Host Adaptation in *Legionellales* Is 1.9 Ga, Coincident with Eukaryogenesis

Eric Hugoson,†,1,2 Andrei Guliaev,†,1 Tea Ammunét,†,1 and Lionel Guy ‡,*,1

1Department of Medical Biochemistry and Microbiology, Science for Life Laboratories, Uppsala University, Uppsala, Sweden
2Department of Microbial Population Biology, Max Planck Institute for Evolutionary Biology, Plön, Germany

*Corresponding author: E-mail: lionel.guy@imbim.uu.se.
†These authors contributed equally to this work.
‡Present address: Medical Bioinformatics Centre, Turku Bioscience, University of Turku, Turku, Finland

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, and the emergence of both mitochondrion and chloroplast. 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. 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). Many metabolic gene families were lost in LLCA and its immediate descendants, including functions directly and indirectly related to molybdenum metabolism. On the other hand, genome sizes increased in the ancestors of the *Legionella* genus. We estimate that LLCA lived approximately 1.89 Ga, probably predating the last eukaryotic common ancestor by approximately 0.4–1.0 Gy. 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 early in eukaryogenesis.

Key words: *Legionella*, eukaryogenesis, phlyogenomics, metagenomics, host adaptation.

Introduction

The recent discovery of Asgard archaea and their placement on the tree of life (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017) sheds light on early eukaryogenesis, confirming that eukaryotes arose from a fusion between an archaeon and a bacterium. However, many questions pertaining to eukaryogenesis remain unanswered. In particular, the timing and the nature of the fusion are vigorously debated (Poole and Gribaldo 2014; Pittis and Gabaldón 2016; Eme et al. 2017; Lópe-García et al. 2017). Mito-late scenarios (López-García and Moreira 2006; Pittis and Gabaldón 2016; Spang et al. 2019) posit that endosymbiosis of the mitochondrion occurred in a eukaryote that was capable of phagocytosis, whereas mito-early scenarios (Martin and Müller 1998; Martin et al. 2017) posit that mitochondrial endosymbiosis triggered eukaryogenesis. A recent study proposed a “mito-intermediate” scenario, where the pre-endosymbiotic eukaryote was somewhat complex, with a dynamic cytoskeleton and a membrane trafficking system, but the endomembrane system, and the transcription regulation and signaling systems occurred postendosymbiosis (Vosseberg et al. 2021). In many scenarios including the latter, phagocytosis is considered a prerequisite for mitochondrion endosymbiosis and therefore a key component needed for eukaryogenesis (Poole and Gribaldo 2014). It is not certain when eukaryotes gained the ability to phagocytose bacteria, but it was most certainly prior to the last eukaryotic common ancestor (LECA) (Poole and Gribaldo 2014; Eme et al. 2017).

The rise of bacteria-phagocytosing eukaryotes created a new ecological opportunity for bacteria, as the eukaryotic cytoplasm is a nutrient-rich environment. Many bacteria have adapted to exploit this niche, by resisting digestion once phagocytosed by their predators. Host-adapted lifestyles have evolved many times in many different taxonomic groups (Toft and Andersson 2010), but there are few large groups comprised solely of host-adapted members; such groups include *Legionellales*, *Chlamydiales*, *Rickettsiales*, and *Mycobacteriaceae*. *Legionellales* is a diverse group of Gammaproteobacteria (Durón et al. 2018; Graells et al. 2018) that lives only intracellularly and includes the well-studied accidental human pathogens *Coxiella burnetii* and *Legionella* spp. *Legionellales* species vary greatly in lifestyle (Durón et al. 2018), ranging from facultative intracellular (like *Legionella* spp.) to obligate intracellular. The most...
reduced genomes in Legionellales include the vertically inherited Coxiella symbionts of ticks (Gottlieb et al. 2015; Guizzo et al. 2017), a louse symbiont in the genus Legionella with a 5-fold reduced genome (Rihova et al. 2017), and a protist endosymbiont performing respiration and providing energy to its host (Graf et al. 2021). Because of their rarity and fastidious nature, these host-adapted bacteria are underrepresented in genomic databases; individual isolates from only seven genera have been sequenced out of an estimated >450 genera (Graells et al. 2018).

Hallmarks of adaptation to intracellular space include smaller population sizes, genome reduction and degradation, and pseudogenization (Toft and Andersson 2010). In addition, the infection process linked to an intracellular lifestyle requires a number of specialized functions. For instance, the Type IVB secretion system (T4BSS) plays a key role in the interaction of Coxiella (van Schaik et al. 2013) and Legionella (Segal et al. 2005; Isberg et al. 2009) with their hosts, injecting a wide diversity of protein effectors into the host cell. These effectors alter the host behavior, preventing the bacterium from being digested and aiding exploitation of host resources, among others. The Legionella genus pangenome contains as many as 18,000 different effectors, but only eight of these are conserved throughout the genus (Burstein et al. 2016; Gomez-Valero, Rusniok, et al. 2019).

To better understand the evolutionary history of Legionellales and its relationships with its early hosts, we gathered 35 genome sequences from novel Legionellales through whole-genome sequencing, binning of metagenome-assembled genomes (MAGs), and database mining. However, Bayesian phylogenies (supplementary figs. 1 and 2, Supplementary Material online) inferred with the CAT model consistently placed Berkiella as sister clade to Legionellales sensu stricto (ss), so Berkiella + Legionellales is hereafter referred to as Legionellales sensu lato (sl). The placement of Piscirickettsia is not as consistent, and further studies are needed to establish whether it is more closely related to Legionellales or the Francisella/Fangia group.

**Evolution of Genome Sizes**

To better understand the evolution of host-adaptation in the order, we inferred ancestral genome sizes from the reconstructed number of protein families (see below), using extant genomes to calibrate the correlation (supplementary fig. 3, Supplementary Material online). As expected from extant genome sizes, Coxiellaceae have the smallest genome size average (1.68 Mb), Aquicella were intermediate (1.92 Mb), and Legionella and Berkiella were larger, roughly double as large (3.42 and 3.40 Mb, respectively). Among deep ancestors, the genome size drops between the last free-living ancestor (LFLA, node 107, 2.71 Mb) and the next ancestor (last Legionellales/Piscirickettsia common ancestor or LLPCA, node 97, 2.16 Mb) toward Legionellales, and remains stable in the two subsequent ancestors (LLCA ss, node 96, 2.19 Mb and LLCA ss, node 95, 2.28 Mb). A stable trend is also seen in the Coxiellaceae, where the LCA of the family (node 94, 2.09 Mb), and the two daughter nodes (LCA of Aquicella, node 93, 2.11 Mb; LCA of Coxiella, node 78, 2.05 Mb) have similar genome sizes. In Legionellaceae, on the other hand, the trend is toward larger genome sizes (LCA of Legionellales, node 57, 2.24 Mb; the two subsequent nodes, 56, 2.34 Mb, and especially node 55, 2.98 Mb). The two MAGs branching as sister groups to the Legionella genus (TARA_PSE_MAG_00004 and RIFCSPHIGHO2_12_FULL_37_14), have small genomes (2.2 and 1.9 Mb, respectively, see supplementary fig. 2, Supplementary Material online), whereas genomes within Legionella (except for "Ca. L. polyplacis," an obligate louse endosymbiont; Rihova et al. 2017) range from 2.3 to 4.9 Mb (Gomez-Valero, Rusniok, et al. 2019). The two MAGs branching first on the way to Legionella provide some information on how the Legionella genus evolved, among others by their small genome size. The first one (TARA_PSE_MAG_00004), branching very shortly after the Coxiellaceae/Legionellaceae divergence, was reconstructed from a coassembly of 16 metagenomes collected in Southeastern Pacific by the Tara Ocean expedition (Delmont et al. 2018). Despite its robust phylogenetic association within Legionellaceae, its T4BSS resembles Coxiellaceae, lacking icmX, icmM/dotJ, and icmG/dotF (supplementary fig. 2, Supplementary Material online). It is also apparently lacking all conserved effectors. The next MAG (RIFCSPHIGHO2_12_FULL_37_14), assembled from groundwater samples taken in the Rifle Aquifer (Colorado), is more similar to Legionella. It lacks part of the icm operon (icmP-icmB), probably due to incomplete binning (estimated completeness: 95%), but it harbors the Legionella-specific icmX and six of the eight conserved Legionella effectors. This confirms the view that icmX and most of the conserved effectors were gained in the
Legionellaceae, and not lost by the other groups. It also sup-
port that the number of protein families which, in turn, tends to be underestimated in very deep nodes. In these nodes, genes are more likely not to be preserved in any of the descendants, due to their scarce representation—in that case the rest of the Gammaproteobacteria.

This pattern suggests that the last common ancestor of Legionellaceae had a smaller genome than extant Legionella species, although it cannot be completely excluded that genome reduction occurred independently in these two MAGs.

Incidentally, both MAGs contained inside the Legionella clade III (Gomez-Valero, Rusniok, et al. 2019) also display smaller genomes than their closest relatives: 1.9 and 2.3 Mb, whereas clade III Legionella spp. range from 3.1 to 3.9 Mb. The same is also true, albeit to a lesser extent, with the MAG branching in clade IV (Legionella_sp_40_6: 3.1 Mb; rest of the clade: 3.4–4.9 Mb). The reduced genome size in these MAGs, associated with the longer branches leading to them, might indicate shifts in lifestyles, either toward stream-lining as in the Pseudomonas (Giovannoni 2017) or toward an increased dependency on their host (Toft and Andersson 2010), similarly to "Ca. Legionella polyplacis" (Rihova et al. 2017). Inferred genome sizes in MAGs may differ from the actual genome sizes, since segments of the genome can be missed by the binning algorithms, and conclusions reached from these should be taken with some caution. However, the MAGs branching within the Legionellaceae are predicted to be almost complete, with the least complete being 89% complete and three being >96% complete (supplementary table 2, Supplementary Material online), which supports the results discussed above.

Evolution of Genome Content in the Legionellales Last Common Ancestors
To further understand evolutionary paths taken by the ancestors of the order, we reconstructed gene flow within Legionellales. The phylogenetic birth-and-death model implemented in Count (Csuros 2010) was used on the same set of
genomes (Legio93) as the tree in figure 1. Count estimates, at each node on the tree, the probability that each gene family is present, present in multiple copies, gained, lost, expanded, and contracted. To analyze specific gene families, we used a 0.5 probability threshold, for each state: for example, we deemed specific families to be gained at a certain node if their probability to be gained was larger than 0.5. However, summary statistics per node (e.g., the total number of families gained on a certain node) are calculated as the sum of gain probabilities for all gene families at that node. These two ways of calculating gains lead to some slight discrepancies when comparing statistics at the node level or specific gene families.

The reconstructed gene flow reveals 553 gene family losses from the LFLA (node 107) to LLCA sl (node 95; fig. 2 and supplementary fig. 2 and table 3, Supplementary Material online), consistent with the decrease in genome size. These losses can be compared with 781 losses on the branch leading to Coxiella/Aquicella, another intracellular group. A subsequent large gain in gene families and in genome size in the last common ancestors of Legionellaceae and Berkea spp. (588 and 556 protein families, respectively) but not in the ancestors of Coxiellaceae (Coxiella/Aquicella) suggests that LLCA had a genome size comparable to that of the latter group (average: 1.78 Mb), compatible with genome sizes of other Gammaproteobacteria intracellular bacteria (Toft and Andersson 2010). The details of the gene gains in Legionellaceae and Berkea are presented in supplementary results, Supplementary Material online.

Functional analysis of protein family losses from LFLA to the LLCA ss reveals large losses across all COG categories (fig. 3). However, more than half (52.1%) of these protein families are involved in metabolism. The categories that account for the largest decreases include indeed metabolic functions (inorganic ion transport and metabolism [category P], energy production and conversion [C], carbohydrate transport and metabolism [G]) but, maybe more unexpectedly, also functions involved in transcription (K) and signal transduction mechanisms (T).

A closer analysis of the metabolic genes lost from LFLA to LLCA ss reveals a large number of genes directly or indirectly linked to Molybdenum metabolism. Molybdenum (Mo) is a second-row transition metal, the only one essential to most organisms. In most cases, enzymes possessing a Mo atom also contain a cofactor, molybdopterin (Mendel and Bittner 2006). Mo-containing enzymes are essential to catalyze key steps of carbon, nitrogen, and sulfur metabolism (Hille 1996). Among the functions that were lost during transition to intracellular lifestyle in Legionellales (LFLA to LLCA), several are linked to Mo metabolism: the molybdopterin synthesis proteins (eight families, including the moaABCDE operon and moeA), the molydate transporter operon modABC and the molybdopterin-guanine dinucleotide synthase mobA are missing from most Legionellales. Consistent with Mo being central in sulfur metabolism, several genes of the cysteine metabolism were lost at the same stage (six families, including three transporters), as already noticed for L. pneumophila and Aquicella spp, both auxotrophic for cysteine (George et al. 1980; Santos et al. 2003). Several other proteins of the sulfur relay system (including thiamine synthesis) are also missing: both subunits of a sulfate adenyltransferase, thiS (immediate sulfur donor in thiazole formation), the sulfurtitransferase tusE, and nudJ (bifunctional thiamine pyrophosphate hydrodrolase and thiamine pyrophosphate hydrodrolase). Another group of function predicted to be lost at the same time is nitrite/nitrate metabolism, of which many reactions are catalyzed by Mo-containing enzymes. These missing genes include the narGHJIKL operon, encoding among others subunits of the nitrate reductase, a nitrite reductase, and a nitric oxide reductase, as well as two enzymes encoding an allophanate hydrodrolase. The carbon metabolism-related functions lost at the same stage appear less coherent and seem to include enzymes in various pathways. One exception is the glycogen metabolism operon glgPACXB (Chandra et al. 2011). Although it is not possible at this stage to infer the exact order of these losses or their consequences, they form a coherent picture of the LLCA, which became dependent on its host for many of the aspects of sulfur, nitrogen, and carbon metabolism. Interestingly, the recently sequenced “Ca. Azoamicus ciliaticola,” an obligate endosymbiont of an anaerobic ciliate, harbors many genes involved in denitrification respiration (including narGHK, modABC, and the nitrite- and nitrate reductases) predicted to be lost on the immediate descendant of the FLCA. This suggests that “Ca. Azoamicus ciliaticola” is only distantly related to Legionellales, although a thorough phylogenomic remains to be performed.

Another group of genes lost on the way to intracellular lifestyle are related to protection against oxidizing entities. In Escherichia coli, the regulator SoxR senses oxidants and activates other genes that protect the bacterium against superoxide and nitric oxide, among others (Koo et al. 2003). After being activated, SoxR is reduced by the SoxR iron–sulfur cluster reduction factor component, encoded by the operon rsxACDGE (Koo et al. 2003). SoxR is missing from the Legionellales genomes and rsxACDGE is part of the genes lost in the transition to intracellular lifestyle. In L. pneumophila, the function of SoxR is replaced by an homolog of OxyR (LeBlanc et al. 2008), widely distributed in the Gammaproteobacteria.

Beyond metabolic functions, DNA repair functions are often lost on the path to endosymbiosis (Toft and Andersson 2010). Although the homologous recombination protein gene recA is present in the large majority of Legionellales, the three genes recBCD, which initiate recombinational repair of double-strand breaks in DNA and are known to be missing in L. pneumophila (Bryan and Swanson 2011), is shown to be lost right after the LLPCA.

In summary, the deep ancestors of Legionellales on the way to LLCA experienced a loss of large portions of central metabolic functions (sulfur, nitrogen, carbon) linked to molybdenum metabolism, and transport of many molecules. This is consistent with these bacteria being adapted to hosts, progressively relying on the latter to provide them with complex molecules.

The flagella, present in the LLCA, is lost in Coxiellaceae, consistent with the observation that neither Coxiella or Aquicella possess one (Santos et al. 2003; Garrity et al. 2010, 2011). This dramatic change is also reflected in the genome size, which is highly reduced in LLCA compared to LFLA. The most important aspect of this genome reduction is that it leads to nearly complete loss of central metabolic pathways that evolved before transition to intracellular lifestyle.
Host Adaptation in *Legionellales* · doi:10.1093/molbev/msac037

**FIG. 2.** Overview of ancestral reconstruction of *Legionellales* genomes, with the T4BSS system and the *Legionella*-conserved effectors detailed. The ancestral reconstruction is based on the Bayesian tree in figure 1. At the left of each node, barplots depict the number of gene families (blue, ranging from 0 to 2,295) inferred at that node, as well as the number of gene family gains (green, from 0 to 589) and losses (red, from 0 to 781) on the branches leading to nodes. On the right panel, squares represent genes in the T4BSS—separated according to the three operons present in R64 (Segal et al. 2005)—and the eight effector genes conserved in *Legionella* (Gomez-Valero, Rusniok, et al. 2019). Rows in-between terminal nodes correspond to the last common ancestor of the two nearest terminal nodes, for example, the row between *Aquicella* and *Berkiella* corresponds to the LLCA sl. The color of the squares represents the posterior probability that each protein was present in the ancestor of each group. A complete tree is shown in supplementary figure 2, Supplementary Material online, and the underlying numbers are available in supplementary tables 3 and 4, Supplementary Material online.

**FIG. 3.** Protein family flow by COG category in the early *Legionellales* ancestors. The upper panels show the absolute number of families present (blue) in the ancestor, lost (red) and gained (green) on the way to its descendant (as indicated on the panel title). The lower panels show the fraction of families lost (red) or gained (green) on the way to the next node, relative to the number of families of that category in the ancestor. COG categories are grouped in four super-categories: information storage and processing (purple), cellular processes and signaling (orange), metabolism (pink), poorly categorized (grey).
Most of the flagellar genes (31 genes, flgABCDEFGHIJKLMNOP, flhABF, flhACDEFGIMNPRQS, motAB, minD) are predicted to be absent from the last common ancestor of Coxiiellaceae. The loss of the flagella represents 14.7% of all gene families lost in that ancestor.

Genome Content of the Legionellaceae Last Common Ancestors
Whereas the early Legionellales ancestors show a net loss of gene families, genome sizes increase in the early Legionellaceae. A trend toward larger genomes is unusual in host-adapted bacteria (Toft and Andersson 2010), due, among others, to the progressive loss of exposure to foreign DNA. However, in the case of Legionella genus, it appears that genome sizes went from approximately 2.2 Mb in the LLCA ss to approximately 3 Mb in the last common ancestor of genus. We suggest that a supplementary conjugation system (trb) and an improved transformation potential (through the initiation complex of the pilus) might have contributed to this increase. The details are discussed in supplementary results, Supplementary Material online.

Type IV Secretion System
A crucial feature for the success of many host-adapted bacteria is the ability to transfer proteins into the host's cytoplasm, typically through secretion systems. Among the proteobacterial VirB4-family of ssDNA conjugation systems (Type IV Secretion Systems, T4SS), one called T4BSS (also called MPF1 or I-type T4SS) probably arose early in the group (Guggielmini et al. 2013). It is present in almost all extant Legionellales, with the exception of the extremely reduced endosymbionts of the group (supplementary fig. 2, Supplementary Material online), where it is either missing completely or pseudogenized. It is partly missing in several of the small, novel Coxiiellaceae MAGs, possibly indicating increased host dependence. However, it is difficult to assess with certainty whether genes absent in MAGs represent true losses or are the result of incomplete binning of metagenomic contigs. T4BSS in Legionellales appears largely collinear (fig. 4 and supplementary fig. 4, Supplementary Material online). Ancestral reconstruction of the order revealed that T4BSS, as it is found in extant Legionellales, originated after the LFLA but before the LLCA, with several proteins added over time (fig. 2). IcmW and IcmS (two small acidic cytoplasmic proteins, part of the coupling protein subcomplex; Sutherland et al. 2012; Kwak et al. 2017), IcmC/DotE, IcmD/DotP, and IcmV (three small integral inner membrane proteins, of uncharacterized functions) were gained in the LLCA ss. IcmQ, a cytoplasmic protein and IcmN/DotK, an outer membrane lipoprotein (Ghosal et al. 2019), were acquired in the LLCA ss, after the divergence with Berkiiella. The limited knowledge about the function of these proteins prevents us to infer exactly how their gain allowed the LLCA to infect eukaryotic cells. However, except for IcmN/DotK, all these proteins are located either in the inner membrane or in the cytoplasm, suggesting that the specific role of the Legionellales-gained proteins is related to the coupling protein complex rather than in the core complex. In Legionella however, the T4BSS harbors three specific proteins: IcmX, IcmM/Dotl, and IcmG/DotF. The first one is subject to high recombination rates (Gomez-Valero, Chiner-Oms, et al. 2019), located on the surface of the bacteria, (Khemiri et al. 2008) and thus likely to be exposed to the host. The second one, IcmM/Dotl, is located in the inner membrane, and is a distant paralog of IcmL/Dotl with which it forms an heteroduplex (Kuroda et al. 2015). The third one, IcmG/DotF, has a periplasmic part which is thought to be under positive selection (Gomez-Valero, Chiner-Oms, et al. 2019), and is presumably involved in substrate recognition.

A homologous T4BSS system is present in Fangia hongkongensis, but mostly absent in the free-living Gammaproteobacteria and in Francisella, a sister clade to Fangia. It is likely that the T4BSS in Fangia has been acquired by horizontal gene transfer, possibly from Piscirickettsia (supplementary figs. 5 and 6, Supplementary Material online). It is however not possible to exclude that a T4BSS was present in the LFLA of Legionellales, and subsequently lost on the branch leading to the free-living Gammaproteobacteria.

A homologous T4BSS system is present in Fangia hongkongensis, but mostly absent in the free-living Gammaproteobacteria and in Francisella, a sister clade to Fangia. It is likely that the T4BSS in Fangia has been acquired by horizontal gene transfer, possibly from Piscirickettsia (supplementary figs. 5 and 6, Supplementary Material online). It is however not possible to exclude that a T4BSS was present in the LFLA of Legionellales, and subsequently lost on the branch leading to the free-living Gammaproteobacteria.

Contrary to T4BSS itself, which is highly conserved throughout the order, effectors are much more versatile. Of the eight effectors conserved in the Legionella genus (Gomez-Valero, Rusniok, et al. 2019), only two are found outside of the Legionellales family (fig. 2): LegA3/AnkH/AnkW (lpg2300), found in Legionellales ss and Berkiiella, and RavC (lpg0107), found in most Legionellales ss, but not in Berkiiella. MavN (lpg2831), which is conserved in Legionella spp. and was found in Rickettsiella (Burstein et al. 2016), is harbored by a few Coxiiellaceae species only (fig. 2 and supplementary fig. 2, Supplementary Material online), and is unlikely to have been acquired in the LLCA. The effector LegA3 (lpg2300) contains ankyrin repeats, and has been suggested to play a role in processes involved in modulation of phagosome biogenesis (Habyarimana et al. 2008). The protein translocates to the nucleus, where it interacts with several host targets. It results in changes in host transcription, which promote intracellular bacterial growth (Dwingelo et al. 2019). The function of RavC (lpg0107) is currently unknown, but it is found—beyond Legionellales—in Chlamydiaceae (Burstein et al. 2016) and has homologs throughout bacteria. The protein belongs to the required for meiotic division (RMD1) family, characterized in Saccharomyces cerevisiae where it is essential for sporulation (Enyenihi and Saunders 2003). It is difficult to draw very specific conclusions based on that level of functional evidence for those two effectors, but their high level of conservation strongly suggests a central role in how Legionellales interact with their hosts.
The effector lpg0140 (lpp0155/lpl0140/CetLp1) displays an interesting phylogenetic distribution, being present in all Legionella species (except “Ca. L. polyplacis”), in one Coxiellaceae MAG and in Piscirickettsia. Its function is currently unknown, but it shares a homologous domain with the protein LciE, whose expression is induced by the two-component system LciRS, in response to copper stimulation (Linsky et al. 2020). Lpg0140 is specific to Legionellales and Piscirickettsia, as no homologs could be found outside of these two groups (PSI-BLAST, E-value inclusion threshold ≤ 10^-5; eight rounds until convergence). Further functional studies are required to reveal its role in the infection process of Legionella and Piscirickettsia.

**Dating the Rise of the Legionellales**

Absolute dating of bacterial trees is often inconclusive, because very few reliable biomarkers can be unambiguously attributed to a bacterial clade and used as calibration points (Brocks and Pearson 2005). However, one such biomarker is okenone, a pigment whose degradation product okenane has been found in the 1.64-Gy-old Barney Creek Formation (Brocks and Schaeffer 2008), and is exclusively produced by Chromatiaceae (part of the purple sulfur bacteria) and Legionellales. The latter group does not produce okenone but a different isoprenoid compound using the same genes.

**Discussion**

The common ability of Legionellales bacteria to invade eukaryotic cells combined with the presence and high conservation
of the crucial host-adaptation genes, namely a complete T4BSS and two effectors, strongly suggest that the last common ancestor of Legionellales (LLCA) was already infecting other cells. The hypothesis that LLCA used these mechanisms to infect or interact with other prokaryotes cannot be completely ruled out, but seems extremely unlikely for several reasons. For example, if LLCA was only infecting prokaryotes, then the ability to infect eukaryotes would have to have evolved independently several times within Legionellales, in all free-living descending clades, with no exception (all experimentally investigated Legionellales are able to live inside the cytoplasm of eukaryotic cells). Furthermore, each of these independent host-adaptation events would have to have involved the same host-adaptation genes (all experimentally investigated Legionellales use the same T4BSS to interact with their hosts). Another indication that Legionellales were associated with early eukaryotes is provided by the presence of SNARE-like proteins, key to the vesicle fusion process, in the genomes of the former (Neveu et al. 2020). Some of these Legionellales-acquired proteins branch deep in the tree, close to root, defined by the homologs belonging to Heimdallarchaeota (Neveu et al. 2020), pointing toward a horizontal gene transfer from early eukaryotes to Legionellales. This last element of evidence is however mostly circumstantial, since proteins from Legionellales, Heimdallarchaeota, and eukaryotes were not present on the same tree.

It is also likely that LLCA, like extant Legionellales, depended on phagocytosis (or its prototypic version), a mechanism exclusively found in eukaryotes. This implies that LLCA appeared after the division of Archaea and the first eukaryotic common ancestor (FECA), thus placing the rise of phagocytosis and FECA at 1.75 Ga at the latest (lower bound of the 95% HPD for the first host-adapted ancestor, LLPCA), but could be as early as 2.12 Ga (higher bound of the 95% HPD for LLFLA). Given a consensual estimate of the age of the LECA of 1.0–1.6 Gy (Eme et al. 2014), the time for eukaryogenesis, defined as the time elapsed between FECA and LECA (Eme et al. 2017), would be at least 250 My, but could be over 1.1 Gy.

The existence of a phagocytosing eukaryote at 1.89 Gy, likely pre-LECA, has implications for hypotheses of eukaryogenesis. Four of the most prominent of these (reviewed, e.g., in López-García et al. 2017) make different assumptions about the timing of the mitochondrial endosymbiosis. In the hydrogen hypothesis (Martin and Müller 1998), the mitochondria arrived early (mito-early), with the mitochondrial endosymbiosis event itself triggering eukaryogenesis. In the phagocytosing archaean model (PhAT) (Poole and Neumann 2011; Martijn and Ettema 2013), the syntrophy hypothesis (López-García and Moreira 2006), and the serial endosymbiosis model (Pittis and Gabaldón 2016), the mitochondrion arrive late (mito-late). Specifically, in the PhAT model, phagocytosis machinery is a prerequisite for the fusion of the mitochondrial progenitor with an Asgard archaeon. The very early emergence of phagocytosis proposed in this contribution supports a mito-late scenario. Indeed, the timing of the mitochondrial endosymbiosis has been recently estimated to 1.21–2.053 Gy (Betts et al. 2018). Although confidence intervals overlap, this estimation would place the phagocytosis of the first Legionellales (1.75–2.12 Gy, see above) before the mitochondrial endosymbiosis. This timing is also consistent with the mito-intermediate scenario described by Vosseberg et al. (2021). These authors propose that the Asgard-related host that would later acquire the mitochondrion already harbored a certain level of complexity (notably, it had a dynamic cytoskeleton and membrane trafficking), whereas the complex signaling and regulation network and the complex organelar endomembrane system were created after the mitochondrial endosymbiosis. They also suggest that the cytoskeleton and membrane remodeling became increasingly more complex pre-endosymbiosis, perhaps
leading to a primitive phagocytosis system that enabled the mitochondrial endosymbiosis. This primitive phagocytosis system might have been used to engulf the first **Legionellales**.

A recent study estimates LECA to be much older, around 2–2.4 Gy old (Strassert et al. 2021). However, such an old age for LECA heavily relies on one single calibration point, fossil remains found in the approximately 1.6 Gy old Vindhyan formation. These fossils have been interpreted as crown-group red algae, thereby very significantly pushing back the minimum age for LECA. The attribution of these fossils to the crown-group red algae is however not unambiguous, and several authors do not to include it as a calibration point (Betts et al. 2018). Further, the age of LECA in Strassert et al. (2021) is also likely dependent on the prior on the age of the root (3.2–1.6 Gy), the rationale for which is not discussed in the article. These parameters do not significantly affect the main message of the article, which is about the (later) timing of plastid acquisition, but presumably have a significant effect on the estimation of the age of LECA. The latter age should be taken with caution. Should these results however hold true, the conclusions presented here would be different: instead of being coincident with eukaryogenesis, the host-adaptation event that created **Legionellales** would have occurred between LECA and the divergence of Amoebozoa, which are known hosts for Legionella and other Legionellales.

In summary, we show here that the LLCA already possessed a T4BSS and two effectors, and existed 1.9 Ga. We propose that LLCA, upon being phagocytosed by eukaryotic cells, already had the ability to resist digestion, owing to its host-adaptation genes. Thus, phagocytosis is likely at least 1.9 Gy old, older than previously thought. This hypothesis is consistent with a scenario in which some early eukaryotes developed phagocytic properties and fed on prokaryotes. Some of these, among them LLCA, rapidly acquired the abilities to resist host digestion and exploit the novel, rich ecological niche that is the eukaryotic cytoplasm.

**Materials and Methods**

**Genome Sequencing of Aquicella Species**

Two species of **A. lusitana**, DSM 16500 and **A. siphonis**, DSM 17428 (Santos et al. 2003) were cultivated, sequenced, and annotated as part of this study. The details are available in supplementary methods, Supplementary Material online.

**Protein Marker Sets**

A set of 139 PFAM protein domains was used as a starting point both to perform quality control of MAGs and for phylogenomic reconstructions. The original set was used by Rinke et al. (2013) (supplementary table 13, Supplementary Material online) to estimate the completeness of their MAGs. We used the same set (referred to as Bact139) to estimate MAG completeness. A large subset of the 139 protein domains is very frequently located in the same protein in a vast majority of Proteobacteria. This was evidenced by investigating 1,083 genomes, using phyloSkeleton 1.1 (Guy 2017) to select one representative per proteobacterial genus and to identify proteins containing the Bact139 domain set. Of these domains which often colocalized in the same protein, only the most widespread one was retained. In total, only 109 domains were used to identify proteins suitable for phylogenomics analysis. This set is referred to as Bact109 and is available in phyloSkeleton 1.1 (Guy 2017).

**Metagenomics and Selection of MAGs**

The details of the metagenome assembly and binning and of the identification of MAGs belonging to **Legionellales** are available in supplementary methods, Supplementary Material online.

**Phylogenomics**

Two sets of genomes (Gamma105 and Legio93) were used for phylogenomics in this contribution, and two extra ones (Bacteria134 and Bacteria94, both built on the Gamma105 set) were used only for the time-constrained and are described below. All sets consist of part or all of the novel **Legionellales** MAGs and genomes, supplemented with varying subsets of outgroups. The initial selection of the representative organisms and identification of markers was done with phyloSkeleton 1.1 (Guy 2017), initially choosing one representative per class in the Betaproteobacteria, the Zetaproteobacteria, and the Acidithiobacillus; one per genus in the Gammaproteobacteria, except in the Piscirickettsiaceae and the Francisellaceae (one per species). That initial selection of genomes was then trimmed down to reduce the computational burden of tree inference.

For both main sets, protein markers from the Bact109 set were identified with phyloSkeleton 1.1 (Guy 2017). Each marker was aligned separately with mafft-linsi v7.273 (Katoh and Standley 2013). The resulting alignments were trimmed with BMGE v1.12 (Criscuolo and Gribaldo 2010), using the BLOSUM30 matrix and the stationary-based trimming algorithm. The alignments were then concatenated. From these alignments, maximum-likelihood trees were reconstructed with FastTreeMP v2.1.10 (Price et al. 2010) using the WAG substitution matrix. Upon visual inspection of the resulting trees, the genome selection was reduced to remove closely related outgroups, and the phylogenomic procedure repeated. Once the genome data set was final, a maximum-likelihood tree was inferred with IQ-TREE v1.6.5 (Nguyen et al. 2015), using the LG (Le and Gascuel 2008) substitution matrix, empirical codon frequencies, four gamma categories, the C60 mixture model (Quang le et al. 2008), and PMSF approximation (Wang et al. 2018); 1,000 ultrafast bootstraps were drawn (Hoang et al. 2018).

The first set of genomes, called Gamma105, was used to correctly place **Legionellales** in Gammaproteobacteria, in particular with respect to Chromatiaceae. It encompasses 110 genomes, of which 105 are Gammaproteobacteria, one is a Zetaproteobacterium, three are Betaproteobacteria, and one is an Acidithiobacillus. The selected Gammaproteobacteria include representatives from Legionellales (22), Chromatiaceae (19), Francisellaceae (17), and Piscirickettsiaceae (4) (supplementary table 7, Supplementary Material online).
The second set, called Legio93, is focused on Legionellales itself and comprises 113 genomes, of which 93 belong to Legionellales and the rest to other Gammaproteobacteria (16 genomes), Betaproteobacteria (two genomes), Acidithiobacillus (one genome), and Zetaproteobacteria (one genome; supplementary table 2, Supplementary Material online).

For both sets, single-gene maximum-likelihood trees were inferred for each marker. The BMGE-trimmed alignments were used to infer a tree using IQ-TREE 1.6.5, using the automatic model finder (Kalyaanamoorthy et al. 2017), limiting the matrices to be tested to the LG matrix and the C10 and C20 mixture models. Single-gene trees were visually inspected for the presence of very long branches resulting from distant paralogs being chosen by phyloSkeleton.

For both sets, a Bayesian phylogeny was inferred with phylobayes MPI 1.5a (Rodrigue and Lartillot 2014) from the concatenated BMGE-trimmed alignments, using a CAT+GTR model. For the Gamma105 set, four parallel chains were run for 9,500 generations. The chains did not fully converge, but three out of four chains yielded the same overall topology, whereas the last one had the Piscirickettsia and the Berkiella clades inverted (supplementary table 1, Supplementary Material online). For the Legio93 set, two chains were run for 16,000 generations. The chains converged in an acceptable way and mixed well (maxdiff <0.15, with 3,000 generations as burn-in; ESS >100 for all parameters). The majority-rule tree obtained from the Bayesian trees for Legio93 was subsequently used for the ancestral reconstruction (see below), whereas the one for Gamma105 was used to place Legionellales in the Gammaproteobacteria and as a ground to build another data set to estimate the time of divergence of the LLCA (see below).

To estimate the date of divergence of Legionellales, we created a third data set (Bacteria134), adding 27 additional bacterial genomes from a recent study by Betts (2018) to the Gamma105 data set (containing 110 genomes) (supplementary table 8, Supplementary Material online). Three MAGs from the Gamma105 data set were removed, due to low completeness, yielding a data set with 134 genomes. Protein markers were identified, aligned, and the alignment trimmed as described above. Trimmed alignments were concatenated and the resulting alignment (26,344 sites) was used as input to infer maximum-likelihood with IQ-TREE as described above, but using a C50 mixture model. The resulting tree was used as input to MCMCTree (see below). A fourth data set (Bacteria94) was obtained by 40 removing fast-evolving intracellular organisms (Legionellales and Francisellales) from the Bacteria134 set and aligning the resulting sequences. The tree based on Bacteria94 was pruned to produce a tree corresponding to the Bacteria94 data set.

Analysis of Genes and Genetic Systems
To relate the earliest documented trace of okenone to a specific ancestor, we investigated literature and searched for homologs of the genes specific to okenone synthesis. We also separately analyzed the genes of the T4BSS, effector proteins, and proteins involved in competence. The details are available in supplementary methods, Supplementary Material online.

Time-Constrained Tree
To estimate the time of divergence of Legionellales, we used MCMCTree from the paml package v4.9j (Yang 2007), taking advantage of the implemented approximate likelihood calculations (dos Reis and Yang 2011). The concatenated alignment was considered as a single partition and analyzed under a LG+G model. We assumed a uniform birth/death rate with an uncorrelated clock (clock = 2). Due to the considerable variation in substitution rates existing among bacteria (Kuo and Ochman 2009; Duchêne et al. 2016; Gibson and Eyre-Walker 2019), and as evidenced by the very different branch lengths obtained here (supplementary figs. 9 and 7, Supplementary Material online), a strict clock model was unlikely to be warranted, and therefore not used. A uniform prior calibration distribution was chosen in all cases, with hard bounds. We chose Gy as time unit. The Dirichlet-gamma prior for the mean substitution rate (rgene_gamma) was estimated using the median branch length from tips to root, \( \beta = 2.55 \), and the mean between the minimum (3.21 Gy) and maximum (4.52) age of the root (\( t = 3.865 \)). This provided an estimated substitution rate (b/t) equal to 0.660 substitutions per site per Gy or 6.60 \times 10^{-8} \text{ per site per year}. In the gamma distribution, the \( \alpha \) parameter sets the width of the distribution, and a value of 2 was selected to cover a broad range of substitution rates to account for the fast-evolving organisms in the data set. The \( \beta \) parameter of the distribution was then calculated by setting the middle of the distribution (\( \alpha/\beta \)) to the estimated substitution rate, yielding a \( \beta \) value of 3.03.

To calibrate the clock, it was assumed that the divergence of all okenone-producing Chromatiaceae was at least as old as the rocks where the earliest traces of okenone (a degradation product of okenone) were found, 1,640 ± 30 Ma old (Page and Sweet 1998; Brocks et al. 2005; Brocks and Schaeffer 2008). The use of a taxonomically broader data set (Bacteria134, see above) allowed to use another four, more distant, calibration points. Divergence time estimates for three of them were taken from the study by Betts et al. (2018) (“Total Cyanobacteria,” min. age 3,225 Ma, max. age 4,520 Ma; “Crown Cyanobacteria,” min. age 1,033 Ma, max. age 4,520 Ma; “Crown Alphaproteobacteria,” min. age 1,033 Ma, max. age 4,520 Ma) and one from a study by Lin et al. (2017) (common ancestor of Nitrosira and Proteobacteria, min. age 3,000 Ma, max. age 4,520 Ma).

MCMCTree was run with the concatenated Bacteria134 data set and the maximum-likelihood tree inferred with IQ-TREE as input. Two chains were run for 5 million generations, discarding the first 10% as burn-in. Tracer 1.7.2 (Rambaut et al. 2018) was used to check that both chains had converged (all parameters with ESS >200) and were mixing well. To estimate how the priors affected the runs, parameters obtained from the actual run were compared with those from a run with 1 million generations drawing only from the priors. All parameters displayed very different
distributions, suggesting that the choice of the priors had very little effect on the obtained values.

An analysis of the variation of branch rates was performed to investigate whether a strict clock was indeed not warranted in this case. The coefficient of rate variation (defined as the standard deviation of the branch rates divided by the mean rate) was equal to 0.27, suggesting that the strict clock model could be rejected (Ho et al. 2015).

To estimate whether the increased evolutionary rate in in Legionellas and Francisellaceae affects the dating of the Legionellas divergence, we created a reduced data set (Bacteria94, see above) by removing the intracellular genera Legionella, Francisella, Fangia, and Piscirickettsia from the Bacteria134 data set. MCMCTree was run similarly as above, using the maximum-likelihood tree mentioned above, pruned of the relevant clades.

An attempt to estimate the time of divergence of LLCA with BEAST 2 (Bouckaert et al. 2014), using a relaxed clock model (uncorrelated log-normal) (Drummond et al. 2006), was unsuccessful. The chains did not converge, after over 55 million generations.

Protein Family Clustering and Annotation

All 113 proteomes from the Legio93 set were clustered into protein families using OrthoMCL 2.0.9 (Li et al. 2003). All proteins were aligned to each other with the BlastP variant of DIAMOND v0.9.8.109 (Buchfink et al. 2015), with the –more-sensitive option, enabling masking of low-complexity regions (–masking 1), retrieving at most 105 target sequences (–k 100,000), with an E-value threshold of 10^-5 (-e 1e-5), and a precalculated database size (–dbsize 98,280,075). In the clustering step, mcl 14-137 (Enright et al. 2002) was called with the inflation parameter equal to 1.5 (+1.5), as recommended by OrthoMCL.

Annotation of the protein families obtained by OrthoMCL was performed by searching for protein accession number in a list of reference genomes (see below). If any protein in a family was present in the first genome, the family was attributed this annotation; else, the second genome was searched for any of the accession numbers, and so on. The reference list was as follows (GenBank accession numbers between parentheses): Legionella pneumophila subsp. pneumophila strain Philadelphia 1 (AE017354), Legionella longbeachae NSW150 (FN650140), Legionella oakridgensis ATCC 33761 (CP004006), Legionella fallonii LLAP-10 (LN614827), Legionella hackeliae (LN681225), Tatlockia micdadei ATCC33218 (LN614830), Coxiala burnetii RSA 493 (AE016828), Coxiala endosymbiont of Amblyomma americanum (CP007541), Escherichia coli str. K-12 substr. MG1655 (U00096), Francisella tularensis subsp. tularensis (AJ749949), Piscirickettsia salmonis LF-89 = ATCC VR-1361 (CP011849), Acidithiobacillus ferrooxidans SS3 (CP011849), Neisseria meningitidis MC58 (AE002098), Rickettsiella gyalli (NZ_AAQJ0000000.2). Legionella effector orthologous groups were identified by comparing protein accession numbers with the list provided by Burstein et al. (2016).

Ancestral Reconstruction

The flow of genes in the order Legionellas was analyzed by reconstructing the genomes of the ancestors. The ancestral reconstruction was performed with Count v10.04 (Csuros 2010), which implements a maximum-likelihood phylogenetic birth- and death model. The Bayesian input tree was obtained from the Legio93 data set (see above) and the OrthoMCL families as described above. The rates (gain, loss, and duplication) were first optimized on the OrthoMCL protein families, to allow different rates on all branches. The family size distribution at the root was set to be Poisson. All rates and the edge length were drawn from a single Gamma category, allowing 100 rounds of optimization, with a 0.1 likelihood threshold. Reconstruction of gene flow was then performed by using posterior probabilities.

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments

This work was supported by the Swedish Research Council (2017-03709 to L.G.), Science for Life Laboratory (SciLifeLab National Project 2015 to L.G.), and the Carl Tryggers Foundation (CTS 15:184, CTS 17:178 to L.G.). We would like to thank Thijs Ettema and Lisa Klasson for constructive discussion; Jennah Dharmashi and Lina Juzokaite for help with the 16S amplicon protocol; Laura Eme for help with dating issues; Joran Martijn and Julian Vosseberg for the help with the Tara Ocean binning; Iker Iriarri for his advice about molecular dating; and Jennifer Ast for her help with editing the manuscript. We would also like to thank the Ag1000G project for releasing early their data.

Author Contributions

L.G. conceived the study and drafted the manuscript. E.H. performed database screening and metagenomics. A.G. performed the functional analysis. T.A. analyzed the T4BSS and effectors. A.G., E.H., and L.G. performed phylogenomics. All authors contributed to writing the final manuscript and approved it.

Data Availability

The raw and assembled data for the Aequorea genomes are deposited at the European Nucleotide Archive (ENA) under study accession number PRJEB29684. The assemblies for A. lusitana and A. siphonis have the accessions GCA_902459475 (replicons LR699114.1–LR699118.1) and GCA_902459485 (replicons LR699119.1–LR699120.1). MAGs, genomes, associated proteomes, and alignments underlying the trees presented in this contribution are available at Zenodo, with doi:10.5281/zenodo.4607174.

References

Betsz HC, Puttick MN, Clark JW, Williams TA, Donoghue PCL, Pisani D. 2018. Integrated genomic and fossil evidence illuminates life’s early evolution and eukaryote origin. Nat Ecol Evol. 2(10):1556–1562.
LeBlanc JJ, Brassinga AKC, Ewann F, Davidson RJ, Hoffman PS. 2008. An ortholog of OsxR in Legionella pneumophila is expressed postexponentially and negatively regulates the alkyl hydroperoxide reductase (ahpC2D) operon. J Bacteriol. 190(10):3444–3455.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178–2189.

Lin W, Paterson GA, Zhu Q, Wang Y, Kopylova E, Li Y, Knight R, Bazyliński DA, Zhu R, Kirschvink JL, et al. 2017. Origin of microbial biomineralization and magnetotaxis during the Archean. Proc Natl Acad Sci U S A. 114(9):2171–2176.

Linsky M, Vitkin Y, Segal G. 2020. A novel Legionella genomic island encodes a copper-responsive regulatory system and a single Icm/Dot effector protein transcriptionally activated by copper. mBio 11(1):11.

López-García P, Eme L, Moreira D. 2017. Symbiosis in eukaryotic evolution. J Theor Biol. 434:20–33.

López-García P, Moreira D. 2006. Selective forces for the origin of the eukaryotic nucleus. Bioessays 28(5):525–533.

Martijn J, Ettema TJ. 2013. From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. Biochem Soc Trans. 41(1):451–457.

Martin W, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392(6671):37–41.

Martin WF, Tielen AGM, Mentel M, Garg SG, Gould SB. 2017. The hydrogen hypothesis. Nat Commun 8:15999.

Mendel RR, Bittner F. 2006. Cell biology of molybdenum. Biochim Biophys Acta. 1763(7):621–635.

Neveu E, Khalifeh D, Salamin N, Fasshauer D. 2020. Prototypic SNARE proteins are encoded in the genomes of Heimdallarchaeota, potentially bridging the gap between the prokaryotes and eukaryotes. Curr Biol. 30(13):2468–2480.e5.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1):268–274.

Page RW, Sweet JP. 1998. Geochronology of basin phases in the western Mt Isa Inlier, and correlation with the McArthur Basin. Aust J Earth Sci. 45(2):219–232.

Pittis AA, Gabaldón T. 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature 531(7592):101–104.

Poole AM, Gribaldo S. 2014. Eukaryotic origins: how and when was the mitochondrion acquired? Curr Biol. 24(8):2305–2314.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson SI, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499(7459):431–437.

Rodríguez N, Lartillot N. 2014. Site-heterogeneous mutation-selection models within the PhyloBayes-MP package. Bioinformatics 30(7):1020–1021.

Santos P, Pinhal I, Rainey FA, Empadinhas N, Costa J, Fields B, Benson R, Verissimo A, Da Costa MS. 2003. Gamma-proteobacteria Aquicella lusitana gen. nov., sp. nov., and Aquicella sphinis sp. nov. infect protozoa and require activated charcoal for growth in laboratory media. Appl Environ Microbiol. 69(11):6533–6540.

Segal G, Feldman M, Zusman T. 2005. The Icm/Dot type-IV secretion systems of Legionella pneumophila and Cauliella burnetti. FEMS Microbiol Rev. 29(1):65–81.

Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521(7551):173–179.

Spang A, Stairs CW, Dombrowski N, Erle M, Lombard J, Caceres EF, Greening C, Baker BJ, Ettema TJG. 2019. Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. Nat Microbiol. 4(7):1138–1148.

Strassert JHF, Irisarri I, Williams TA, Burki F. 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nat Commun. 12(1):1879.

Sutherland MC, Nguyen TL, Tseng V, Vogel JP. 2012. The Legionella lcsMsw complex directly interacts with DotL to mediate translocation of adaptor-dependent substrates. PLoS Pathog. 8(9):e1002910.

Toft C, Andersson SCE. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. Nat Rev Genet. 11(7):465–475.

van Schalkwijk CJ, Chen C, Mertens K, Weber MM, Samuel JE. 2013. Molecular pathogenesis of the obligate intracellular bacterium Cauliella burnetti. Nat Rev Microbiol. 11(8):561–573.

Vogel K, Bryant DA. 2012. Biosynthesis of the biomarker okenone: chi-ring formation. Geobiology 10(3):205–215.

Wesenberg J, van Hoof JJE, Marcet-Houben M, van Vlimmeren A, van Wijk LM, Gabaldón T, Snel B. 2021. Timing the origin of eukaryotic cellular complexity with ancient duplications. Nat Ecol Evol. 5(1):92–100.

Wang HC, Minh BQ, Susko E, Roger AJ. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst Biol. 67(2):216–235.

Williams KP, Gillepsie JJ, Sobral BW, Nordberg EK, Snyder EE, Shallom JM, Dickerman AW. 2010. Phylogeny of gammaproteobacteria. J Bacteriol. 192(9):2305–2314.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Backstrom D, Juzokaite L, van Sijl R, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541(7637):353–358.