned with MAFFT L-INS-I v7.273 [43], and the alignments concatenated through the
266 RP15 pipeline. The resulting concatenated alignment was used to infer a rough phylogenetic
267 tree using FastTree 2.1.8 [44], using the WAG model. R-contigs that were clearly unrelated to
268 *Legionellales* were removed from the dataset, while the remaining proteins were realigned
269 and concatenated, to generate a new tree. This process was iterated until all remaining MAGs
270 and r-contigs were deemed belonging to *Legionellales*. A final phylogeny was inferred with
271 RAxML 8.2.8 [45] using the PROTGAMMA model which allowed automatic selection of best
272 substitution matrix to infer a more robust phylogeny.

273 Metagenomic binning and quality control

274 Metagenomes containing r-contigs belonging to *Legionellales* were then binned with ESOM
275 [46,47]. An ESOM map was generated from contigs over 2 kb with a 10 kb sliding window, and
276 visualized using the ESOM analyzer, highlighting the location of the r-contig and selecting
277 the corresponding bin. The completeness and redundancy of all MAGs were then controlled
278 using miComplete [48]

279 Some MAGs were very similar to each other, due to different versions of the same bin
280 being present in databases. The most complete MAG was then selected for each set, provided
281 they were not more redundant than another genome within the same set. If equal in

282 completeness the NCBI- submitted MAG was preferred.

283 Phylogenomics

284 Two sets of genomes were used for phylogenomics in this contribution. Both sets consist of
285 part or all of the novel *Legionellales* MAGs and genomes, supplemented with varying subsets
286 of outgroups. The initial selection of the representative organisms and identification of
287 markers was done with phyloSkeleton v1.1 [37], initially choosing one representative per class
288 in the *Betaproteobacteria*, the *Zetaproteobacteria* and the *Acidithiobacillia*; one per genus in
289 the Gammaproteobacteria, except in the *Piscirickettsiaceae* and the *Francisellaceae* (one per
290 species). That initial selection of genomes was then trimmed down to reduce the
291 computational burden of tree inference.

292 For both sets, protein markers from the Bact109 set were identified with
293 phyloSkeleton v1.1 [37]. Each marker was aligned separately with mafft-linsi v7.273 [43]. The
294 resulting alignments were trimmed with BMGE v1.12 [49], using the BLOSUM30 matrix and
295 the stationary-based trimming algorithm. The alignments were then concatenated. From these
296 alignments, maximum-likelihood trees were reconstructed with FastTreeMP v2.1.10 [44] using
297 the WAG substitution matrix. Upon visual inspection of the resulting trees, the genome
298 selection was reduced to remove closely related outgroups, and the phylogenomic procedure
299 repeated. Once the genome dataset was final, a maximum-likelihood tree was inferred with
300 IQ-TREE v1.6.5 [50], using the LG [51] substitution matrix, empirical codon frequencies, four
301 gamma categories, the C60 mixture model [52] and PMSF approximation [53]; 1000 ultrafast
302 bootstraps were drawn [54].

303 The first set of genomes, referred to as Gamma105 was used to correctly place the
304 *Legionellales* in the *Gammaproteobacteria*, in particular with respect to the *Chromatiaceae*. It
305 encompasses 110 genomes, of which 105 are *Gammaproteobacteria*, one is a

306 *Zetaproteobacterium*, 3 are *Betaproteobacteria*, and one is an *Acidithiobacillia*. The selected
307 *Gammaproteobacteria* include representatives from *Legionellales* (22), *Chromatiaceae* (19),
308 *Francisellaceae* (17) and *Piscirickettsiaceae* (4) (**Supp. Table 6**).

309 The second set, referred to as Legio93, is focused on the *Legionellales* itself and
310 encompasses 113 genomes, of which 93 belong to *Legionellales* and the rest to
311 Gammaproteobacteria (16 genomes), Betaproteobacteria (2 genomes), Acidithobacillia (1
312 genome) and Zetaproteobacteria (1 genome) (**Supp. Table 7**).

313 For both sets, single-gene maximum-likelihood trees were inferred for each marker.
314 The BMGE-trimmed alignments were used to infer a tree using IQ-TREE 1.6.5, using the
315 automatic model finder [55], limiting the matrices to be tested to the LG matrix and the C10 and
316 C20 mixture models. Single-gene trees were visually inspected for the presence of very long
317 branches resulting from distant paralogs being chosen by phyloSkeleton.

318 For both sets, a Bayesian phylogeny was inferred with phylobayes MPI 1.5a [56] from
319 the concatenated BMGE-trimmed alignments, using a CAT+GTR model. Four parallel chains
320 were run for 9500 generations (Gamma105 set) and 3500 generations (Legio93 set),
321 respectively. The chains did not converge in either set. For Legio93, all four chains yielded
322 the same topology for all deep nodes in and around *Legionellales*. For Gamma105, 3 out of 4
323 chains yielded the same overall topology, while the last one had the *Piscirickettsia* and the
324 *Berkiella* clades inverted (**Supp. Table 1**). The majority-rule tree obtained from the Bayesian
325 trees for Legio93 was subsequently used for the ancestral reconstruction (see below), while
326 the one for Gamma105 was used as input tree to estimate the time of divergence of the LLCA
327 (see below).

328 ## Identification of okenone-producing *Chromatiaceae*

329 To relate the earliest documented trace of okenone [27,57] to a specific ancestor, we took two
330 approaches. First, we compared a list of okenone-producing *Chromatiaceae* [58] with the
331 genomes available in Genbank and at the JGI Genome Portal [59]. We found five okenone-
332 producing *Chromatiaceae* genomes. Second, we used homologs of the proteins encoded by
333 the genes *crtU* and *crtY*, which are considered essential to synthesize okenone [60], to screen the
334 nr database. We used the proteins from *Thiodictyon syntrophicum* str. Cad16 (accession
335 number: CrtU, AEO72326.1; CrtY, AEO72327.1) as queries in a PSI-BLAST search [61]
336 restricted to the order *Chromatiales*, keeping only proteins with at least 50% similarity.
337 Gathered homologs (n=7 in both cases) were aligned with MAFFT L-INS-I v7.273 [43] and the
338 alignment was pressed into a hidden Markov model (HMM) with HMMer 3.1b2 [62]. Using
339 phyloSkeleton [37], we collected all *Chromatiales* genomes from NCBI (n=130) and queried
340 these, together with the Gamma105 set with both HMMs. A significant similarity (E-value <
341 10^{-10}) for CrtU and CrtY was found in 42 and 25 genomes, respectively. A hit for both
342 proteins was found in 8 genomes only, of which 6 belong to the *Chromatiaceae*. Of these 6, 4
343 were already identified by the first method listed above; one was a MAG that was too
344 incomplete to include in further phylogenomics analysis; the last one was another MAG that
345 was included in later analysis. The other two genome that had proteins similar to CrtU and
346 CrtY are *Kushneria aurantia* DSM 21353, a member of the *Halomonadaceae* and
347 *Immundisolibacter cernigliae* TR3-2, which branches very close to the root of the
348 *Gammaproteobacteria*: these two might represent cases of horizontal gene transfer.

349 ## Time-constrained tree

350 To estimate the time of divergence of LLCA, we used BEAST 2 [63], using a strict and a
351 relaxed clock model [64]. We also attempted to use a random local clock, but it was not

352 computationally tractable. To reduce computational load, we used a fixed input tree
353 (Bayesian, Gamma105; see above). The following parameters were used: the concatenated
354 alignment (104 proteins) was used as a single partition; for site model, a LG matrix with 4
355 gamma categories and 10% invariant was used, estimating all parameters; for the relaxed
356 clock model, an uncorrelated log-normal model was used, estimating the clock rate. The
357 following prior distributions were used: shape of the gamma distributions for the site model:
358 exponential; mutation rate: 1/x; proportion of invariants: uniform; ucld mean: 1/X, offset: 0,
359 ucld std: gamma. To calibrate the clock, it was assumed that the last common ancestor of all
360 okenone-producing *Chromatiaceae* was at least as old as the rocks where the earliest traces of
361 okenane (a degradation product of okenone) was found, 1.640 ± 0.03 Gy old [27,57,65]. A prior
362 was thus set for that the divergence time of the monophyletic group comprising these
363 organisms: the divergence time was drawn from an exponential distribution of mean 0.1 Gy
364 with an offset of 1.64 Gy. Although it cannot be completely excluded, no evidence suggests
365 that other organisms had produced okenone earlier than the last common ancestor of
366 *Chromatiaceae*: the only mentions of okenone in the scientific literature are linked to the
367 presence of *Chromatiaceae*.

368 The four chains run for the strict clock model (10 million generations each) converged
369 but the strict clock assumption seemed unrealistic when examining the variation of mutation
370 rates obtained from the relaxed clock model. For the relaxed clock model, five chains were
371 run, one for 55.7 million generations, and four for 10 million generations each. Traces were
372 examined with Tracer 1.6.0 [66]. Most of the parameters achieved an ESS score > 100. The
373 mutation rate and the mean of the ucld parameter did not reach very high ESS scores but were
374 negatively correlated with each other. Trees were extracted with TreeAnnotator 2.4.7 (part of
375 BEAST), deciding the burn-in fraction after visual examination of the traces (see **Supp.**
376 **Table D**). Despite not fully converging, the five trees from the five chains yielded similar

377 results.

378 To estimate how the chosen priors affected the runs, a run with 10 million generations
379 was performed, drawing from the priors. Parameters were compared in Tracer. All parameters
380 displayed very different distributions, suggesting that the choice of the priors had very little
381 effect on the obtained values.

382 Protein family clustering and annotation

383 All 113 proteomes from the Legio93 set were clustered into protein families using OrthoMCL
384 [67]. All proteins were aligned to each other with the blastp variant of DIAMOND v0.9.8.109 [68],
385 with the --more-sensitive option, enabling masking of low-complexity regions (--masking 1),
386 retrieving at most 10^5 target sequences (-k 100000), with a E-value threshold of 10^{-5} (-e 1e-5),
387 and a pre-calculated database size (--dbsize 98280075). In the clustering step, mcl 14-137 [69]
388 was called with the inflation parameter equal to 1.5 (-I 1.5), as recommended by OrthoMCL.

389 Annotation of the protein families obtained by OrthoMCL was performed by
390 searching for protein accession number in a list of reference genomes (see below). If any
391 protein in a family was present in the first genome, the family was attributed this annotation;
392 else, the second genome was searched for any of the accession numbers, and so on. The
393 reference list was as follow (Genbank accession numbers between parentheses): *Legionella*
394 *pneumophila* subsp. *pneumophila* strain Philadelphia 1 (AE017354), *Legionella longbeachae*
395 NSW150 (FN650140), *Legionella oakridgensis* ATCC 33761 (CP004006), *Legionella*
396 *fallonii* LLAP-10 (LN614827), *Legionella hackeliae* (LN681225), *Tatlockia micdadei*
397 ATCC33218 (LN614830), *Coxiella burnetii* RSA 493 (AE016828), *Coxiella* endosymbiont
398 of *Amblyomma americanum* (CP007541), *Escherichia coli* str. K-12 substr. MG1655
399 (U00096), *Francisella tularensis* subsp. *tularensis* (AJ749949), *Piscirickettsia salmonis* LF-
400 89 = ATCC VR-1361 (CP011849), *Acidithiobacillus ferrivorans* SS3 (CP011849), *Neisseria*

401 *meningitidis* MC58 (AE002098), *Rickettsiella grylli* (NZ_AAQJ00000000.2). *Legionella*
402 effector orthologous groups were identified by comparing protein accession numbers with the
403 list provided by [20].

404 Ancestral reconstruction

405 The flow of genes in the order *Legionellales* was analyzed by reconstructing the genomes of
406 the ancestors. The ancestral reconstruction was performed with Count [23], which implements a
407 maximum-likelihood phylogenetic birth- and death-model. The Bayesian input tree was
408 obtained from the Legio93 dataset (see above) and the OrthoMCL families as described
409 above. The rates (gain, loss, and duplication) were first optimized on the OrthoMCL protein
410 families, to allow different rates on all branches. The family size distribution at the root was
411 set to be Poisson. All rates and the edge length were drawn from a single Gamma category,
412 allowing 100 rounds of optimization, with a 0.1 likelihood threshold. The reconstruction of
413 the gene flow was then performed by using posterior probabilities.

414 Analysis of the Type IVB secretion system

415 Homologs of the Type IVB secretion system were identified with MacSyFinder v1.0.3 [70]. We
416 used the pT4SSi profile, corresponding to the Type IV Secretion System encoded by the IncI
417 R64 plasmid and *Legionella pneumophila* (T4BSS; also referred to as MPF_I) [71], as query to
418 search all proteomes in the Gamma105 and the Legio93 datasets. This profile includes models
419 for the 17 proteins of the IncI R64 plasmid that have a homolog in *L. pneumophila*. The
420 profile does not cover 6 of the Dot/Icm proteins shared among *L. pneumophila* and *C. burnetii*
421 (IcmFHNQSW). MacSyFinder was run in "ordered_replicon" mode, and using the following
422 HMMer options: --coverage-profile 0.33, --i-evalue-select 0.001. A T4BSS was identified if,
423 within 30 neighboring genes: (i) a IcmB/DotU/VirB4 homolog was present, (ii) at least 6 out
424 of the other 16 were also present, (ii) no relaxase (MOBB) were identified. Since the Icm/Dot

425 system is often split in several operons, not all proteins could be identified by MacSyFinder in
426 "ordered_replicon" mode, but most were since the operon which contains IcmB/DotU/VirB4
427 is the largest.

428 To inspect the gene organization further in a selection of organisms, a subset of 12
429 representative genomes were aligned with tblastx 2.2.30 [72], with an E-value threshold of 0.01,
430 not filtering the sequences for low complexity regions, and splitting the genomes in smaller
431 pieces when reaching the maximum number of hits between two segments. Results were first
432 visualized in ACT [73], and a final version of the comparison was drawn with genoPlotR [74]. In
433 the cases where MacSyFinder failed to detect a T4BSS, a similar visual inspection of the
434 comparison of the genome with both *L. pneumophila* and *C. burnetii* was performed.

435 For the genomes where the T4BSS was detected by MacSyFinder, twelve proteins
436 (IcmBCDEGKLOPT, DotAC) were retrieved, aligned with mafft-linsi v7.273 [43], trimmed
437 with trimAl v1.4.rev15 [75], using the -automated1 setting, and concatenated. The concatenated
438 alignment was used to infer a maximum-likelihood tree with IQ-TREE v1.6.8 [50], forcing the
439 use of the LG [51] substitution matrix, but letting ModelFinder [55] select the across-site
440 categories, which resulted in the C60 mixture model [52]; 1000 ultrafast bootstraps were drawn
441 [54]. For the genomes not belonging to *Legionellales lato sensu* (R64, *Fangia hongkongensis*,
442 *Acidithiobacillus ferrivorans*, *Acidiferrobacter thiooxydans* and *Alteromonas stellipolaris*),
443 only a few proteins were retrieved by the automated approach. A preliminary tree displayed
444 very long branches, and the procedure above was repeated without these four genomes. For
445 the 12 genomes for which the T4BSS was manually annotated, a tree based on the alignment
446 of 26 proteins belonging to the T4BSS was performed as above, resulting in a LG+F+I+G4
447 model being selected.

448 ## Data availability

449 The raw and assembled data for the *Aquicella* genomes is deposited at the European
450 Nucleotide Archive (ENA) under study accession PRJEB29684. The assemblies for *A.*
451 *lusitana* and *A. siphonis* have the accessions GCA_902459475 (replicons LR699114.1-
452 LR699118.1) and GCA_902459485 (replicons LR699119.1- LR699120.1).

453 MAGs, associated proteomes and alignments underlying the trees presented in this
454 contribution are available at Zenodo, with DOI 10.5281/zenodo.3543579

455 # References

456 1. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and
457 eukaryotes. *Nature* **521**, 173--179 (2015).

458 2. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic
459 cellular complexity. *Nature* **541**, 353–358 (2017).

460 3. Sagan, L. On the origin of mitosing cells. *J. Theor. Biol.* **14**, 225-IN6 (1967).

461 4. Duron, O., Doublet, P., Vavre, F. & Bouchon, D. The Importance of Revisiting
462 Legionellales Diversity. *Trends Parasitol.* **34**, 1027–1037 (2018).

463 5. Eme, L., Sharpe, S. C., Brown, M. W. & Roger, A. J. On the age of eukaryotes:
464 evaluating evidence from fossils and molecular clocks. *Cold Spring Harb Perspect Biol* **6**,
465 (2014).

466 6. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin
467 of eukaryotes. *Nat. Rev. Microbiol.* **15**, 711–723 (2017).

468 7. Poole, A. M. & Gribaldo, S. Eukaryotic Origins: How and When Was the Mitochondrion
469 Acquired? *Cold Spring Harb. Perspect. Biol.* **6**, a015990 (2014).

470 8. Pittis, A. A. & Gabaldón, T. Late acquisition of mitochondria by a host with chimaeric
471 prokaryotic ancestry. *Nature* **531**, 101–104 (2016).

472 9. López-García, P., Eme, L. & Moreira, D. Symbiosis in eukaryotic evolution. *J. Theor.*
473 *Biol.* **434**, 20–33 (2017).

474 10. López-García, P. & Moreira, D. Selective forces for the origin of the eukaryotic nucleus.
475 *BioEssays* **28**, 525–533 (2006).

476 11. Spang, A. *et al.* Proposal of the reverse flow model for the origin of the eukaryotic cell
477 based on comparative analyses of Asgard archaeal metabolism. *Nat. Microbiol.* **4**, 1138–
478 1148 (2019).

479 12. Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**,
480 37–41 (1998).

481 13. Toft, C. & Andersson, S. G. E. Evolutionary microbial genomics: insights into bacterial
482 host adaptation. *Nat. Rev. Genet.* **11**, 465–475 (2010).

483 14. Graells, T., Ishak, H., Larsson, M. & Guy, L. The all-intracellular order Legionellales is
484 unexpectedly diverse, globally distributed and lowly abundant. *FEMS Microbiol. Ecol.*
485 **94**, (2018).

486 15. Guizzo, M. G. *et al.* A Coxiella mutualist symbiont is essential to the development of
487 Rhipicephalus microplus. *Sci. Rep.* **7**, 17554 (2017).

488 16. Gottlieb, Y., Lalzar, I. & Klasson, L. Distinctive Genome Reduction Rates Revealed by
489 Genomic Analyses of Two Coxiella-Like Endosymbionts in Ticks. *Genome Biol Evol* **7**,
490 1779–96 (2015).

491 17. van Schaik, E. J., Chen, C., Mertens, K., Weber, M. M. & Samuel, J. E. Molecular
492 pathogenesis of the obligate intracellular bacterium *Coxiella burnetii*. *Nat. Rev.*

493 *Microbiol.* **11**, 561–73 (2013).

494 18. Isberg, R. R., O'Connor, T. J. & Heidtman, M. The Legionella pneumophila replication
495 vacuole: making a cosy niche inside host cells. *Nat. Rev. Microbiol.* **7**, 12–24 (2009).

496 19. Segal, G., Feldman, M. & Zusman, T. The Icm/Dot type-IV secretion systems of
497 Legionella pneumophila and Coxiella burnetii. *FEMS Microbiol. Rev.* **29**, 65–81 (2005).

498 20. Burstein, D. *et al.* Genomic analysis of 38 Legionella species identifies large and diverse
499 effector repertoires. *Nat. Genet.* **48**, 167–75 (2016).

500 21. Gomez-Valero, L. *et al.* More than 18,000 effectors in the Legionella genus genome
501 provide multiple, independent combinations for replication in human cells. *Proc. Natl.*
502 *Acad. Sci.* **116**, 2265–2273 (2019).

503 22. Rihova, J., Novakova, E., Husnik, F. & Hypsa, V. Legionella Becoming a Mutualist:
504 Adaptive Processes Shaping the Genome of Symbiont in the Louse Polyplax serrata.
505 *Genome Biol Evol* **9**, 2946–2957 (2017).

506 23. Csuros, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and
507 likelihood. *Bioinformatics* **26**, 1910–2 (2010).

508 24. Guglielmini, J., de la Cruz, F. & Rocha, E. P. Evolution of conjugation and type IV
509 secretion systems. *Mol. Biol. Evol.* **30**, 315–31 (2013).

510 25. Habyarimana, F. *et al.* Role for the Ankyrin eukaryotic-like genes of Legionella
511 pneumophila in parasitism of protozoan hosts and human macrophages. *Environ.*
512 *Microbiol.* **10**, 1460–1474 (2008).

513 26. Brocks, J. J. & Pearson, A. Building the Biomarker Tree of Life. *Rev. Mineral. Geochem.*
514 **59**, 233–258 (2005).

515 27. Brocks, J. J. & Schaeffer, P. Okenane, a biomarker for purple sulfur bacteria

516 (Chromatiaceae), and other new carotenoid derivatives from the 1640 Ma Barney Creek
517 Formation. *Geochim. Cosmochim. Acta* **72**, 1396–1414 (2008).

518 28. Williams, K. P. *et al.* Phylogeny of gammaproteobacteria. *J. Bacteriol.* **192**, 2305–14
519 (2010).

520 29. Martijn, J. & Ettema, T. J. From archaeon to eukaryote: the evolutionary dark ages of the
521 eukaryotic cell. *Biochem. Soc. Trans.* **41**, 451–7 (2013).

522 30. Santos, P. *et al.* Gamma-proteobacteria Aquicella lusitana gen. nov., sp. nov., and
523 Aquicella siphonis sp. nov. infect protozoa and require activated charcoal for growth in
524 laboratory media. *Appl. Environ. Microbiol.* **69**, 6533–40 (2003).

525 31. Guy, L., Karamata, D., Moreillon, P. & Roten, C.-A. H. Genometrics as an essential tool
526 for the assembly of whole genome sequences: The example of the chromosome of
527 Bifidobacterium longum NCC2705. *BMC Microbiol.* **5**, (2005).

528 32. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–9
529 (2014).

530 33. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
531 identification. *BMC Bioinformatics* **11**, 119 (2010).

532 34. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA
533 genes in nucleotide sequences. *Nucleic Acids Res* **32**, 11–6 (2004).

534 35. Seemann, T. *Barrnap: BAsic Rapid Ribosomal RNA Predictor*. (2013).

535 36. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter.
536 *Nature* **499**, 431–437 (2013).

537 37. Guy, L. phyloSkeleton: taxon selection, data retrieval and marker identification for
538 phylogenomics. *Bioinformatics* **33**, 1230–1232 (2017).

539 38. Krueger, F. *Trim Galore!* (Babraham Bioinformatics, 2017).

540 39. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
541 sequence data. *Bioinformatics* **30**, 2114–20 (2014).

542 40. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to
543 single-cell sequencing. *J. Comput. Biol.* **19**, 455–77 (2012).

544 41. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin
545 outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).

546 42. Delmont, T. O. *et al.* Nitrogen-Fixing Populations Of Planctomycetes And Proteobacteria
547 Are Abundant In The Surface Ocean. *bioRxiv* (2017).

548 43. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
549 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–80 (2013).

550 44. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood
551 trees for large alignments. *PloS One* **5**, e9490 (2010).

552 45. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
553 large phylogenies. *Bioinformatics* **30**, 1312–3 (2014).

554 46. Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures.
555 *Genome Biol.* **10**, R85 (2009).

556 47. Ultsch, A. & Moerchen, F. *ESOM-Maps: tools for clustering, visualization, and*
557 *classification with Emergent SOM*. (2005).

558 48. Hugoson, E., Lam, W. T. & Guy, L. miComplete: weighted quality evaluation of
559 assembled microbial genomes. *Bioinformatics* doi:10.1093/bioinformatics/btz664.

560 49. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new
561 software for selection of phylogenetic informative regions from multiple sequence

562 alignments. *BMC Evol Biol* **10**, 210 (2010).

563 50. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and
564 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol.*
565 *Evol.* **32**, 268–74 (2015).

566 51. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol.*
567 *Evol.* **25**, 1307–20 (2008).

568 52. Quang le, S., Gascuel, O. & Lartillot, N. Empirical profile mixture models for
569 phylogenetic reconstruction. *Bioinformatics* **24**, 2317–23 (2008).

570 53. Wang, H. C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with
571 Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation.
572 *Syst Biol* **67**, 216–235 (2018).

573 54. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2:
574 Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

575 55. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
576 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**,
577 587–589 (2017).

578 56. Rodrigue, N. & Lartillot, N. Site-heterogeneous mutation-selection models within the
579 PhyloBayes-MPI package. *Bioinformatics* **30**, 1020–1 (2014).

580 57. Brocks, J. J. *et al.* Biomarker evidence for green and purple sulphur bacteria in a stratified
581 Palaeoproterozoic sea. *Nature* **437**, 866–870 (2005).

582 58. Takaichi, S. Distribution and Biosynthesis of Carotenoids. in *The Purple Phototrophic*
583 *Bacteria* (eds. Hunter, C. N., Daldal, F., Thurnauer, M. C. & Beatty, J. T.) 97–117
584 (Springer Netherlands, 2009).

585 59. Nordberg, H. *et al.* The genome portal of the Department of Energy Joint Genome
586 Institute: 2014 updates. *Nucleic Acids Res* **42**, D26-31 (2014).

587 60. Vogl, K. & Bryant, D. A. Biosynthesis of the biomarker okenone: chi-ring formation.
588 *Geobiology* **10**, 205–15 (2012).

589 61. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein
590 database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

591 62. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology
592 search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* **41**,
593 e121 (2013).

594 63. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis.
595 *PLoS Comput Biol* **10**, e1003537 (2014).

596 64. Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and
597 dating with confidence. *PLoS Biol* **4**, e88 (2006).

598 65. Page, R. W. & Sweet, I. P. Geochronology of basin phases in the western Mt Isa Inlier,
599 and correlation with the McArthur Basin. *Aust. J. Earth Sci.* **45**, 219–232 (1998).

600 66. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior
601 Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* **67**, 901–904
602 (2018).

603 67. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of Ortholog Groups for
604 Eukaryotic Genomes. *Genome Res.* **13**, 2178–2189 (2003).

605 68. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
606 DIAMOND. *Nat Methods* **12**, 59–60 (2015).

607 69. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale

608 detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).

609 70. Abby, S. S., Neron, B., Menager, H., Touchon, M. & Rocha, E. P. MacSyFinder: a
610 program to mine genomes for molecular systems with an application to CRISPR-Cas
611 systems. *PloS One* **9**, e110726 (2014).

612 71. Guglielmini, J. *et al.* Key components of the eight classes of type IV secretion systems
613 involved in bacterial conjugation or protein secretion. *Nucleic Acids Res* **42**, 5715–27
614 (2014).

615 72. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
616 search tool. *J. Mol. Biol.* **215**, 403–10. (1990).

617 73. Carver, T. J. *et al.* ACT: the Artemis comparison tool. *Bioinformatics* **21**, 3422–3423
618 (2005).

619 74. Guy, L., Kultima, J. R. & Andersson, S. G. E. genoPlotR: Comparative gene and genome
620 visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).

621 75. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated
622 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–3
623 (2009).

624 76. Gast, R. J., Moran, D. M., Dennett, M. R., Wurtsbaugh, W. A. & Amaral-Zettler, L. A.
625 Amoebae and Legionella pneumophila in saline environments. *J. Water Health* **9**, 37–52
626 (2011).

627 77. Giovannoni, S. J. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Annu.*
628 *Rev. Mar. Sci.* **9**, 231–255 (2017).

629 78. Kwak, M.-J. *et al.* Architecture of the type IV coupling protein complex of *Legionella*
630 *pneumophila*. *Nat. Microbiol.* **2**, 17114 (2017).

631 79. Sutherland, M. C., Nguyen, T. L., Tseng, V. & Vogel, J. P. The Legionella IcmSW
632 Complex Directly Interacts with DotL to Mediate Translocation of Adaptor-Dependent
633 Substrates. *PLOS Pathog.* **8**, e1002910 (2012).

634 80. Ghosal, D. *et al.* Molecular architecture, polar targeting and biogenesis of the Legionella
635 Dot/Icm T4SS. *Nat. Microbiol.* **4**, 1173–1182 (2019).

636

637 Acknowledgments

638 This work was supported by the Swedish Research Council [2017-03709 to L.G.], Science for
639 Life Laboratory [SciLifeLab National Project 2015 to L.G.], and the Carl Tryggers
640 Foundation [CTS 15:184, CTS 17:178 to L.G.].

641 We would like to thank Thijs Ettema and Lisa Klasson for constructive discussion;
642 Jennah Dharamshi and Lina Juzokaite for help with the 16S amplicon protocol; Laura Eme
643 for help with dating issues; Joran Martijn and Julian Vos for the help with the Tara Ocean
644 binning. We would also like to thank the Ag1000G project for releasing early their data.

645 Author contributions

646 L.G. conceived the study and drafted the manuscript. E.H. performed database screening and
647 metagenomics; T.A. analyzed the T4BSS and effectors; E.H. and L.G. performed
648 phylogenomics. All authors contributed to writing the final manuscript and approve it.

649 # Supplementary material

650 ## Supplementary discussion

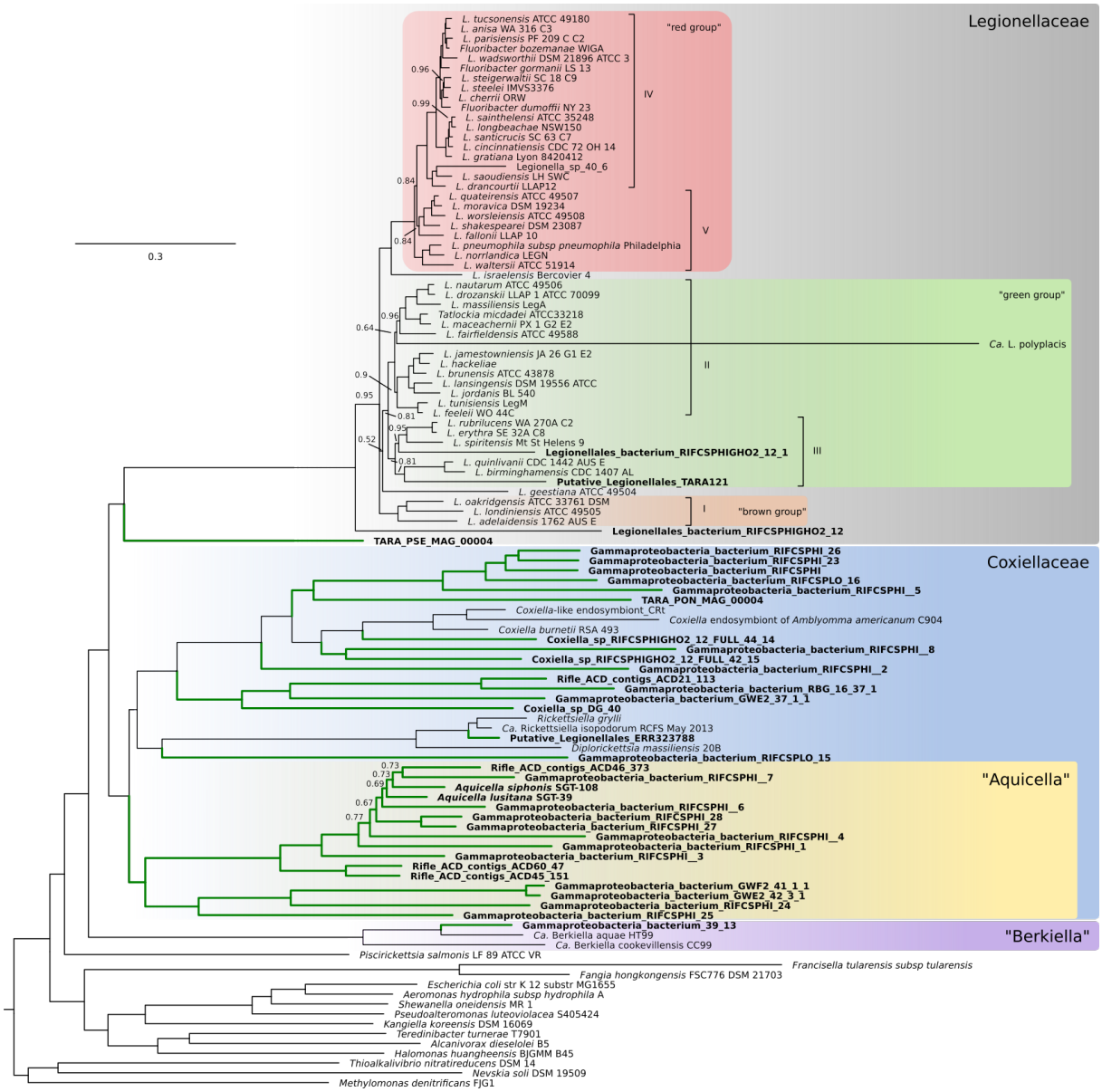
651 ### MAGs in *Legionellaceae*

652 The phylogeny shows a comprehensive, previously unknown diversity among *Legionellales*
653 (**Fig. 1**). The *Coxiellaceae* is divided in two large clades, one including the known genera
654 *Coxiella*, *Rickettsiella* and *Diplorickettsia*, and another ("Aquicella"), about which very little
655 is known, includes the amoebal pathogen *Aquicella* and environmental MAGs. Interestingly,
656 two MAGs recovered from the TARA Ocean sampling campaign are included in the
657 *Legionellaceae*, one as a sister clade to the *Legionella* genus, and one clustering within the
658 "green group", having *L. birminghamensis* and *L. quinlivanii* as closest relatives (**Fig. 1**). The
659 discovery of two *Legionella* genomes in a marine environment supports rising evidence that
660 *Legionella* and other genera of the *Legionellales* can colonize saline waters [14,76]. Strikingly,
661 both MAGs branching as sister groups to the *Legionella* genus have small genomes (2.2 and
662 1.9 Mb, respectively), whereas genomes in the *Legionella* genus (except for *Ca.* L. polyplacis,
663 which is an obligate endosymbiont of a louse [22]) range from 2.3 to 4.9 Mb [21]. This argues in
664 favor of a *Legionellaceae* last common ancestor that had a smaller genome than the extant
665 *Legionella* species. Incidentally, both MAGs contained inside the *Legionella* clade III [21] also
666 display smaller genomes than their closest relatives: 1.9 and 2.3 Mb, whereas clade III
667 *Legionella* spp. range from 3.1 to 3.9 Mb. The same is also true, albeit to a lesser extent, with
668 the MAG branching in clade IV (Legionella_sp_40_6: 3.1 Mb; rest of the clade: 3.4-4.9 Mb).
669 The reduced genome size in these MAGs, associated with the longer branches leading to
670 them, might indicate shifts in lifestyles, either towards streamlining as in the *Pelagibacterales*
671 [77] or towards an increased dependency on their host [13], similarly to *Ca.* Legionella polyplacis
672 [22].

673 **Evolution of the T4BSS**

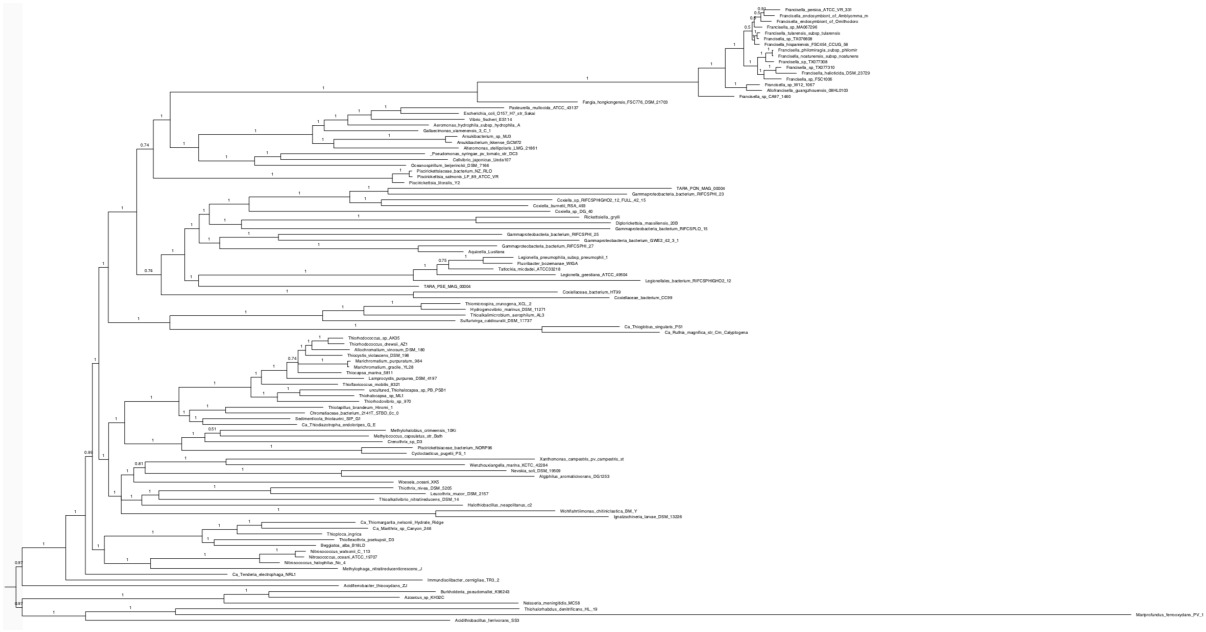
674 IcmW and IcmS (two small acidic cytoplasmic proteins, part of the coupling protein
675 subcomplex [78,79]), IcmC/DotE, IcmD/DotP and IcmV (three small integral inner membrane
676 proteins, of uncharacterized functions) were gained in the LLCA sensu lato. IcmQ, a
677 cytoplasmic protein and IcmN/DotK, an outer membrane lipoprotein [80], were acquired in the
678 LLCA sensu stricto, after the divergence with *Berkiella*. The limited amount of characterized
679 functions for these proteins prevents us to infer exactly how their gain allowed the LLCA to
680 infect eukaryotic cells. However, except for IcmN/DotK, all these proteins are located either
681 in the inner membrane or in the cytoplasm, suggesting that the specific role of the
682 *Legionellales*-gained proteins is related to the coupling protein complex rather than in the
683 core complex.

Supplementary Figures

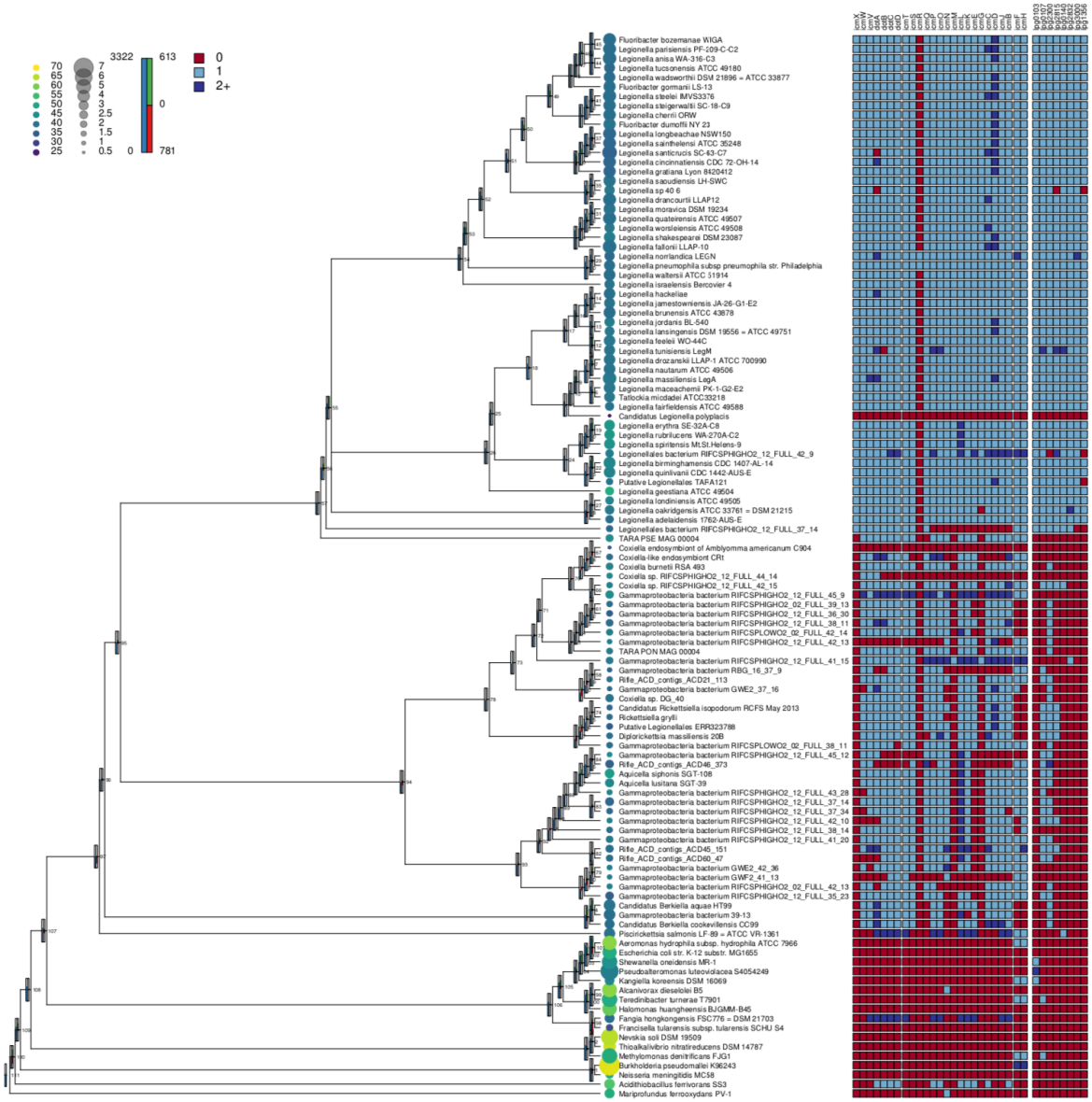


Supplementary Figure 1: Bayesian phylogeny of the order *Legionellales*. Based on a concatenated amino-acid alignment of 105 single-copy orthologs, the tree was inferred with PhyloBayes using a GTR+CAT model. The tree encompasses 93 *Legionellales* genomes and is rooted with an outgroup of 4 genomes from *Proteobacteria* other than *Gammaproteobacteria*. Numbers on branches show posterior probabilities (pp). Branches without number indicate pp = 1. Terminal nodes in bold indicate taxa recovered in this study. Green branches indicate clades where no genomic data was available prior to this study. The

scale shows the average number of substitutions per site. The colored groups and Roman numerals in the *Legionellaceae* correspond to the groupings in [20] and [21], respectively.



Supplementary Figure 2: Bayesian estimation of the phylogenetic relationships of 105 *Gammaproteobacteria*. The tree is inferred from the Gamma105 dataset, which encompasses 110 genomes, including 105 *Gammaproteobacteria* and 5 other *Proteobacteria.* The selected *Gammaproteobacteria* include 22 *Legionellales*, 19 *Chromatiaceae*, 17 *Francisellaceae* and 4 *Piscirickettsiaceae*. Four phylobayes chains were run from the concatenated BMGE-trimmed alignments of 104 protein markers, using a CAT+GTR model. Numbers on the branches represent the posterior probability. The tree was rooted with the five outgroup genomes. This tree was used as an input for the time-constrained tree show on **Fig. 4.**



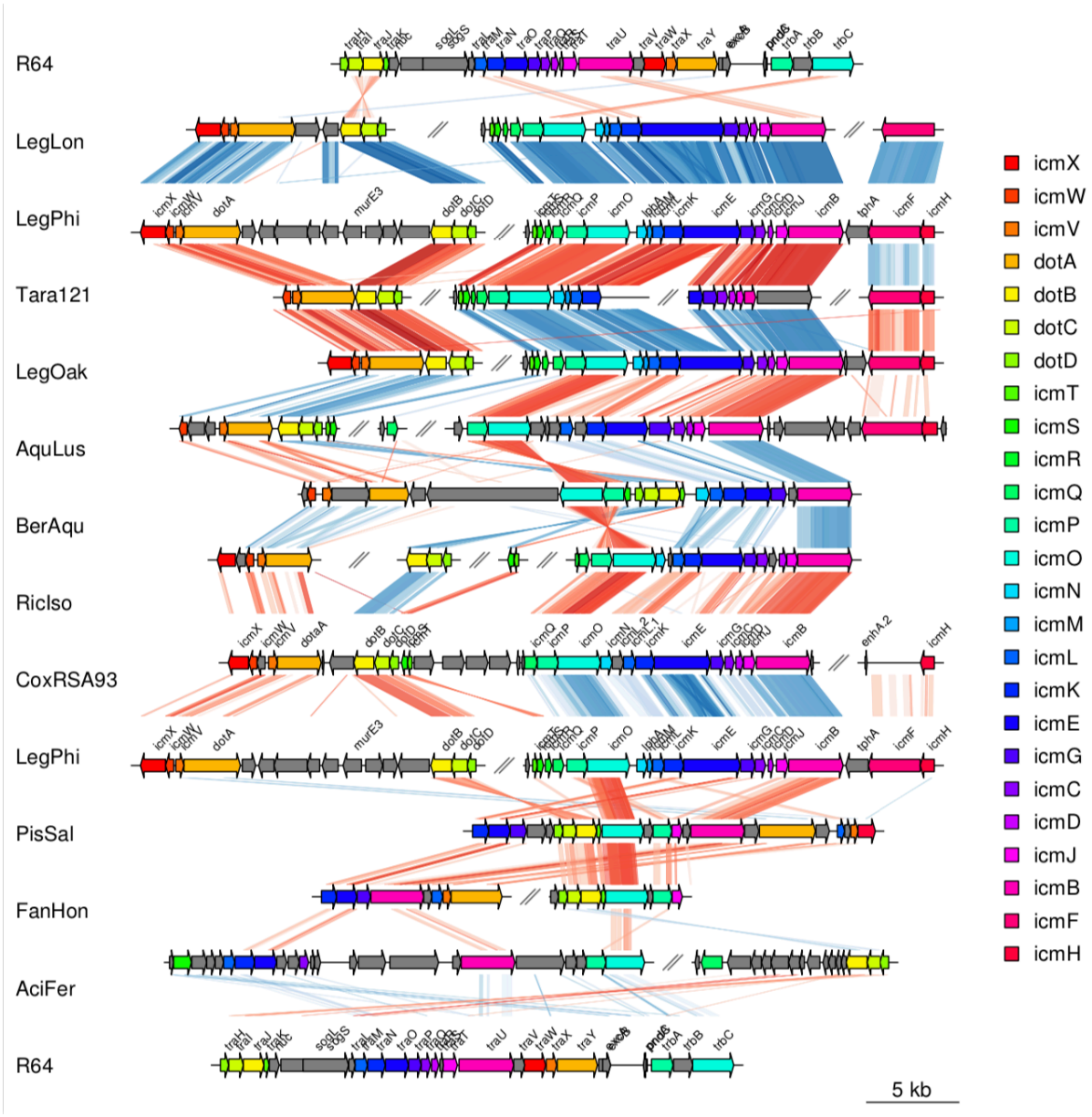
Supplementary Figure 3: Gene flow analysis in the *Legionellales* and other *Gammaproteobacteria* genomes, with focus on the T4BSS system and the *Legionella*-conserved effectors. At the left of each node, barplots depict the number of gene families (blue, ranging from 0 to 3322) inferred in that ancestor, as well as the number of family gains (green, from 0 to 613) and losses (red, from 0 to 781) occurring on the branch leading to that ancestor. To the right of each terminal node, a circle depicts the size of the corresponding genome (proportional to the diameter of the circle) and its G+C content (color of the circle).

On the right panel, squares represent proteins included in the T4BSS (separated according to

714 the three operons present in R64 [19]) and the eight effectors conserved in *Legionella*. The

715 colors represent the number of homologs of each gene in the genome: red, absent; light blue,

716 one copy; dark blue, several copies.



717

718 **Supplementary Figure 4:** Collinearity of the T4BSS in *Legionellales*. Genes are colored as

719 they appear in *Legionella pneumophila* Philadelphia (LegPhi). Areas between the rows

720 display similarities between the sequences, as identified by tblastx. The sequences displayed

721 here in addition are the IncI plasmid R64 from *Salmonella enterica* serovar Typhimurium

722 (R64), *Legionella longbeacheae* (LegLon), a putative Legionellales TARA121 (Tara121),

723 *Berkiella aquae* (BerAqu), *Rickettsiella isopodorum* (RicIso), *Coxiella burnetii* RSA93

724 (CoxRSA93), *Piscirickettsia salmonis* (PisSal), *Fangia hongkongensis* (FanHon) and

725 *Acidithiobacillus ferrivorans* (AciFer).

726

727

728 **Supplementary Figure 5:** Maximum-likelihood phylogeny of the T4BSS in the

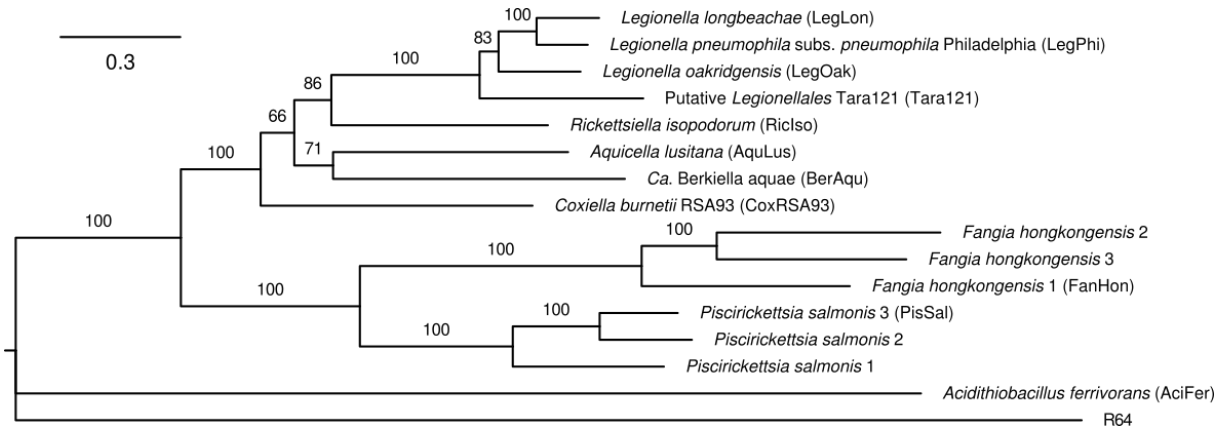
729 *Legionellales*. The phylogeny is based on 12 genes of the T4BSS automatically detected by

730 MacSyFinder. Number on branches represent the percent of bootstrap support. Branches

731 without number indicate 100% bootstrap support. The scale represents the number of

732 substitutions per site. Taxa in bold were selected for manual curation and are represented in

733 **Supplementary Figure 6** and **4**.



734

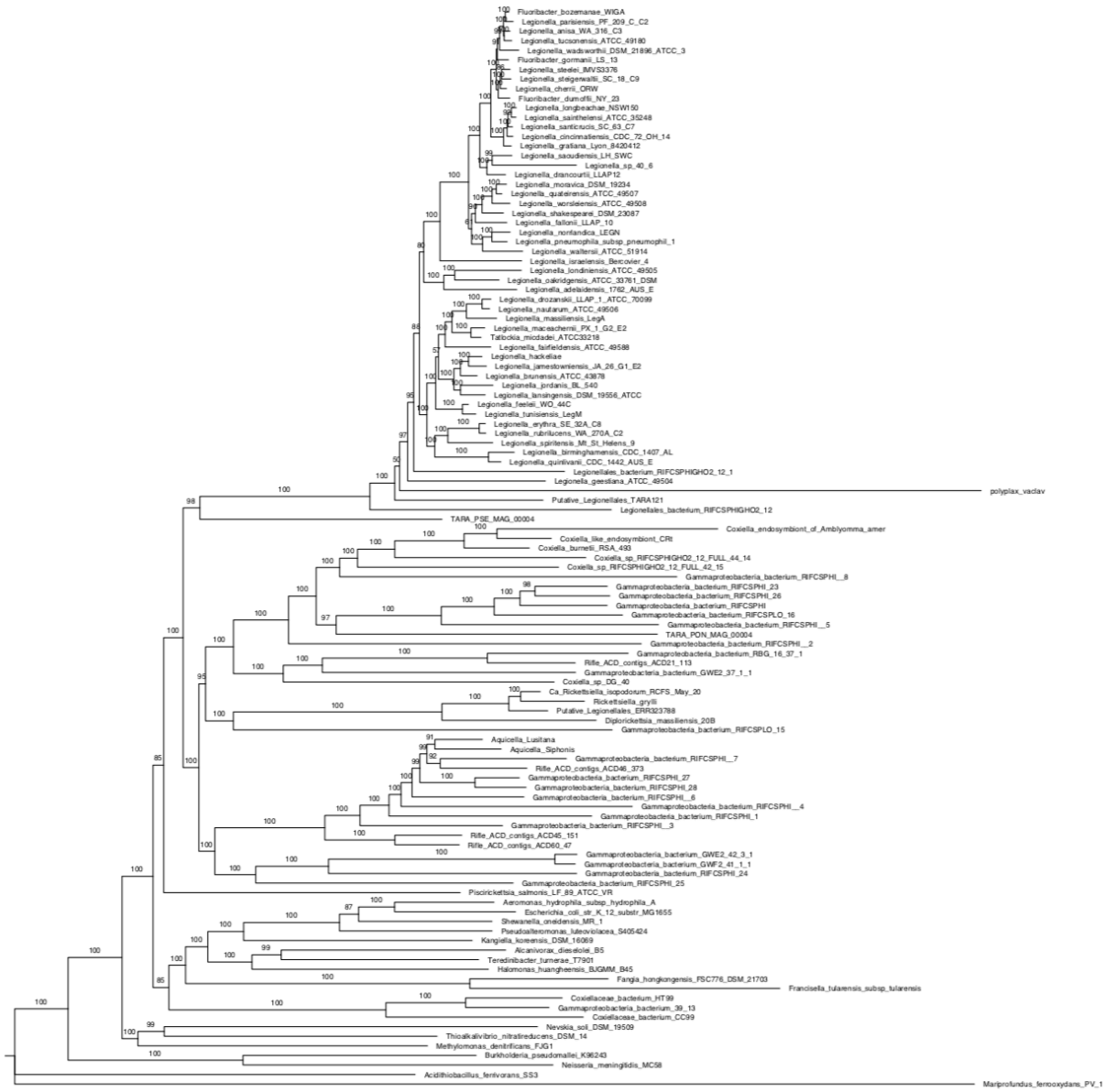
735 **Supplementary Figure 6:** Maximum-likelihood phylogeny of the T4BSS. The phylogeny is

736 based on the concatenation of the alignments of the 25 genes displayed on **Fig. 2** and

737 **Supplementary Figure 4**, manually curated. Numbers on branches indicate the percentage of

738 bootstrap support. The scale represents the number of substitutions per site.

Supplementary Figure 7: Maximum-likelihood tree of 105 *Gammaproteobacteria*. The underlying alignment is the same as in **Supplementary Figure 2**. The tree was inferred with IQ-tree, using the LG substitution matrix, empirical codon frequencies, four gamma categories, the C60 mixture model and PMSF approximation. A thousand UltraFast bootstrap were drawn. The numbers on the branches represent the percentage of bootstrap tree supporting that branch. The scale represents the average number of substitutions per site.



Supplementary Figure 8: Maximum-likelihood tree of the *Legionellales*. The underlying alignment is the same as in **Supplementary Figure 1**. The tree was inferred with IQ-tree, using the LG substitution matrix, empirical codon frequencies, four gamma categories, the C60 mixture model and PMSF approximation. A thousand UltraFast bootstrap were drawn. The numbers on the branches represent the percentage of bootstrap tree supporting that branch. The scale represents the average number of substitutions per site.

753 ## Supplementary tables

754 **Supplementary Table 1:** Summary of the phylogenetic placement of *Berkiella* and
755 *Piscirickettsia* according to different methods.

756 **Supplementary Table 2:** Summary of the age of LLCA in different BEAST runs.

757 **Supplementary Table 3:** Ancestral reconstruction of T4BSS and conserved effectors.

758 **Supplementary Table 4:** Ancestral reconstruction, probabilities of families
759 present/gained/lost at each node.

760 **Supplementary Table 5**: Metagenomes containing high fractions of reads mapping to
761 *Legionellales* 16S rDNA.

762 **Supplementary Table 6**: Gamma105 dataset, genomes used to place *Legionellales* in the
763 *Gammaproteobacteria*.

764 **Supplementary Table 7**: Legio93 dataset, genomes used to reconstruct the *Legionellales*
765 genomic tree.

766

767

768

769