Host-adaptation in *Legionellales* is 2.4 Ga, 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

Bacteria adapting to living in a host cell caused the most salient events in the evolution of eukaryotes, namely the seminal fusion with an archaeon [1], and the emergence of both the mitochondrion and the chloroplast [2]. A bacterial clade that may hold the key to understanding these events is the deep-branching gammaproteobacterial order *Legionellales* – containing among others *Coxiella* and *Legionella* – of which all known members grow inside eukaryotic cells [3]. Here, by analyzing 35 novel *Legionellales* genomes mainly acquired through metagenomics, we show that this group 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 were lost in LLCA and its immediate descendants. We estimate that LLCA lived circa 2.4 Ga ago, predating the last eukaryotic common ancestor (LECA) by at least 0.5 Ga [4]. These elements strongly indicate that host-adaptation arose only once in *Legionellales*, and that these bacteria were using advanced molecular machinery to exploit and manipulate host cells very early in eukaryogenesis.

32 # Introduction

33 The recent discovery of Asgard archaea and their placement on the tree of life [5,6] sheds light
34 on early eukaryogenesis, confirming that eukaryotes arose from a fusion between an archaeon
35 and a bacterium. However, many questions pertaining to eukaryogenesis remain unanswered.
36 In particular, the timing and the nature of the fusion is vigorously debated [1,7–9]. Mito-late
37 scenarios [8,10,11] posit that endosymbiosis of the mitochondrion occurred in a eukaryote that
38 was capable of phagocytosis, while mito-early scenarios [12] posit that mitochondrial
39 endosymbiosis triggered eukaryogenesis. In many scenarios, phagocytosis is considered a
40 prerequisite for mitochondrion endosymbiosis and therefore a key component needed for
41 eukaryogenesis [7]. It is not certain when eukaryotes gained the ability to phagocytose bacteria,
42 but it was most certainly prior to the last eukaryotic common ancestor (LECA) [1,7].

43 The rise of bacteria-phagocytosing eukaryotes created a new ecological opportunity
44 for bacteria, as the eukaryotic cytoplasm is a rich environment. Many bacteria have adapted to
45 exploit this niche, by resisting digestion once phagocytosed by their predators. A host-adapted
46 lifestyle has evolved many times in many different taxonomic groups [13], but there are few
47 large groups comprised solely of host-adapted members; such groups include *Legionellales*,
48 *Chlamydiales*, *Rickettsiales* and *Mycobacteriaceae*. *Legionellales* is a diverse group of
49 *Gammaproteobacteria* [3,14] that lives only intracellularly and include the well-studied
50 accidental human pathogens *Coxiella burnetii* and *Legionella* spp. *Legionellales* species vary
51 greatly in lifestyle [3], ranging from facultatively intracellular (like *Legionella* spp.) to
52 obligatorily intracellular (like the vertically inherited *Coxiella* symbionts of ticks [15,16]).
53 Because of their rarity and fastidious nature, these host-adapted bacteria are underrepresented
54 in genomic databases; individual isolates from only 7 genera have been sequenced out of an
55 estimated > 450 genera [14].

56 Hallmarks of adaptation to intracellular space include smaller population sizes,
57 genome reduction and degradation, and pseudogenization [13]. In addition, the infection process
58 linked to an intracellular lifestyle requires a number of specialized functions. For instance, the
59 Type IVB secretion system (T4BSS) plays a key role in the interaction of *Coxiella* [17] and
60 *Legionella* [18,19] with their hosts, injecting a wide diversity of protein effectors into the host
61 cell. These effectors alter the host behavior, preventing the bacterium from being digested and
62 aiding exploitation of host resources, among others. Some *Legionella* species encode as many
63 as 18 000 types of effectors, but only eight of these are conserved throughout the genus [20,21].

64 To better understand the evolutionary history of *Legionellales* and its relationships
65 with their early hosts, we gathered 35 genomic sequences from novel *Legionellales* through
66 whole-genome sequencing, binning of metagenome-assembled genomes (MAGs), and
67 database mining.

68 Results and Discussion

69 Phylogenetic relationships and diversity among *Legionellales*

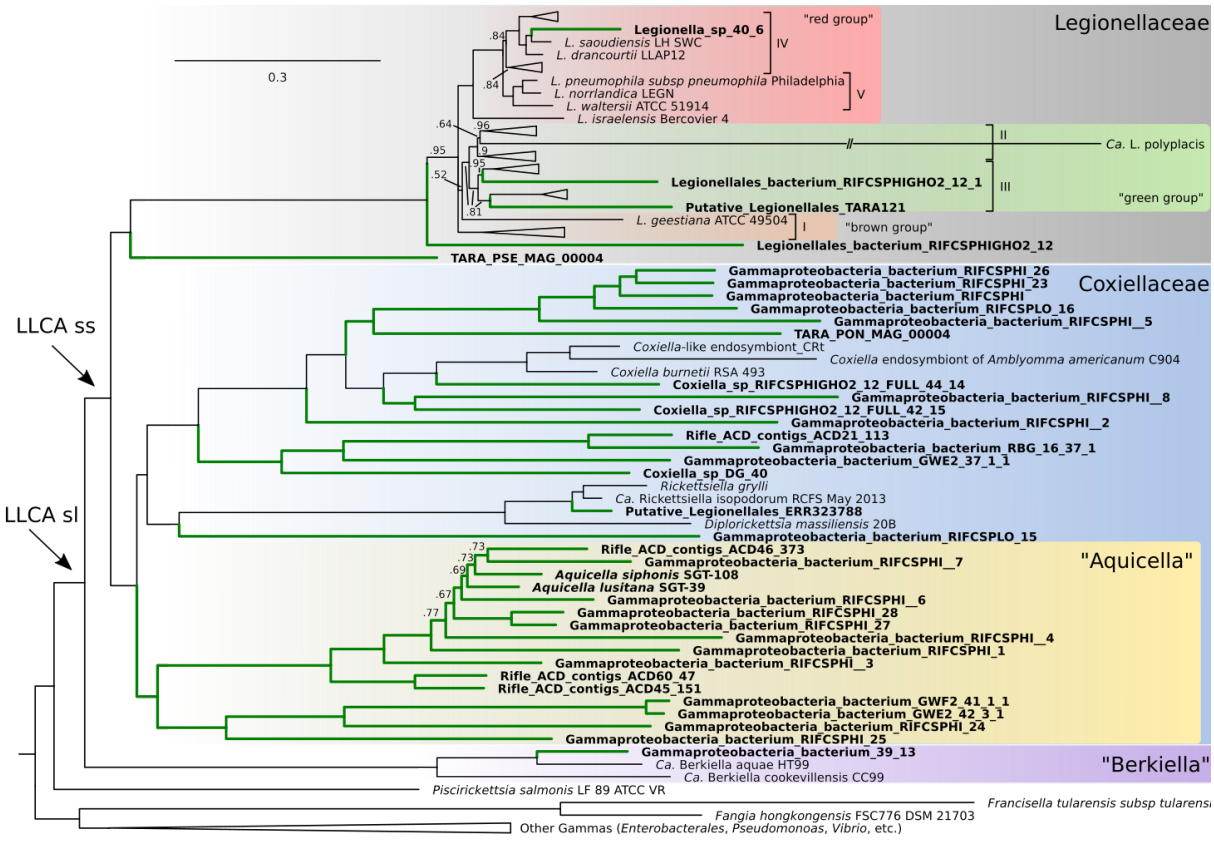
70 Bayesian and maximum-likelihood phylogenies confirm that *Legionellaceae* and
71 *Coxiellaceae* are monophyletic sister clades (hereafter jointly referred as *Legionellales* sensu
72 stricto) are monophyletic, sister clades, and diverged rapidly from each other after the last
73 *Legionellales* common ancestor (LLCA) (**Fig. 1**). The phylogeny reveals extensive,
74 previously unknown diversity among *Legionellales* (**Fig. 1**). *Coxiellaceae* comprises two
75 large clades, one including the known genera *Coxiella*, *Rickettsiella* and *Diplorickettsia*, and
76 the other including the amoebal pathogen *Aquicella*[22] and environmental MAGs. Notably, the
77 two MAGs in *Legionellales* that are resolved as outside *Legionella* have small genomes (2.2
78 and 1.9 Mb, respectively, see **Supp. Figure 3**), whereas genomes within *Legionella* (except

79 for *Ca.* L. polyplacis, an obligate louse endosymbiont [23]) range from 2.3 to 4.9 Mb [21]. This

80 pattern suggests that the last common ancestor of *Legionellaceae* had a smaller genome than

81 extant *Legionella* species, although it cannot be completely excluded that genome reduction

82 occurred independently in these two MAGs.



83

84 **Figure 1 | Bayesian phylogeny of *Legionellales*.** Tree inferred with PhyloBayes using a

85 GTR+CAT model and a concatenated amino acid alignment of 109 single-copy orthologs

86 (Bact109) comprising 93 *Legionellales* genomes, 16 genomes from other

87 *Gammaproteobacteria*, and rooted with an outgroup of 4 genomes from non-

88 *Gammaproteobacteria Proteobacteria* (dataset Legio93). Numbers on branches show

89 posterior probabilities (pp). Branches without numbers indicate pp = 1. Terminal nodes in

90 bold indicate taxa recovered in this study either by database mining or sequencing. Green

91 branches indicate clades where no genomic data was available prior to this study. The scale

92 shows the average number of substitutions per site. The colored groups and Roman numerals

93 in *Legionellaceae* correspond to the groupings in [20] and [21], respectively. The full tree is
94 available in **Supp. Figure 1**.

95 Interestingly, two MAGs recovered from the TARA Ocean sampling campaign are
96 included in the *Legionellaceae*, one as a sister clade to the *Legionella* genus, and one
97 clustering within the "green group", having *L. birminghamensis* and *L. quinlivanii* as closest
98 relatives (**Fig. 1**). The discovery of two *Legionella* genomes in a marine environment supports
99 rising evidence that *Legionella* and other genera of the *Legionellales* can colonize saline
100 waters [14,77]. Strikingly, both MAGs branching as sister groups to the *Legionella* genus have
101 small genomes (2.2 and 1.9 Mb, respectively), whereas genomes in the *Legionella* genus
102 (except for *Ca.* L. polyplacis, which is an obligate endosymbiont of a louse [23]) range from 2.3
103 to 4.9 Mb [21]. This argues in favor of a *Legionellaceae* last common ancestor that had a
104 smaller genome than the extant *Legionella* species. Incidentally, both MAGs contained inside
105 the *Legionella* clade III [21] also display smaller genomes than their closest relatives: 1.9 and
106 2.3 Mb, whereas clade III *Legionella* spp. range from 3.1 to 3.9 Mb. The same is also true,
107 albeit to a lesser extent, with the MAG branching in clade IV (Legionella_sp_40_6: 3.1 Mb;
108 rest of the clade: 3.4-4.9 Mb). The reduced genome size in these MAGs, associated with the
109 longer branches leading to them, might indicate shifts in lifestyles, either towards
110 streamlining as in the *Pelagibacterales* [78] or towards an increased dependency on their host [13],
111 similarly to *Ca.* Legionella polyplacis [23].

112 Two other groups, *Berkiella* spp. and *Piscirickettsia* spp. could not be placed with
113 very high confidence in the tree (**Supp. Table 1**). However, Bayesian phylogenies (**Supp.**
114 **Figure 1 and 2**) inferred with the CAT model consistently placed *Berkiella* as sister clade to
115 *Legionellales* sensu stricto (ss), so *Berkiella* + *Legionellales* is hereafter referred to as
116 *Legionellales* sensu lato (sl). The placement of *Piscirickettsia* is not as consistent, and further

117 studies are needed to establish whether it is more closely related to *Legionellales* or the

118 *Francisella*/*Fangia* group.

119 Evolution of genome content in *Legionellales*

120 To better understand the evolution of host-adaptation in the order, we reconstructed gene flow

121 within *Legionellales*. The phylogenetic birth-and-death model implemented in Count [24] was

122 used on the same set of genomes (Legio93) as the tree in **Fig. 1**. The reconstructed gene flow

123 reveals 553 gene family losses from the last free-living ancestor to LLCA sl (**Fig. 2, Supp.**

124 **Figure 3, Supp. Table 2**). These 553 losses can be compared to 781 losses on the branch

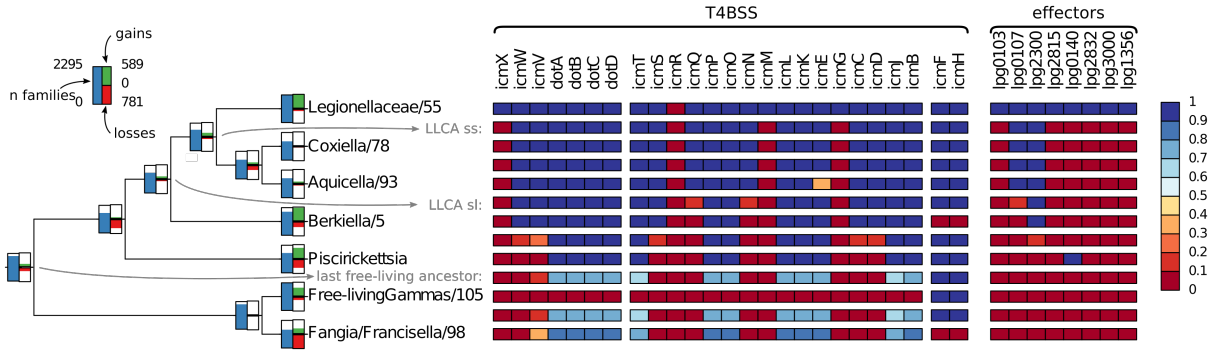
125 leading to *Fangia*/*Francisella*, another intracellular group. A subsequent large gain in gene

126 families in the last common ancestors of *Legionellaceae* and *Berkiella* spp. (588 and 556,

127 respectively) but not in the ancestors of *Coxiellaceae* (*Coxiella*/*Aquicella*) suggests that

128 LLCA had a genome size comparable to that of the latter group (average: 1.8 Mb), compatible

129 with genome sizes of gammaproteobacterial intracellular bacteria [13].



130

131 **Figure 2 | Overview of ancestral reconstruction of *Legionellales* genomes, with the**

132 **T4BSS system and the *Legionella*-conserved effectors detailed.** The ancestral

133 reconstruction is based on the Bayesian tree in **Fig. 1**. At the left of each node, barplots depict

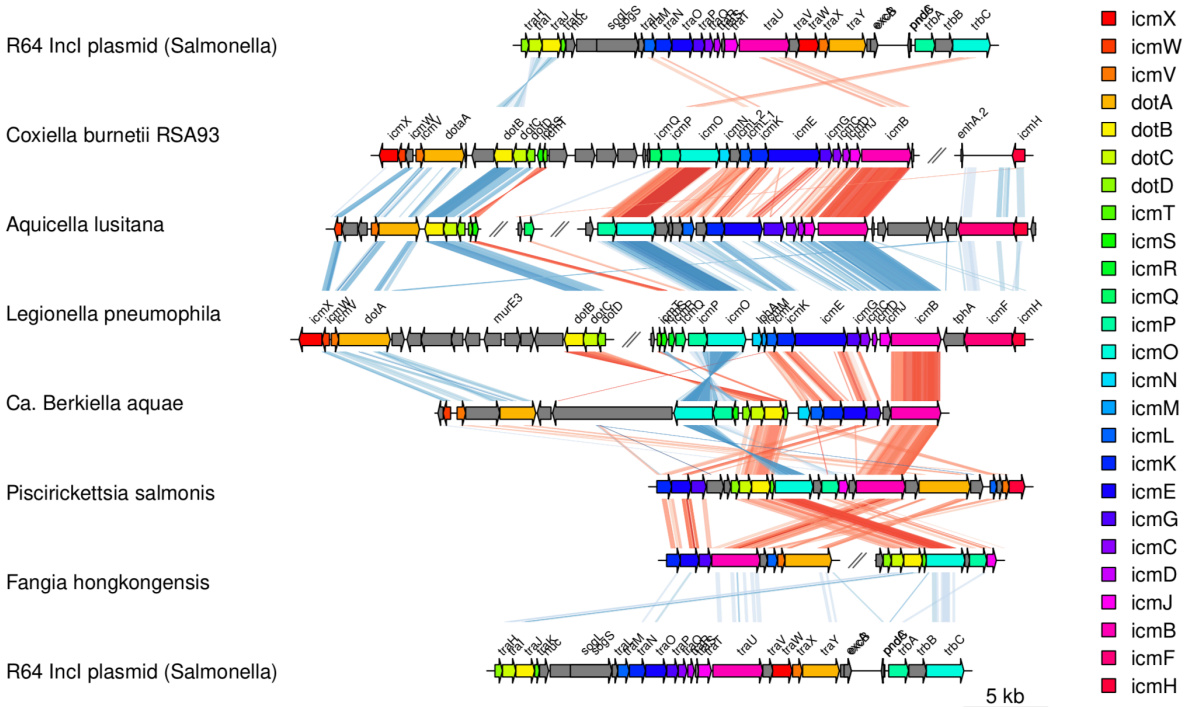
134 the number of gene families (blue, ranging from 0 to 2295) inferred at that node, as well as

135 the number of gene family gains (green, from 0 to 589) and losses (red, from 0 to 781) on the

136 branches leading to nodes. On the right panel, squares represent genes in the T4BSS
137 (separated according to the three operons present in R64 [19]) and the eight effector genes
138 conserved in *Legionella* [21]. Rows in-between terminal nodes correspond to the last common
139 ancestor of the two nearest terminal nodes, e.g. the row between *Aquicella* and *Berkiella*
140 corresponds the LLCA ss. The color of the squares represents the posterior probability that
141 each protein was present in the ancestor of each group. A complete tree is shown in **Supp.**
142 **Figure 3** and the underlying numbers are available in **Supp. Tables 2 and 3**.

143 A crucial feature for the success of many host-adapted bacteria is the ability to transfer
144 proteins into the host's cytoplasm, typically through secretion systems. Among the
145 proteobacterial VirB4-family of ssDNA conjugation systems (Type IV Secretion Systems,
146 T4SS), one called T4BSS (also called MPF_I or I-type T4SS) probably arose early in the group
147 [25]. It is present in almost all extant *Legionellales*, with the exception of the extremely reduced
148 endosymbionts of the group (**Supp. Figure 3**), where it is either missing completely or
149 pseudogenized. It is partly missing in several of the small, novel *Coxiellaceae* MAGs,
150 possibly indicating increased host dependence. However, it is difficult to assess with certainty
151 whether genes absent in MAGs represent true losses or are the result of incomplete binning of
152 metagenomic contigs. T4BSS in *Legionellales* appears largely collinear (**Fig. 3, Supp. Figure**
153 **4**). Ancestral reconstruction of the order revealed that T4BSS, as it is found in extant
154 *Legionellales*, originated after the last free-living ancestor but before the LLCA, with several
155 proteins added over time (**Fig. 2**). IcmW and IcmS (two small acidic cytoplasmic proteins,
156 part of the coupling protein subcomplex [79,80]), IcmC/DotE, IcmD/DotP and IcmV (three small
157 integral inner membrane proteins, of uncharacterized functions) were gained in the LLCA sl.
158 IcmQ, a cytoplasmic protein and IcmN/DotK, an outer membrane lipoprotein [81], were
159 acquired in the LLCA sl stricto, after the divergence with *Berkiella*. The limited amount of
160 characterized functions for these proteins prevents us to infer exactly how their gain allowed

161 the LLCA to infect eukaryotic cells. However, except for IcmN/DotK, all these proteins are
162 located either in the inner membrane or in the cytoplasm, suggesting that the specific role of
163 the *Legionellales*-gained proteins is related to the coupling protein complex rather than in the
164 core complex.



165

166 **Figure 3 | Collinearity of T4BSS in *Legionellales*.** Genes are colored as they appear in
167 *Legionella pneumophila* Philadelphia. Lines connecting rows show similarities between the
168 sequences, as identified by tblastx. Red lines indicate a direct hit, blue lines a complementary
169 one. Darker shades indicate higher bit scores of the tblastx hit.

170 A homologous T4BSS system is present in *Fangia hongkongensis*, but mostly absent
171 in the free-living *Gammaproteobacteria* and in *Francisella*, a sister clade to *Fangia*. It is
172 likely that the T4BSS in *Fangia* has been acquired by horizontal gene transfer, possibly from
173 *Piscirickettsia* (**Supp. Figure 5** and **6**).

174 Contrarily to T4BSS itself, which is highly conserved throughout the order, effectors

175 are much more versatile. Of the eight effectors conserved in the *Legionella* genus [21], only two
176 are found outside of the *Legionellaceae* family (**Fig. 2**): RavC (lpg0107) is found in most
177 *Legionellales* ss, while LegA3/AnkH/AnkW (lpg2300) is found in *Legionellales* ss and
178 *Berkiella*. MavN (lpg2815), which is conserved in *Legionella* spp. and was found in
179 *Rickettsiella* [20], is found only in a few *Coxiellaceae* species (**Fig. 2**; **Supp. Figure 3**), and is
180 unlikely to have been acquired in the LLCA. Although the function of these effectors is not
181 precisely known, LegA3/AnkH contains ankyrin repeats, and has been suggested to play a
182 role in process involved in modulation of phagosome biogenesis [26].

183 Dating the rise of the *Legionellales*

184 Absolute dating of bacterial trees is often inconclusive, because very few reliable biomarkers
185 can be unambiguously attributed to a bacterial clade and used as calibration points [27].
186 However, one such biomarker is okenone, a pigment whose degradation product okenane has
187 been found in the 1.64-Ga-old Barney Creek Formation [28], and is exclusively produced by a
188 subset of *Chromatiaceae*. *Chromatiaceae* (part of the purple sulfur bacteria) and
189 *Legionellales* are phylogenetically close to each other within *Gammaproteobacteria* [29]. We
190 reasoned that the last common ancestor of all okenone-producing *Chromatiaceae* had to be at
191 least as old as the earliest known trace of okenone, i.e. ≥1.64 Ga (**Fig. 4**). Using this
192 constraint and a relaxed molecular clock model, we were able to calibrate a Bayesian tree
193 based on a dataset including genomes from 105 *Gammaproteobacteria* (of which 22 belong
194 *Legionellales* and 19 to *Chromatiaceae*) and 5 outgroups (Gamma105 dataset). We estimate
195 that LLCA existed ca. 2.43 Ga ago (**Fig. 4**), with a 95% high posterior density ranging from
196 2.70 to 2.26 Ga (see **Supp. Table 4**).

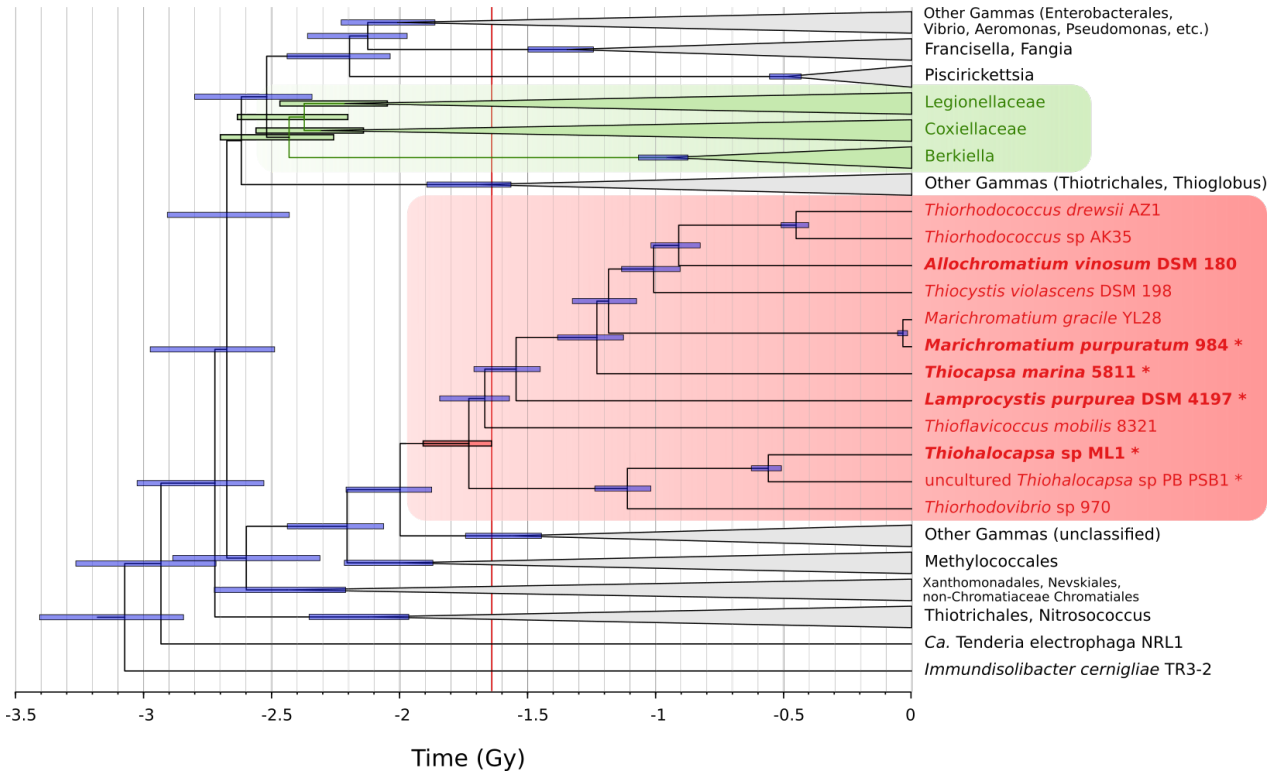


Figure 4 | Bayesian estimation of the appearance of the LLCA. *Chromatiaceae* are highlighted in red, with okenone-producing species in bold. *Legionellales* are highlighted in green. The red vertical line is drawn at 1.64 Ga, the age of the Barney Creek Formation which contains okenane, which gives the lower estimate for the divergence of *Chromatiaceae*. Horizontal bars at interior nodes represent the highest posterior density interval (95%) as determined by BEAST. The red horizontal bar indicates the last common ancestor of all okenone-producing *Chromatiaceae*, the green bars indicate ancestors of *Legionellales*, with, from top to bottom, the last common ancestor of *Legionellaceae*, *Legionellales* ss, *Coxiellaceae*, and *Legionellales* sl (LLCA). The underlying tree is shown in **Supp. Figure 2**.

We took several steps to mitigate confounding factors in estimating the age of LLCA. Firstly, we carefully selected genomes included in the Gamma105 dataset: (i) we included a large outgroup, both in the *Gammaproteobacteria* and outside; (ii) whenever possible, we included genomes that were breaking long branches; (iii) we removed fast-evolving genomes, but kept those breaking long branches.

212 Secondly, we tested several clock models: strict, relaxed with uncorrelated log-normal
213 distribution, and relaxed with random local distribution. The latter was not computationally
214 tractable. The four chains run for the strict clock model (10 million generations each)
215 converged but the strict clock assumption appeared invalid when examining the variation of
216 mutation rates obtained from the relaxed clock model. The results of the strict clock model
217 were therefore not used. For the relaxed clock model (uncorrelated log-normal), five chains
218 were run, one for 55.7 million generations, and four for 10 million generations each. Most of
219 the parameters achieved an ESS score > 100, a good indication that the chains had efficiently
220 sampled the parameter space. The mutation rate and the mean of the ucld parameter did not
221 reach high ESS scores but were negatively correlated with each other. Despite not fully
222 converging, the five trees from the five chains yielded similar results. In particular, the
223 estimation of the age of LLCA was consistent between the chains (see **Supp. Table 4**), which
224 strongly indicates that the chains correctly explored the parameter space. To estimate how the
225 chosen priors affected the runs, the parameters obtained from the actual run (57 million
226 generations) were compared with those from a run with 10 million generations drawing only
227 from the priors. All parameters displayed very different distributions, suggesting that the
228 choice of the priors had very little effect on the obtained values.

229
Conclusion

230 The common ability of *Legionellales* bacteria to invade eukaryotic cells combined with the
231 presence and high conservation of the crucial host-adaptation genes, namely a complete
232 T4BSS and two effectors, strongly suggest that the last common ancestor of *Legionellales*
233 was already infecting other cells. The hypothesis that LLCA used these mechanisms to infect
234 or interact with other prokaryotes cannot be completely ruled out, but seems extremely
235 unlikely for several reasons. For example, if LLCA was only infecting prokaryotes, then the

236 ability to infect eukaryotes would have to have evolved independently several times within
237 *Legionellales*, in all free-living descending clades, with no exception (all experimentally
238 investigated *Legionellales* are able to live inside the cytoplasm of eukaryotic cells).
239 Furthermore, each of these independent host-adaptation events would have to have involved
240 the same host-adaptation genes (all experimentally investigated *Legionellales* use the same
241 T4BSS to interact with their hosts).

242 It is also likely that LLCA, like extant *Legionellales*, depended on phagocytosis (or its
243 prototypic version), a mechanism exclusively found in eukaryotes. This implies that LLCA
244 appeared after the division of Archaea and the first eukaryotic common ancestor (FECA), thus
245 placing rise of phagocytosis and FECA at 2.26 Ga at the latest. Given a consensual estimate
246 of the last eukaryotic common ancestor's (LECA) age of 1.0–1.6 Ga [4], the time for
247 eukaryogenesis, defined as the time elapsed between FECA and LECA [1], would be at least
248 660 Ma, but could be >1.2 Ga.

249 The existence of a phagocytosing eukaryote at 2.26 Ga has implications for
250 hypotheses of eukaryogenesis. Four of the most prominent of these (reviewed e.g. in [9]) make
251 different assumptions about the timing of the mitochondrial endosymbiosis. In the hydrogen
252 hypothesis [12], the mitochondria arrived early ('mito-early'), with the mitochondrial
253 endosymbiosis event itself triggering eukaryogenesis. In the phagocytosing archaeon model
254 (PhAT) [30], the syntrophy hypothesis [10], and the serial endosymbiosis model [8], the
255 mitochondrion arrive late ('mito-late'). Specifically, in the PhAT model, phagocytosis
256 machinery is a pre-requisite for the fusion of the mitochondrial progenitor with an Asgard
257 archaeon. The very early emergence of phagocytosis proposed in this contribution supports a
258 mito-late scenario.

259 In summary, we show here that the last *Legionellales* common ancestor already

260 possessed a T4BSS and two effectors, and existed over 2.26 Ga ago. We propose that LLCA,
261 upon being phagocytosed by eukaryotic cells, already had the ability to resist digestion, owing
262 to its host-adaptation genes. Thus, phagocytosis is at least 2.26 Ga old, earlier than previously
263 thought. This hypothesis is consistent with a scenario in which some early eukaryotes
264 developed phagocytic properties and fed on prokaryotes. Some of these, among them LLCA,
265 rapidly acquired the abilities to resist host digestion and exploit the novel, rich ecological
266 niche that is the eukaryotic cytoplasm.

267 Material and Methods

268 Bacterial strains

269 Two species of *Aquicella* (*A. lusitana*, DSM 16500 and *A. siphonis*, DSM 17428) [22] were
270 ordered from the DSMZ strain culture and grown at 37°C for 5 days on Buffered Charcoal
271 Yeast Extract (BCYE) agar plates with *Legionella* BCYE supplement (OXOID).

272 DNA Extraction and sequencing, assembly and annotation

273 DNA was extracted from pure culture isolates (*Aquicella* spp.) using the Genomic Tip 100
274 DNA extraction kit (QIAGEN), following the manufacturer's instructions. DNA extracted
275 from pure cultures was sequenced by PacBio, using one SMRT cell for each isolate. Reads
276 were assembled using the HGAP.3 pipeline, resulting in one unitig per replicon. The
277 assembly was confirmed by performing cumulative GC skews on each replicon [31]. All
278 replicons had repetitive sequences at the beginning and the end of the sequences, suggesting
279 that they were circular. One unitig in *A. lusitana* had multiple repeats of the same sequences
280 and was edited manually. All chromosomes were rotated to have the origin of replication, as
281 determined by cumulative GC skews, at the beginning of the sequence. The final sequences
282 were confirmed by mapping ~500k shorter reads (2x300 bp, obtained from an Illumina MiSeq

283 run) to the sequences obtained from PacBio. No SNP could be found in any of the replicons.
284 All sequencing, as well as quality control and assembly of the PacBio reads, was performed at
285 NGI, SciLifeLab, Uppsala and Stockholm, Sweden.

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

290 Protein marker sets

291 A set of 139 PFAM protein domains was used as a starting point both to perform quality
292 control of MAGs and for phylogenomic reconstructions. The original set was used by Rinke
293 et al (Supplementary Table 13 in [36]) to estimate the completeness of their MAGs. We used the
294 same set (referred to as Bact139) to estimate MAG completeness. A large subset of the 139
295 protein domains are very frequently located in the same protein in a vast majority of
296 Proteobacteria. This was evidenced by investigating 1083 genomes, using phyloSkeleton 1.1
297 [37] to select one representative per proteobacterial genus and to identify proteins containing the
298 Bact139 domain set. Of these domains which often co-localized in the same protein, only the
299 most widespread one was retained. In total, only 109 domains were used to identify proteins
300 suitable for phylogenomics analysis. This set is referred to as Bact109 and is available in
301 phyloSkeleton 1.1 [37].

302 Metagenome assembly

303 The 86 metagenomes with the highest fraction of reads belonging to *Legionellales* [14] were
304 selected (**Supp. Table 5**) and downloaded from MGnify [38]. Metagenomes sequenced only
305 with 454 (Roche), or that were too complex to be assembled on the computational cluster at

306 our disposition (512 Gb RAM) were not included. Raw reads were downloaded from the
307 European Nucleotide Archive (ENA) and trimmed using Trim Galore (v0.4.2) [39] to remove
308 standard adapter sequences. Additionally, reads were trimmed using Trimmomatic (v0.36) [40]
309 with a sliding window (4:15) to only discard low-quality bases without losing whole reads.
310 Trimmed reads were then assembled using SPAdes (St. Petersburg genome assembler)
311 (v3.7.1) [41], using the --meta option and kmer sizes 21, 33, and 55 for metagenome sizes that
312 did not require running on a HPCC (High Performance Computer Cluster). More complex
313 metagenomes requiring a HPCC were assembled on UPPMAX with a single kmer, 31.
314 Contigs smaller than 500 bp were then discarded.

315 Identification of MAGs belonging to *Legionellales*

316 To identify MAGs belonging to *Legionellales*, the ggkBase (https://ggkbase.berkeley.edu/)
317 and NCBI Genome databases were screened for MAGs attributed to Gammaproteobacteria.
318 MAGs obtained through metagenomic binning from a previous published study [42] and from a
319 prepublication [43] that branched close to *Legionella pneumophila* were also included.

320 The selected MAGs and the assembled metagenomes were screened using the same
321 ribosomal protein (r-protein) pipeline as described in [6], hereafter referred to as RP15. Briefly,
322 this pipeline uses first PSI-BLAST 2.9.0+ to search the given metagenomes and MAGs for 15
323 ribosomal proteins, which are universally located in a single chromosomal locus. Contigs
324 containing at least 8 of the 15 r-proteins are then retained, and referred to as r-contigs.

325 A backbone of known organisms (with complete genomes and established
326 taxonomies) was used to help identifying to which known clades the r-contigs belong. The
327 package phyloSkeleton 1.1 [37] was used to select one representative genome in each family
328 within Gammaproteobacteria, one in each order of the Beta- and Alphaproteobacteria, as well
329 as *Mariprofundus* and *Acidithiobacillus*. phyloSkeleton also identified orthologs of 15

330 ribosomal proteins in the gathered genomes.

331 The RP15 orthologs gathered from both the metagenomic sources and the backbone
332 were aligned with MAFFT L-INS-I v7.273 [44], and the alignments concatenated through the
333 RP15 pipeline. The resulting concatenated alignment was used to infer a rough phylogenetic
334 tree using FastTree 2.1.8 [45], using the WAG model. R-contigs that were clearly unrelated to
335 *Legionellales* were removed from the dataset, while the remaining proteins were realigned
336 and concatenated, to generate a new tree. This process was iterated until all remaining MAGs
337 and r-contigs were deemed belonging to *Legionellales*. A final phylogeny was inferred with
338 RAxML 8.2.8 [46] using the PROTGAMMA model which allowed automatic selection of best
339 substitution matrix to infer a more robust phylogeny.

340 Metagenomic binning and quality control

341 Metagenomes containing r-contigs belonging to *Legionellales* were then binned with ESOM
342 1.1 [47,48]. An ESOM map was generated from contigs over 2 kb with a 10 kb sliding window,
343 and visualized using the ESOM analyzer, highlighting the location of the r-contig and
344 selecting the corresponding bin. The completeness and redundancy of all MAGs were then
345 controlled using miComplete 0.1 [49]

346 Some MAGs were very similar to each other, due to different versions of the same bin
347 being present in databases. The most complete MAG was then selected for each set, provided
348 they were not more redundant than another genome within the same set. If equal in
349 completeness the NCBI-submitted MAG was preferred.

350 Phylogenomics

351 Two sets of genomes were used for phylogenomics in this contribution. Both sets consist of
352 part or all of the novel *Legionellales* MAGs and genomes, supplemented with varying subsets

353 of outgroups. The initial selection of the representative organisms and identification of
354 markers was done with phyloSkeleton 1.1 [37], initially choosing one representative per class in
355 the *Betaproteobacteria*, the *Zetaproteobacteria* and the *Acidithiobacillia*; one per genus in the
356 Gammaproteobacteria, except in the *Piscirickettsiaceae* and the *Francisellaceae* (one per
357 species). That initial selection of genomes was then trimmed down to reduce the
358 computational burden of tree inference.

359 For both sets, protein markers from the Bact109 set were identified with
360 phyloSkeleton 1.1 [37]. Each marker was aligned separately with mafft-linsi v7.273 [44]. The
361 resulting alignments were trimmed with BMGE v1.12 [50], using the BLOSUM30 matrix and
362 the stationary-based trimming algorithm. The alignments were then concatenated. From these
363 alignments, maximum-likelihood trees were reconstructed with FastTreeMP v2.1.10 [45] using
364 the WAG substitution matrix. Upon visual inspection of the resulting trees, the genome
365 selection was reduced to remove closely related outgroups, and the phylogenomic procedure
366 repeated. Once the genome dataset was final, a maximum-likelihood tree was inferred with
367 IQ-TREE v1.6.5 [51], using the LG [52] substitution matrix, empirical codon frequencies, four
368 gamma categories, the C60 mixture model [53] and PMSF approximation [54]; 1000 ultrafast
369 bootstraps were drawn [55].

370 The first set of genomes, called Gamma105, was used to correctly place *Legionellales*
371 in *Gammaproteobacteria*, in particular with respect to *Chromatiaceae*. It encompasses 110
372 genomes, of which 105 are *Gammaproteobacteria*, one is a *Zetaproteobacterium*, 3 are
373 *Betaproteobacteria*, and one is an *Acidithiobacillia*. The selected *Gammaproteobacteria*
374 include representatives from *Legionellales* (22), *Chromatiaceae* (19), *Francisellaceae* (17)
375 and *Piscirickettsiaceae* (4) (**Supp. Table 6**).

376 The second set, called Legio93, is focused on *Legionellales* itself and comprises 113

377 genomes, of which 93 belong to *Legionellales* and the rest to other *Gammaproteobacteria* (16
378 genomes), *Betaproteobacteria* (2 genomes), *Acidithobacillia* (1 genome) and
379 *Zetaproteobacteria* (1 genome) (**Supp. Table 7**).

380 For both sets, single-gene maximum-likelihood trees were inferred for each marker.
381 The BMGE-trimmed alignments were used to infer a tree using IQ-TREE 1.6.5, using the
382 automatic model finder [56], limiting the matrices to be tested to the LG matrix and the C10 and
383 C20 mixture models. Single-gene trees were visually inspected for the presence of very long
384 branches resulting from distant paralogs being chosen by phyloSkeleton.

385 For both sets, a Bayesian phylogeny was inferred with phylobayes MPI 1.5a [57] from
386 the concatenated BMGE-trimmed alignments, using a CAT+GTR model. Four parallel chains
387 were run for 9500 generations (Gamma105 set) and 3500 generations (Legio93 set),
388 respectively. The chains did not converge in either set. For Legio93, all four chains yielded
389 the same topology for all deep nodes in and around *Legionellales*. For Gamma105, 3 out of 4
390 chains yielded the same overall topology, while the last one had the *Piscirickettsia* and the
391 *Berkiella* clades inverted (**Supp. Table 1**). The majority-rule tree obtained from the Bayesian
392 trees for Legio93 was subsequently used for the ancestral reconstruction (see below), while
393 the one for Gamma105 was used as input tree to estimate the time of divergence of the LLCA
394 (see below).

395 Identification of okenone-producing *Chromatiaceae*

396 To relate the earliest documented trace of okenone [28,58] to a specific ancestor, we took two
397 approaches. First, we compared a list of okenone-producing *Chromatiaceae* [59] with the
398 genomes available in Genbank and at the JGI Genome Portal [60]. We found five okenone-
399 producing *Chromatiaceae* genomes. Second, we used homologs of the proteins encoded by
400 the genes *crtU* and *crtY*, which are considered essential to synthesize okenone [61], to screen the

401 nr database. We used the proteins from *Thiodictyon syntrophicum* str. Cad16 (accession
402 number: CrtU, AEO72326.1; CrtY, AEO72327.1) as queries in a PSI-BLAST 2.9.0+ search [62]
403 restricted to the order *Chromatiales*, keeping only proteins with at least 50% similarity.
404 Gathered homologs (n=7 in both cases) were aligned with MAFFT L-INS-I v7.273 [44] and the
405 alignment was pressed into a hidden Markov model (HMM) with HMMer 3.1b2 [63]. Using
406 phyloSkeleton [37], we collected all *Chromatiales* genomes from NCBI (n=130) and queried
407 these, together with the Gamma105 set with both HMMs. A significant similarity (E-value <
408 10^{-10}) for CrtU and CrtY was found in 42 and 25 genomes, respectively. A hit for both
409 proteins was found in 8 genomes only, of which 6 belong to the *Chromatiaceae*. Of these 6, 4
410 were already identified by the first method listed above; one was a MAG that was too
411 incomplete to include in further phylogenomics analysis; the last one was another MAG that
412 was included in later analysis. The other two genome that had proteins similar to CrtU and
413 CrtY are *Kushneria aurantia* DSM 21353, a member of the *Halomonadaceae* and
414 *Immundisolibacter cernigliae* TR3-2, which branches very close to the root of the
415 *Gammaproteobacteria*: these two might represent cases of horizontal gene transfer.

416 Time-constrained tree

417 To estimate the time of divergence of LLCA, we used BEAST 2 [64], using a strict and a
418 relaxed clock model (uncorrelated log-normal) [65]. To reduce computational load, we used a
419 fixed input tree (Bayesian, Gamma105; see above). The following parameters were used: the
420 concatenated alignment (104 proteins) was used as a single partition; for site model, a LG
421 matrix with 4 gamma categories and 10% invariant was used, estimating all parameters; for
422 the relaxed clock model, an uncorrelated log-normal model was used, estimating the clock
423 rate. The following prior distributions were used: shape of the gamma distributions for the site
424 model: exponential; mutation rate: 1/x; proportion of invariants: uniform; ucld mean: 1/X,

425 offset: 0, ucld std: gamma. To calibrate the clock, it was assumed that the last common
426 ancestor of all okenone-producing *Chromatiaceae* was at least as old as the rocks where the
427 earliest traces of okenane (a degradation product of okenone) was found, 1.640 ± 0.03 Ga old
428 [28,58,66]. A prior was thus set for that the divergence time of the monophyletic group
429 comprising these organisms: the divergence time was drawn from an exponential distribution
430 of mean 0.1 Ga with an offset of 1.64 Ga. Although it cannot be completely excluded, no
431 evidence suggests that other organisms had produced okenone earlier than the last common
432 ancestor of *Chromatiaceae*: the only mentions of okenone in the scientific literature are linked
433 to the presence of *Chromatiaceae*.

434 The four chains run for the strict clock model (10 million generations each) converged
435 but the strict clock assumption seemed unrealistic when examining the variation of mutation
436 rates obtained from the relaxed clock model. For the relaxed clock model, five chains were
437 run, one for 55.7 million generations, and four for 10 million generations each. Traces were
438 examined with Tracer 1.6.0 [67]. Most of the parameters achieved an ESS score > 100. The
439 mutation rate and the mean of the ucld parameter did not reach very high ESS scores but were
440 negatively correlated with each other. Trees were extracted with TreeAnnotator 2.4.7 (part of
441 BEAST), deciding the burn-in fraction after visual examination of the traces (see **Supp.**
442 **Table 4**). Despite not fully converging, the five trees from the five chains yielded similar
443 results.

444 To estimate how the chosen priors affected the runs, a run with 10 million generations
445 was performed, drawing from the priors. Parameters were compared in Tracer 1.6.0. All
446 parameters displayed very different distributions, suggesting that the choice of the priors had
447 very little effect on the obtained values.

448 ## Protein family clustering and annotation

449 All 113 proteomes from the Legio93 set were clustered into protein families using OrthoMCL
450 2.0.9 [68]. All proteins were aligned to each other with the blastp variant of DIAMOND
451 v0.9.8.109 [69], with the --more-sensitive option, enabling masking of low-complexity regions
452 (--masking 1), retrieving at most 10^5 target sequences (-k 100000), with a E-value threshold
453 of 10^{-5} (-e 1e-5), and a pre-calculated database size (--dbsize 98280075). In the clustering
454 step, mcl 14-137 [70] was called with the inflation parameter equal to 1.5 (-I 1.5), as
455 recommended by OrthoMCL.

456 Annotation of the protein families obtained by OrthoMCL was performed by
457 searching for protein accession number in a list of reference genomes (see below). If any
458 protein in a family was present in the first genome, the family was attributed this annotation;
459 else, the second genome was searched for any of the accession numbers, and so on. The
460 reference list was as follow (Genbank accession numbers between parentheses): *Legionella*
461 *pneumophila* subsp. *pneumophila* strain Philadelphia 1 (AE017354), *Legionella longbeachae*
462 NSW150 (FN650140), *Legionella oakridgensis* ATCC 33761 (CP004006), *Legionella*
463 *fallonii* LLAP-10 (LN614827), *Legionella hackeliae* (LN681225), *Tatlockia micdadei*
464 ATCC33218 (LN614830), *Coxiella burnetii* RSA 493 (AE016828), *Coxiella* endosymbiont
465 of *Amblyomma americanum* (CP007541), *Escherichia coli* str. K-12 substr. MG1655
466 (U00096), *Francisella tularensis* subsp. *tularensis* (AJ749949), *Piscirickettsia salmonis* LF-
467 89 = ATCC VR-1361 (CP011849), *Acidithiobacillus ferrivorans* SS3 (CP011849), *Neisseria*
468 *meningitidis* MC58 (AE002098), *Rickettsiella grylli* (NZ_AAQJ00000000.2). *Legionella*
469 effector orthologous groups were identified by comparing protein accession numbers with the
470 list provided by [20].

471 ## Ancestral reconstruction

472 The flow of genes in the order *Legionellales* was analyzed by reconstructing the genomes of
473 the ancestors. The ancestral reconstruction was performed with Count v10.04 [24], which
474 implements a maximum-likelihood phylogenetic birth- and death-model. The Bayesian input
475 tree was obtained from the Legio93 dataset (see above) and the OrthoMCL families as
476 described above. The rates (gain, loss, and duplication) were first optimized on the
477 OrthoMCL protein families, to allow different rates on all branches. The family size
478 distribution at the root was set to be Poisson. All rates and the edge length were drawn from a
479 single Gamma category, allowing 100 rounds of optimization, with a 0.1 likelihood threshold.
480 Reconstruction of gene flow was then performed by using posterior probabilities.

481 ## Analysis of the Type IVB secretion system

482 Homologs of the Type IVB secretion system were identified with MacSyFinder v1.0.3 [71]. We
483 used the pT4SSi profile, corresponding to the Type IV Secretion System encoded by the IncI
484 R64 plasmid and *Legionella pneumophila* (T4BSS; also referred to as MPF_I) [72], as query to
485 search all proteomes in the Gamma105 and the Legio93 datasets. This profile includes models
486 for the 17 proteins of the IncI R64 plasmid that have a homolog in *L. pneumophila*. The
487 profile does not cover 6 of the Dot/Icm proteins shared among *L. pneumophila* and *C. burnetii*
488 (IcmFHNQSW). MacSyFinder was run in "ordered_replicon" mode, and using the following
489 HMMer options: --coverage-profile 0.33, --i-evalue-select 0.001. A T4BSS was identified if,
490 within 30 neighboring genes: (i) a IcmB/DotU/VirB4 homolog was present, (ii) at least 6 out
491 of the other 16 were also present, (ii) no relaxase (MOBB) were identified. Since the Icm/Dot
492 system is often split in several operons, not all proteins could be identified by MacSyFinder in
493 "ordered_replicon" mode, but most were since the operon which contains IcmB/DotU/VirB4
494 is the largest.

495 To inspect the gene organization further in a selection of organisms, a subset of 12
496 representative genomes were aligned with tblastx 2.2.30 [73], with an E-value threshold of 0.01,
497 not filtering the sequences for low complexity regions, and splitting the genomes in smaller
498 pieces when reaching the maximum number of hits between two segments. Results were first
499 visualized in ACT [74], and a final version of the comparison was drawn with genoPlotR [75]. In
500 the cases where MacSyFinder failed to detect a T4BSS, a similar visual inspection of the
501 comparison of the genome with both *L. pneumophila* and *C. burnetii* was performed.

502 For the genomes where the T4BSS was detected by MacSyFinder, twelve proteins
503 (IcmBCDEGKLOPT, DotAC) were retrieved, aligned with mafft-linsi v7.273 [44], trimmed
504 with trimAl v1.4.rev15 [76], using the -automated1 setting, and concatenated. The concatenated
505 alignment was used to infer a maximum-likelihood tree with IQ-TREE v1.6.8 [51], forcing the
506 use of the LG [52] substitution matrix, but letting ModelFinder [56] select the across-site
507 categories, which resulted in the C60 mixture model [53]; 1000 ultrafast bootstraps were drawn
508 [55]. For the genomes not belonging to *Legionellales* sl (R64, *Fangia hongkongensis*,
509 *Acidithiobacillus ferrivorans*, *Acidiferrobacter thiooxydans* and *Alteromonas stellipolaris*),
510 only a few proteins were retrieved by the automated approach. A preliminary tree displayed
511 very long branches, and the procedure above was repeated without these four genomes. For
512 the 12 genomes for which the T4BSS was manually annotated, a tree based on the alignment
513 of 26 proteins belonging to the T4BSS was performed as above, resulting in a LG+F+I+G4
514 model being selected.

515 Data availability

516 The raw and assembled data for the *Aquicella* genomes is deposited at the European
517 Nucleotide Archive (ENA) under study accession PRJEB29684. The assemblies for *A.*
518 *lusitana* and *A. siphonis* have the accessions GCA_902459475 (replicons LR699114.1-

519 LR699118.1) and GCA_902459485 (replicons LR699119.1- LR699120.1).

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

522 Acknowledgments

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

526 We would like to thank Thijs Ettema and Lisa Klasson for constructive discussion;
527 Jennah Dharamshi and Lina Juzokaite for help with the 16S amplicon protocol; Laura Eme
528 for help with dating issues; Joran Martijn and Julian Vosseberg for the help with the Tara
529 Ocean binning; and Jennifer Ast for her help with editing the manuscript. We would also like
530 to thank the Ag1000G project for releasing early their data.

531 Author contributions

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

535 References

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

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

539 3. Duron, O., Doublet, P., Vavre, F. & Bouchon, D. The Importance of Revisiting

540 Legionellales Diversity. *Trends Parasitol.* **34**, 1027–1037 (2018).

541 4. Eme, L., Sharpe, S. C., Brown, M. W. & Roger, A. J. On the age of eukaryotes:
542 evaluating evidence from fossils and molecular clocks. *Cold Spring Harb Perspect Biol* **6**,
543 (2014).

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

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

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

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

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

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

556 11. Spang, A. *et al.* Proposal of the reverse flow model for the origin of the eukaryotic cell
557 based on comparative analyses of Asgard archaeal metabolism. *Nat. Microbiol.* **4**, 1138–
558 1148 (2019).

559 12. Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**,
560 37–41 (1998).

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

563 14. Graells, T., Ishak, H., Larsson, M. & Guy, L. The all-intracellular order Legionellales is
564 unexpectedly diverse, globally distributed and lowly abundant. *FEMS Microbiol. Ecol.*
565 **94**, (2018).

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

568 16. Gottlieb, Y., Lalzar, I. & Klasson, L. Distinctive Genome Reduction Rates Revealed by
569 Genomic Analyses of Two Coxiella-Like Endosymbionts in Ticks. *Genome Biol Evol* **7**,
570 1779–96 (2015).

571 17. van Schaik, E. J., Chen, C., Mertens, K., Weber, M. M. & Samuel, J. E. Molecular
572 pathogenesis of the obligate intracellular bacterium *Coxiella burnetii*. *Nat. Rev.*
573 *Microbiol.* **11**, 561–73 (2013).

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

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

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

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

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

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

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

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

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

596 27. Brocks, J. J. & Pearson, A. Building the Biomarker Tree of Life. *Rev. Mineral. Geochem.*
597 **59**, 233–258 (2005).

598 28. Brocks, J. J. & Schaeffer, P. Okenane, a biomarker for purple sulfur bacteria
599 (Chromatiaceae), and other new carotenoid derivatives from the 1640 Ma Barney Creek
600 Formation. *Geochim. Cosmochim. Acta* **72**, 1396–1414 (2008).

601 29. Williams, K. P. *et al.* Phylogeny of gammaproteobacteria. *J. Bacteriol.* **192**, 2305–14
602 (2010).

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

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

608 32. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–9

609 (2014).

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

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

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

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

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

619 38. Mitchell, A. L. *et al.* EBI Metagenomics in 2017: enriching the analysis of microbial
620 communities, from sequence reads to assemblies. *Nucleic Acids Res* **46**, D726–D735
621 (2018).

622 39. Krueger, F. *Trim Galore!* (Babraham Bioinformatics, 2017).

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

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

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

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

631 44. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:

632 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–80 (2013).

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

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

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

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

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

643 50. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new
644 software for selection of phylogenetic informative regions from multiple sequence
645 alignments. *BMC Evol Biol* **10**, 210 (2010).

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

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

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

653 54. Wang, H. C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with
654 Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation.

655 *Syst Biol* **67**, 216–235 (2018).

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

658 56. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
659 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**,
660 587–589 (2017).

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

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

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

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

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

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

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

677 64. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis.

678 *PLoS Comput Biol* **10**, e1003537 (2014).

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

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

683 67. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior
684 Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* **67**, 901–904
685 (2018).

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

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

690 70. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale
691 detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).

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

695 72. Guglielmini, J. *et al.* Key components of the eight classes of type IV secretion systems
696 involved in bacterial conjugation or protein secretion. *Nucleic Acids Res* **42**, 5715–27
697 (2014).

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

700 74. Carver, T. J. *et al.* ACT: the Artemis comparison tool. *Bioinformatics* **21**, 3422–3423

701 (2005).

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

704 76. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated
705 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–3
706 (2009).

707 77. Gast, R. J., Moran, D. M., Dennett, M. R., Wurtsbaugh, W. A. & Amaral-Zettler, L. A.
708 Amoebae and Legionella pneumophila in saline environments. *J. Water Health* **9**, 37–52
709 (2011).

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

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

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

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

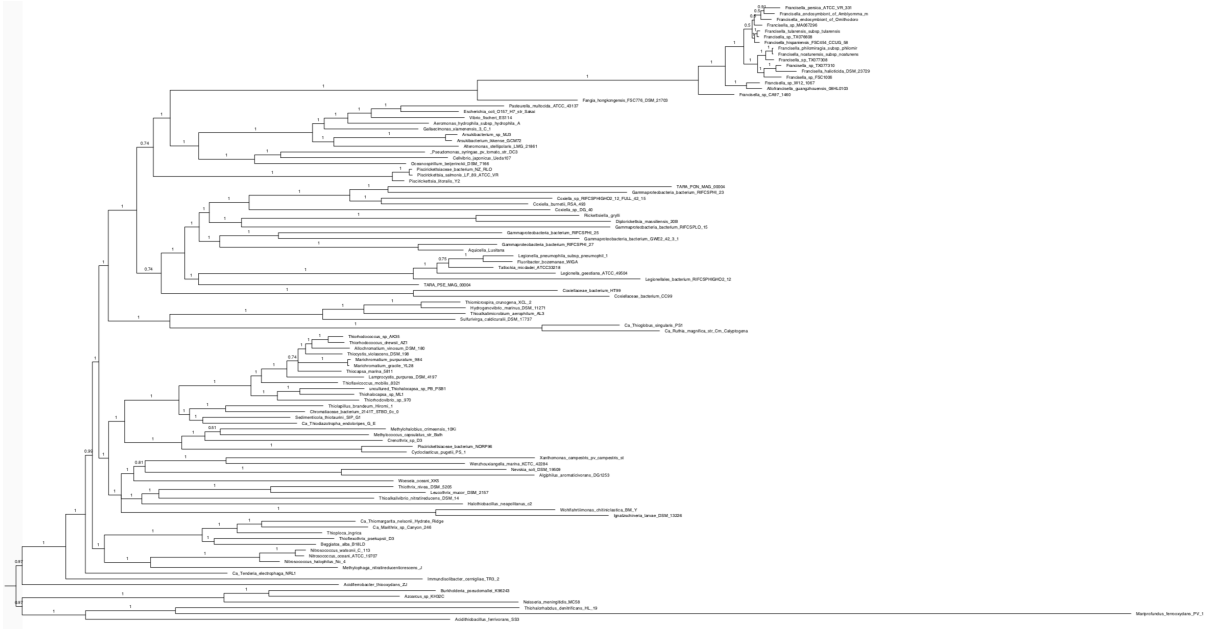
719

Supplementary material

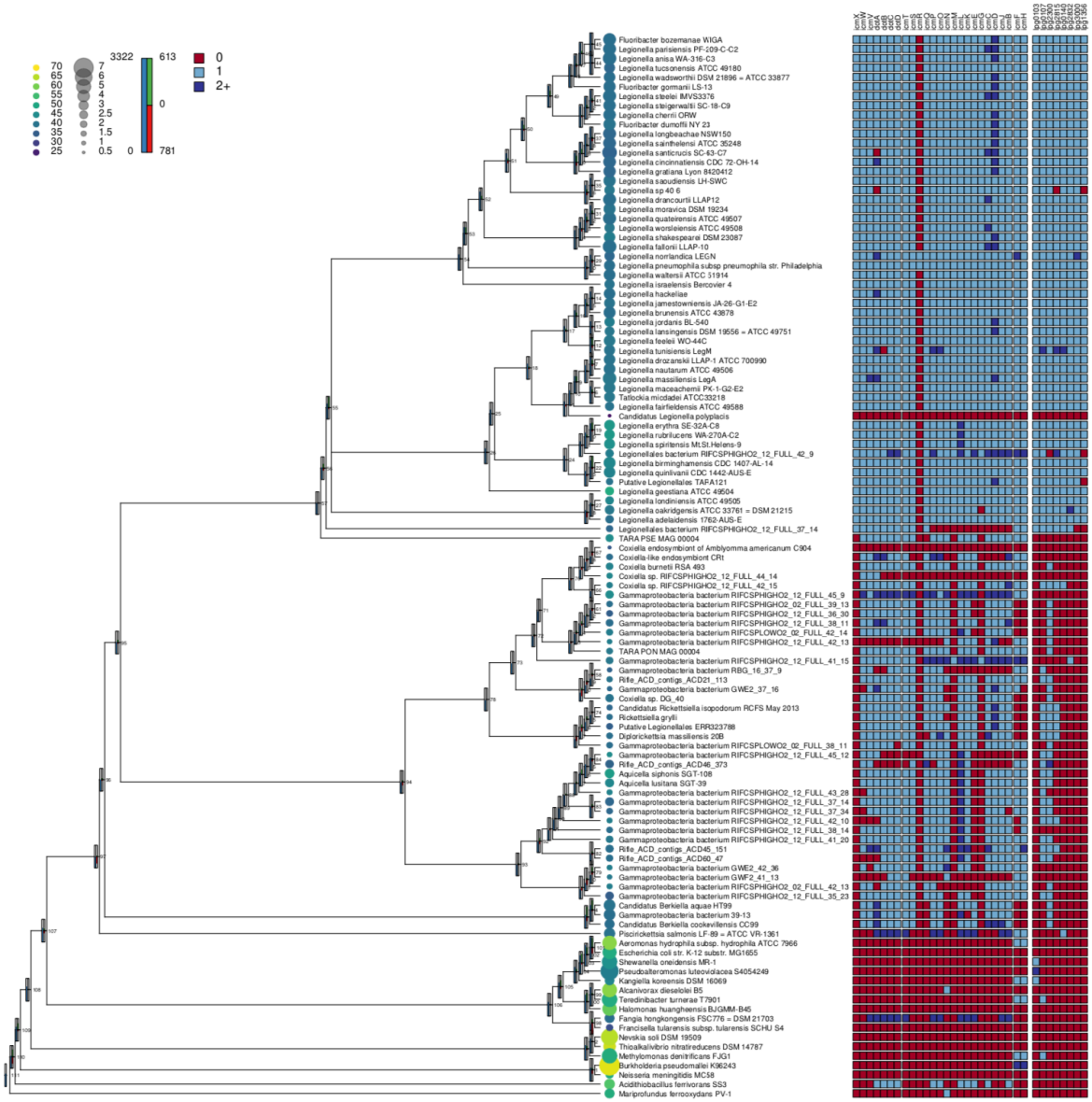
Supplementary Figures

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.**

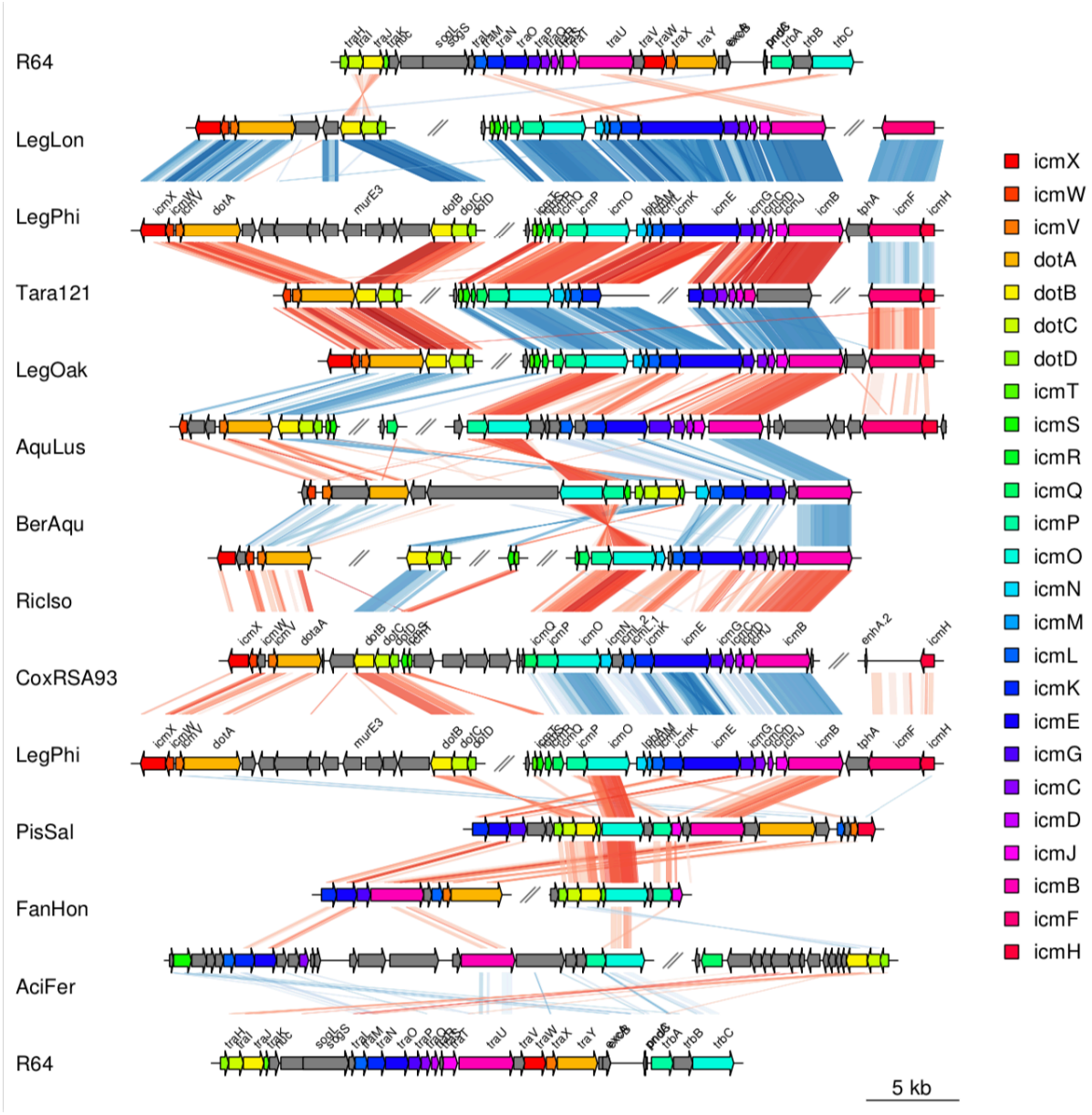


Supplementary Figure 3: Gene flow analysis in the *Legionellales* and other *Gammaproteobacteria* genomes, with focus on the T4BSS system and the *Legionella*-conserved effectors. At the left of each node, barplots depict the number of gene families (blue, ranging from 0 to 3322) inferred in that ancestor, 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

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

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

753 one copy; dark blue, several copies.



754

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

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

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

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

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

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

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

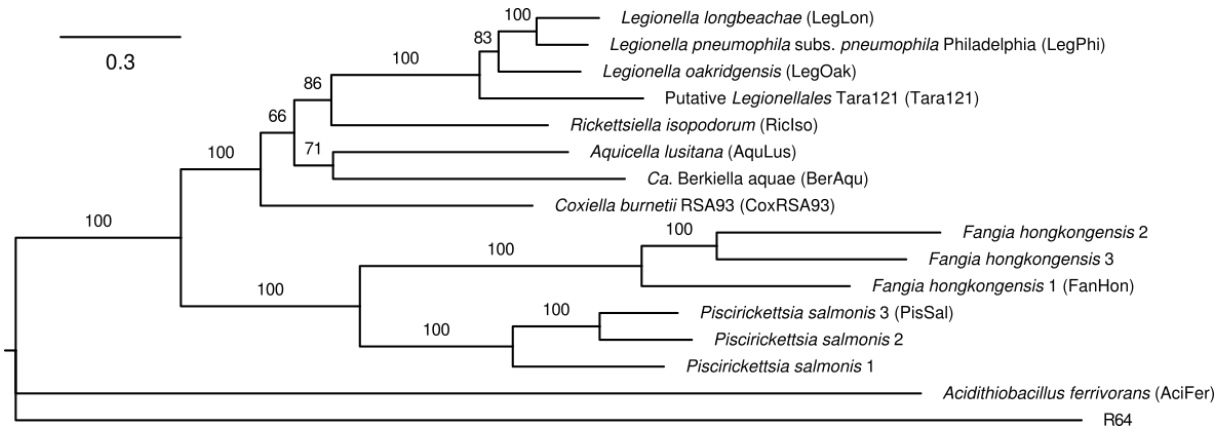
762 *Acidithiobacillus ferrivorans* (AciFer).

763

764

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

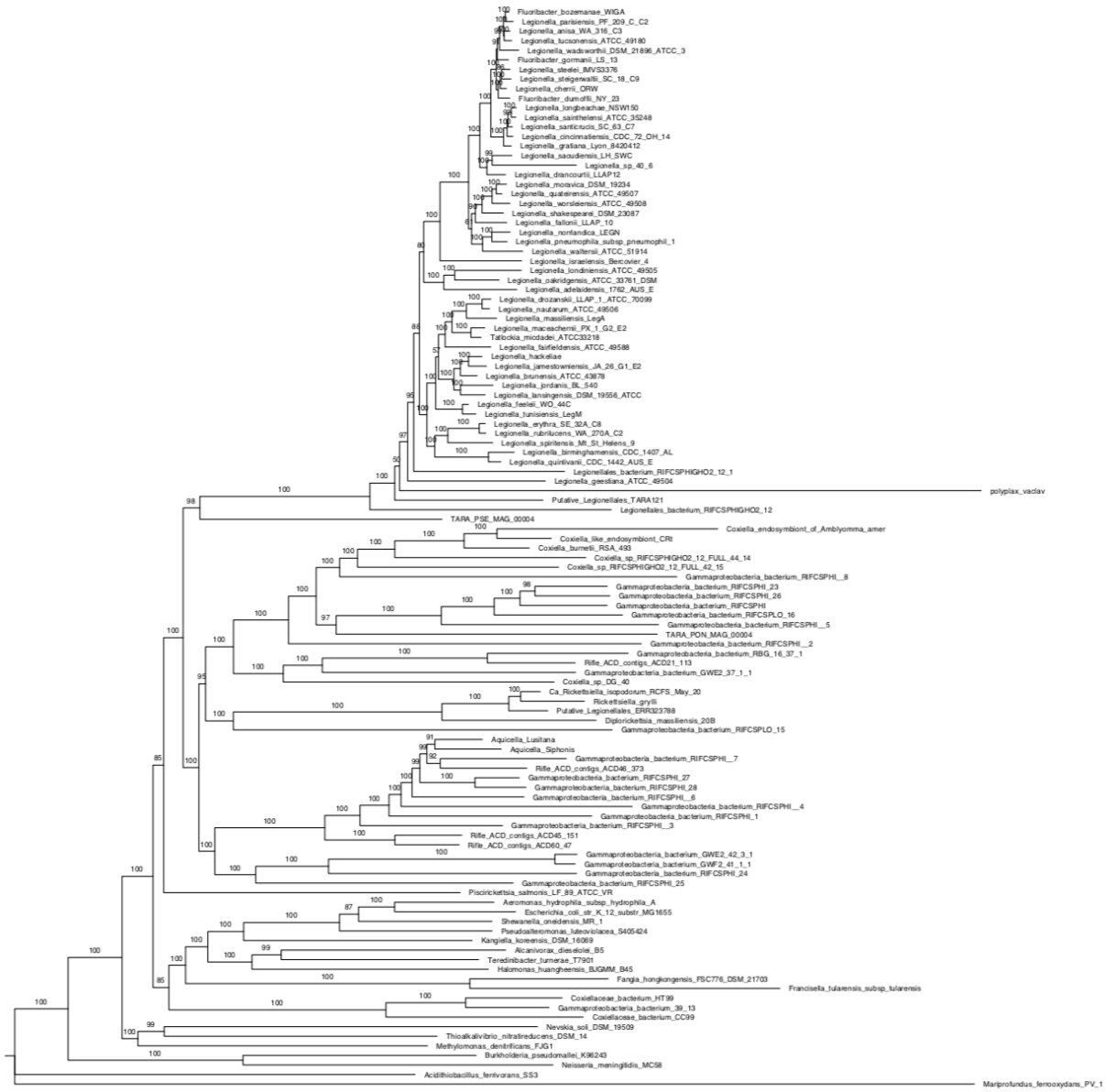
766 *Legionellales*. The phylogeny is based on 12 genes of the T4BSS automatically detected by
767 MacSyFinder. Number on branches represent the percent of bootstrap support. Branches
768 without number indicate 100% bootstrap support. The scale represents the number of
769 substitutions per site. Taxa in bold were selected for manual curation and are represented in
770 **Supp. Figure 6** and **4**.



771

772 **Supplementary Figure 6:** Maximum-likelihood phylogeny of the T4BSS. The phylogeny is
773 based on the concatenation of the alignments of the 25 genes displayed on **Fig. 2** and **Supp.**
774 **Figure 4**, manually curated. Numbers on branches indicate the percentage of bootstrap
775 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 **Supp. 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 **Supp. 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.

790 # Supplementary tables

791 ## Supplementary Table 1.

792 Summary of the phylogenetic placement of *Berkiella* and *Piscirickettsia* according to

793 different methods.

		Berkiella		***Piscirickettsia***		
Method	**Set**	**Sister clade**	**Support ***	**Sister clade**	**Support ***	**Full tree**
PhyloBayes, CAT+GTR	Legio93	*Legionellales* ss	1	*Legionellales* ss + *Berkiella*	1	Fig S1
	Gamma105	*Legionellales* ss	0.76 †	NBG + *Francisella*	0.76 †	Fig S2
IQ-TREE, LG+C60+PMSF	Legio93	NBG + *Francisella* ‡	85	*Legionellales* ss	85	Fig S8
	Gamma105	NBG + *Francisella*	99	*Legionellales* ss	99	Fig S7

* : support is expressed as posterior probability for Bayesian trees and % bootstrap support for maximum likelihood trees
†: in both cases, 3 out of 4 chains support the branch with pp = 1; the last chain supports the opposite (*Piscirikettsia* with *Legionellales*; *Berkiella* with *Gammaproteobacteria*)
‡: NBG, non-basal *Gammaproteobacteria* (*Vibrionales*, *Enterobacteriales*, *Pseudomonadales*, etc).

794

795 ## Supplementary Table 2 (Separate file).

796 Ancestral reconstruction of the *Legionellales*. Output from Count. For each taxon and

797 ancestor, the table shows the estimated probabilities of families present, present in multiple

798 copies (multi) gained, lost, expanded and contracted at each node. Ancestor IDs (numbers) are

799 shown on the tree on **Supp. Figure 3**.

800 ## Supplementary Table 3 (Separate file).

801 Ancestral reconstruction of the genes in T4BSS and conserved effectors in *Legionellales*. For

802 each of these genes (columns) and each of the genomes and ancestors (rows), the table shows

803 the number of copies, the probability of presence, presence in multiple copies, gain, loss,

804 expansion and contractions (in different blocks of rows). Ancestor IDs (numbers) are shown

805 on the tree on **Supp. Figure 3**.

806

807 **Supplementary Table 4.**

808 Summary of the age of LLCA in different BEAST runs.

Chain number	Generations (M)	Burn-in (%)	LLCA Age (Ga)	95% credible interval – lower bound (GA)	95% credible interval – higher bound (GA)	*Chromatiaceae* LCA Age (Ga)
0	55.7	25	2.4315	2.2569	2.7001	1.7295
1	10.0	20	2.2483	2.0757	2.506	1.7299
2	10.0	25	2.2816	2.1249	2.5345	1.7304
3	10.0	25	2.4669	2.29	2.7393	1.7292
4	10.0	50	2.3331	2.1554	2.5967	1.729

809

810

811 **Supplementary Table 5**: Metagenomes containing high fractions of reads mapping to

812 *Legionellales* 16S rDNA.

Accession
ERR598943
SRR1185414
ERR589611
SRR1202095
ERR481104
ERR346687
ERR346672
ERR346678
ERR346689
ERR346671
SRR350864
ERR695596
SRR350867
ERR272377
ERR571345
ERR249384
ERR532617
ERR571347
SRR1209976
SRR1199270
SRR1202089
ERR476939
ERR481106
ERR272378
ERR650528
SRR066138
SRR062218
ERR059353
ERR430972
SRR062168
SRR512582
SRR000905
ERR059351
ERR317255
SRR062149
ERR059347
ERR323780
ERR059350
SRR062243
ERR071289
SRR062159
SRR035082
ERR358544
SRR512809
ERR589374
ERR323707
ERR589381
SRR062130
ERR149036
ERR599003
ERR738547
ERR598952
ERR599098

ERR599063
ERR599016
ERR358549
ERR598950
ERR598945
ERR738548
ERR738549
ERR599113
ERR598997
ERR599104
ERR599137
ERR505057
ERR599118
ERR257715
ERR505054
ERR315863
ERR505056
ERR599032
ERR599050
ERR317264
ERR599139
ERR864075
ERR599163
ERR599038
ERR599081
ERR599065
ERR599095
ERR272375
ERR599052
ERR598978
ERR599077
ERR598992
ERR599023

Supplementary Table 6 (Separate file).

Gamma105 dataset, genomes used to place *Legionellales* in the *Gammaproteobacteria*.

Supplementary Table 7 (Separate file).

Legio93 dataset, genomes used to reconstruct the *Legionellales* genomic tree.