Pinpointing the origin of mitochondria

Zhang Wang
Hanchuan, Hubei, China

B.S., Wuhan University, 2009

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Biology

University of Virginia
August, 2014

Abstract

The explosive growth of genomic data presents both opportunities and challenges for the study of evolutionary biology, ecology and diversity. Genome-scale phylogenetic analysis (known as phylogenomics) has demonstrated its power in resolving the evolutionary tree of life and deciphering various fascinating questions regarding the origin and evolution of earth's contemporary organisms. One of the most fundamental events in the earth's history of life regards the origin of mitochondria. Overwhelming evidence supports the endosymbiotic theory that mitochondria originated once from a free-living α-proteobacterium that was engulfed by its host probably 2 billion years ago. However, its exact position in the tree of life remains highly debated. In particular, systematic errors including sparse taxonomic sampling, high evolutionary rate and sequence composition bias have long plagued the mitochondrial phylogenetics. This dissertation employs an integrated phylogenomic approach toward pinpointing the origin of mitochondria. By strategically sequencing 18 phylogenetically novel α-proteobacterial genomes, using a set of "well-behaved" phylogenetic markers with lower evolutionary rates and less composition bias, and applying more realistic phylogenetic models that better account for the systematic errors, the presented phylogenomic study for the first time placed the mitochondria unequivocally within the *Rickettsiales* order of α-proteobacteria, as a sister clade to the *Rickettsiaceae* and *Anaplasmataceae* families, all subtended by the *Holosporaceae* family. Using this refined mitochondrial phylogeny as framework, gene content reconstruction provides strong evidence that the last common ancestor of mitochondria and α-proteobacteria is an obligate endosymbiont possessing an

ATP/ADP translocase that imports ATP from the host, which directly contrasts with the current role of mitochondria as the cell's energy producer. In addition, it was predicted to possess a flagellum and be capable of oxidative phosphorylation under low oxygen condition. Our ancestral state reconstruction shines light on the driving force of the initial endosymbiosis event. We find features consistent with the "oxygen scavenger hypothesis" but no support for the alternative "hydrogen hypothesis". Furthermore, characterization of individual bacterial genomes provides valuable insights into bacterial predation and endosymbiosis in general.

Acknowledgement

Working as a Ph.D. student has been one of the most magnificent as well as challenging experience to me. There are a great number of people without whom this research might not have been accomplished, and to whom I am greatly indebted.

My first gratitude must go to my advisor, Dr. Martin Wu. It has been an honor to be his first Ph.D. student. Martin not only provided me with the opportunity to pursue my Ph.D. study in the United States, but also introduced me to the field of genomics and bioinformatics that are to become my career choice. Throughout my dissertation research, Martin has been a constant source of support both intellectually and spiritually. His guidance has been throughout every possible area of my Ph.D. research, from troubleshooting a single experiment, optimizing a Perl script, to formulating scientific ideas and the writing of this thesis. His advice has influenced me profoundly both in academic and personal level. For these, I feel so much blessed and will always be grateful.

I would also like to thank all my committee members, Dr. Michael Timko, Dr. Lei Li, Dr. Cameron Mura, Dr. Douglas Taylor and Dr. Robert Cox. Their valuable inputs at each committee meeting always inspire me to think more deeply and broadly on my thesis project and the field in general. Meanwhile, I am also grateful to all the past and present lab mates. They have been making the lab an enjoyable place to work in. When I first joined lab, I knew almost nothing about programming and bioinformatics. Alex Koeppel

spent a lot of his time teaching me all the Unix basics and computational skills, without which this research could not even be started. I am also thankful to Sasha Scott, who has performed substantial preliminary work on the phylum-level bacterial phylogenetic marker database, and Tiantian Ren, who has provided many helpful suggestions to both wet bench and bioinformatic aspects of this project.

This dissertation project is impossible without the considerable help from other members of UVa. I would like to especially thank John Chuckalovcak and AnhThu Nguyen from the Genomics Core Facility, who have put in many efforts generating sequencing data for this project, and Katherine Holcomb from UVACSE, who has helped me solve all kinds of technical difficulties in cluster computing. In addition, Xiaozeng Yang and Huiyong Zhang have also taught me a lot of bioinformatics and molecular biology skills that greatly facilitated the accomplishment of this project.

Collaboration is an indispensable part of this project. Dr. Daniel Kadouri from UMDNJ, Dr. John Dustin Loy from University of Nebraska, and Drs. Wei Lin and Yongxin Pan from Institute of Geology and Geophysics, Chinese Academy of Sciences have all provided valuable bacterial samples and genomic DNA. Dr. Daniel Kadouri has also contributed to the manuscript preparation of the *Micavibrio* genome project. I would also like to thank Dr. Toni Gabaldon from CRG Barcelona who generously provided alignment data and Dr. Daniel Barker from University of St. Andrews who provided the BayesTrait program along with many helpful suggestions on the Bayesian inferences.

Finally, I would like to express my deepest gratitude to my parents in China. My father has been my life tutor. Although he knows nothing about my research, he is always able to find a way to push me forward whenever I face difficulties. On the other hand, I have received the best care and consideration from my mother in almost every possible way. Despite being thousands of miles away, I could always feel that they are standing closely behind me. Their love has provided me with the greatest inspiration and has been driving me through the entire process. This work is dedicated to them.

Table of Contents

Introduction

Ever since Charles Darwin published his theory of evolution in *The Origin of Species* in 1859 (Darwin 1859), phylogenetics — the reconstruction of evolutionary relationships among groups of organisms — has been a prerequisite of almost any studies in the fields of ecological and evolutionary biology. A phylogenetic tree of life essentially delineates a hierarchical classification of the extant organisms into distinct subgroups of a common ancestor. Thus it provides a comparative and predictive framework allowing us to infer the trait evolution and reconstruct the ancestral states from features of contemporary species. Assembling an accurate phylogenetic tree has proven to be extremely insightful in deciphering the diversity of life as well as interpreting the origin and subsequent evolution of contemporary organisms. Over the past few decades, our understanding of the tree of life has advanced rapidly fueled by enormous progress in the field of genomics and powerful sequencing technology. In particular, evolutionary histories of numerous branches of the tree of life that were traditionally considered to be irresolvable have now become fully elucidated by genome-scale molecular phylogenetics (known as phylogenomics) (Delsuc, et al. 2005). With the explosive growth of genome sequence data, we have entered an era in which it is increasingly possible to answer various fascinating evolutionary questions regarding the history of life.

One of the most fundamental events in the history of life is the origin of mitochondria. The origin of mitochondria has been studied since over a century ago. It was firstly articulated in the 1920s by Ivan Wallin as part of the endosymbiotic theory, which proposed that both mitochondria and plastid, two key organelles of eukaryotes, evolved from free-living bacteria via symbiosis with their primitive host cells. Being largely overlooked over the next decades, the endosymbiotic theory was resurrected and popularized by Lynn Margulis in 1970 in the paper

On the origin of mitosing cells (Sagan 1967). The remarkable similarity between organelles and bacteria in terms of morphology, physiology, biochemistry and genome organization lent a strong support for the bacterial origin of these organalles. Further characterization of DNA sequences of both organelles unambiguously confirmed the endosymbiotic theory in that both mitochondrial and plastid genomes bear a striking resemblance to their bacterial counterparts. With increasing organelle and bacterial sequences available, molecular phylogenetics of rRNA gene as well as a few proteins placed mitochondria and plastids at different branches of the bacterial tree of life. Mitochondria have been placed within the α-class of Proteobacteria, while plastids were thought to have evolved from Cyanobacteria. However, exactly when the endosymbiosis happened is still the subject of intense debate.

Mitochondria are usually viewed as oxygen-consuming, ATP-producing organelles that metabolize pyruvate through tricarboxylic acid (TCA) cycle and oxidative phophorylation. Yet mitochondria are known to exist in various forms across eukaryotic lineages. For example, the mitochondria of several unicellular protists and parasitic nematodes rely on terminal electron acceptors other than oxygen, such as NO_3^- and NO_2^-, and are therefore referred as anaerobic mitochondria (Finlay, et al. 1983; Kobayashi, et al. 1996; Zumft 1997; Takaya, et al. 1999). Another type of anaerobic ATP-producing organelle, the hydrogenosome, has been described in a wide variety of anaerobic protists, such as ciliates, amoeboflagellates, chytridiomycete fungi and parabasalids (Muller 1993; van der Giezen, et al. 1997; Akhmanova, et al. 1998; Hackstein, et al. 1999; Voncken, et al. 2002). The hydrogenosome lacks pyruvate dehydrogenase (PDH) and membrane-associated electron-transport chain typically present in aerobic mitochondria. Instead it possesses a pyruvate:ferrodoxin oxidoreductase (PFO) and a hydrogenase. In hydrogenosome, ATP is generated from pyruvate via substrate-level phosphorylation with the production of molecular hydrogen. A fourth type of mitochondria-like organelle, known as mitosome, is an

even more degenerated organelle that is not involved in ATP synthesis at all. Mitosome was first described in *Entamoeba histolytica* (Clark and Roger 1995; Mai, et al. 1999; Tovar, et al. 1999) and has been subsequently identified in several species of *Microsporidia* such as *Glardia lamblia* (Tovar, et al. 2003) and *Encephalitozoon cuniculi* (Goldberg, et al. 2008; Tsaousis, et al. 2008). Unlike most other forms of mitochondria, mitosome does not have its own genomic DNA. In addition, it is devoid of any proteins involved in major mitochondrial metabolism with the exception of a number of proteins involved in Fe-S cluster assembly (Tovar, et al. 2003). Despite such a great metabolic diversity among different forms of mitochondria, genetic material of these organelles and sequences of several nuclear-encoded genes of mitochondrial origin clearly support a common origin of all forms of mitochondria from an ancestral bacterium within α-proteobacteria (Bui, et al. 1996; Germot, et al. 1996; Roger, et al. 1996; Rosenthal, et al. 1997; Peyretaillade, et al. 1998; Roger, et al. 1998).

One key difference between the origin of plastid and mitochondria is that while plastids are restricted to certain eukaryotic lineages (such as plants and algae), mitochondria are ubiquitously present in all the extant eukaryotes. The absence of recognizable mitochondria in certain unicellular eukaryotes (i.e *Microsporidia*) have led to the classical serial endosymbiotic theory that the host of the mitochondrial endosymbiont was a primitive nucleus-containing eukaryote, termed “archezoan”, which should form the basal branch in the eukaryotic tree of life (Cavaliersmith 1987, 1989). Subsequently, this hypothesis has been explicitly refuted by the evidence that, 1) those eukaryotes that lack fully-fledged mitochondria nonetheless possess a mitochondrial remnant (in this case mitosome), as well as multiple nuclear genes clearly of mitochondrial origin (Peyretaillade, et al. 1998; Roger, et al. 1998; Mai, et al. 1999), suggesting mitochondria were present in these lineages at some point and were only secondarily lost, 2) the basal placement of those eukaryotes is likely a result of tree artifact of the SSU rRNA gene

phylogeny, and the "archezoa" appear to be a group of highly derived lineages in phylogenies of protein-coding genes (Hirt, et al. 1999; Keeling, et al. 2000). Thus there is currently no known extant eukaryotic lineage that is convincingly amitochondriate, which implicates that the endosymbiosis event leading to the origin of mitochondria likely occurred prior to the divergence of all eukaryotes.

Regarding the driving force of the endosymbiosis, the traditional serial endosymbiotic theory proposes that the symbiosis was driven by the production of ATP by the aerobic symbiont in exchange for the organic compounds provided by the anaerobic host. However, this view has also been challenged on the biochemical and physiological basis that none of the free-living bacteria known so far encodes ATP exporters, and nor would it necessarily provide excess ATP to initiate such symbiont-host association (Andersson, et al. 1998; Martin and Muller 1998; Vellai, et al. 1998; Keeling, et al. 2000).

Alternative hypotheses have been proposed to account for the circumstances of the founding endosymbiotic events (Embley and Martin 2006; Koonin 2010). For example, the "hydrogen hypothesis", proposed by Martin et al, hypothesizes that the metabolic syntrophy between a H_2-producing anaerobic α-proteobacterium and an autrotrophic H_2-dependent archaeon as the driving force behind the endosymbiosis (Martin and Muller 1998). The hydrogen hypothesis is particularly appealing in that it allows the possibility of a simultaneous origination of mitochondria and the nucleus, with the same α-proteobacterium also contributing to the rise of eukaryotic nucleus by fusing its genome with the host genome. Therefore, the "hydrogen hypothesis" directly challenges the traditional archezoan hypothesis in which the host is posited to be a full-fledged, nucleus-containing eukaryote. In addition, the hydrogen hypothesis explicitly accounts for the origin of hydrogenosome, which has a metabolic feature highly

resembling the H_2-producing bacterial symbiont proposed in the hypothesis. On the other hand, the "oxygen scavenger hypothesis" proposes that the mitochondrial ancestor was an aerobic symbiont that consumed the oxygen that was toxic for its anaerobic host. In return, the heterotrophic host made the pyruvate accessible for energy production of the endosymbiont (Andersson, et al. 2003). In this case, the removal of oxygen by the mitochondrial ancestor from its anaerobic host has driven the initial symbiosis. Therefore, the "oxygen scavenger hypothesis" differs fundamentally from the "hydrogen hypothesis" in that it proposes an aerobic mutualism instead of an anaerobic syntrophy as the driving force. The "oxygen scavenger hypothesis" also has its unique merit in that mitochondria originated concurrently with the dramatic rising of global oxygen levels in earth's atmosphere, roughly 2 billion years ago, about the same time when mitochondria were originated (Kurland and Andersson 2000).

α-proteobacteria represent one of the most diversified bacteria subdivisions, exhibiting substantial variation in lifestyle, metabolic capacity and genome feature. Phylogenetics based upon the small subunit (SSU) rRNA gene have classified α-proteobacteria into six main subgroups, *Rhizobiales, Rhodobacterales, Caulobacterales, Rhodospirillales, Sphingomonadales* and *Rickettsiales*, each with its unique features and versatility (Ettema and Andersson 2009). For example, members of *Rhizobiales* contain a group of nitrogen-fixing bacteria highly abundant in the soil where they maintain a nitrogen-driven symbiotic relationship with the plant root nodules (Galibert, et al. 2001). *Rhodospirillales* include a group of purple non-sulfur bacteria capable of producing energy through phototrophy. Endosymbiosis and pathogenesis have occurred multiple times within α-proteobacteria, once in the lineage leading to the genera *Bartonella* and *Brucella* in *Rhizobiales*, and once in the lineage of the order *Rickettsiales* (Boussau, et al. 2004). *Rickettsiales* contain a group of obligate intracellular bacteria with a large number of notorious pathogens such as *Rickettsia* spp. causing a variety of human diseases (Raoult and Roux 1997).

One subgroup of *Rickettsiales* named *Wolbachia* contain members living inside insects where they are able to manipulate the reproductive system of their hosts (Stouthamer, et al. 1999). In addition, α-proteobacteria contain a group of bacteria that are most abundant on the planet, known as the SAR11 clade (Giovannoni, et al. 1990). Up to 50% of the cells in the upper ocean layers come from this specific subgroup (Giovannoni, et al. 1990). More interestingly, one member of the SAR11 clade, *Candidatus Pelagibacter ubique*, is one of the smallest free-living bacteria ever found, with its 1,308,759 bp genome also being the smallest free-living bacteria genome ever sequenced (Giovannoni, et al. 1990). Members of *Rhodospirillales*, *Rickettsiales* and SAR11 clade have all been suggested to be the close relatives of mitochondria (Andersson, et al. 1998; Esser, et al. 2004; Wu, et al. 2004; Fitzpatrick, et al. 2006; Williams, et al. 2007; Georgiades, et al. 2011; Thrash, et al. 2011; Rodriguez-Ezpeleta and Embley 2012).

A well-supported mitochondrial phylogeny is critical to understanding the endosymbiosis and evolution of mitochondria. A robust species tree essentially serves as a phylogenetic framework on to which traits of individual species can be mapped and their ancestral states can be inferred in an evolutionary context. Therefore, a reliable mitochondrial phylogeny is a prerequisite to reconstructing the gene complement of mitochondrial ancestor, through which a better test of alternative hypotheses is possible. For example, a key piece of support for the hydrogen hypothesis necessitates that the mitochondrial ancestor possessed a hydrogen-producing machinery. Earlier phylogenomic analyses supported two alternative hypotheses regarding the position of mitochondria: 1) grouping with the *Rhodospirillales* order, 2) grouping with the *Rickettsiales* order. Within *Rhodospirillales*, *Rhodospirillum rubrum* is a free-living bacterium capable of producing H_2 by fermentation, and has an overall physiology that is virtually identical to that found among eukaryotes that lack mitochondria and that possess anaerobic mitochondria (Tielens, et al. 2002). In comparison, members of *Rickettsiales* do not produce H_2 and their

genomes all lack hydrogenase genes. Placing mitochondria with *Rhodospirillum rubrum* and related genus certainly will lend stronger support to the "hydrogen hypothesis".

Refining the phylogenetic position of mitochondria within a particular α-proteobacterial group will also give us insight into the relative timing of endosymbiosis and reductive evolution of mitochondrial genome. For instance, SAR11 group has traditionally been place within the *Rickettsiales* order. Members of the SAR11 clade have a free-living lifestyle distinct from the rest of *Rickettsiales* that are all obligate intracellular bacteria. Therefore, a sister-clade relationship with the intracellular *Rickettsiales* would suggest that the endosymbiosis likely happened once, and the mitochondrial ancestor was an endosymbiont that had already reduced its genome to some extent. On the other hand, a sister-clade relationship with the SAR11 clade would implicate that the mitochondrial ancestor was a free-living bacterium and the endosymbiosis happened independently of the intracellular *Rickettsiales* members. Distinguishing between these two scenarios will be insightful in understanding the reductive genome evolution underlying the transition from a fully-fledged bacterium to a highly degenerated organelle and will have a direct impact on the prediction of genetic complement of the mitochondrial ancestor.

Interestingly, by sequencing the genome of *Candidatus Midichloria mitochondrii*, a novel and phylogenetically divergent member of *Rickettsiales*, Sassera et al. identified a number of 26 flagella biosynthesis genes and several cbb3-type cytochrome oxidases involved in oxidative phosphorylation under micro-oxic condition, which were otherwise absent in other *Rickettsiales* at the time of their study (Sassera, et al. 2011). Based on these findings they further predicted that these genes were likely present in the free-living mitochondrial ancestor, thereby suggesting that the mitochondrial ancestor can be motile and capable of oxidative phosphorylation under

low level of oxygen in the earth's atmosphere at the time of endosymbiosis. Traditional views of endosymbiosis all assume that the mitochondrial ancestor was engulfed by a predatory host capable of phagocytosis (Cavalier-Smith 2009). In this context, the presence of flagella in the mitochondrial ancestor would provide an alternative mechanism of host cell entry. With flagella, the mitochondrial ancestor could play a more active role in the endosymbiosis, potentially as a motile predatory or parasitic bacterium like today's bacterial predators *Bdellovibrio* and like organisms (BALOs) that searched, attached to and penetrated the host cells. On the other hand, the presence of cbb3 oxidases in the mitochondrial ancestor has important implications for understanding the ecological context in which the endosymbiosis occurred (Sassera, et al. 2011). Searching and sequencing of additional *Rickettsiales* genomes closely related to mitochondria would have a great potential to improve the accuracy of mitochondrial ancestral reconstruction, which may shed additional light on the nature of the mitochondrial ancestor and the founding endosymbiosis event.

In the presented dissertation, we took advantage of an integrated phylogenomic approach to pinpoint the origin of mitochondria in the tree of life. Recent advance of genomics makes it possible to combine sequences of hundreds of genes to reconstruct fairly robust evolutionary history. As an introduction of phylogenomics in general, Chapter 1 of this thesis presents a phylum-level bacterial phylogenetic marker database and highly resolved bacteria genome trees reconstructed using these markers. Chapter 2 discusses extensively the using of an integrated phylogenomic approach toward pinpointing the origin of mitochondria by 1) filling the gaps in the tree of life through sequencing genomes of 18 α-proteobacteria that represent a broad range of phylogenetic diversity, 2) identifying a number of "well-behaved" phylogenetic markers with lower evolutionary rates and less compositional bias, and 3) applying more sophisticated phylogenetic models that better account for LBA and sequence compositional bias. Using the

refined mitochondrial phylogeny as well as novel genomes closely related to mitochondria, we reconstructed the ancestral gene content of mitochondria and evaluated the alternative hypotheses pertaining to the driving force of the endosymbiosis event. Chapter 3 of this thesis presents a phylogenomic reconstruction of the mitochondrial ancestors.

α-proteobacteria have been dubbed as the Darwin's finches of the microbial world. Phylogenetically novel lineages sequenced in this study are also useful for understanding the tremendous ecological diversity displayed in this group. In Chapter 4, we illustrate this using *Micavibrio aeruginosavorus* ARL-13 as an example. *M. aeruginosavorus* is one of few known bacteria species that prey on other bacterial species. We demonstrate how determining its genome and transcriptome has helped us understand the molecular basis of the predation and the evolution of bacteria predation in general. Appendix 1 describes the results of comparative genomic analyses of four amoeba endosymbionts belonging to the *Rickettsiales* order, which provides new insights into the adaptation of these bacteria to their spectacular intracellular niches, and further supports the role of amoeba as a "melting pot" facilitating the lateral gene transfers among multiple distantly related bacteria residing within the same host.

Collectively, the results of this research show that using an integrated phylogenomic approach, we are able to refine the position of mitochondria and move one step closer toward pinpointing its origin. With this refined phylogeny, reconstruction of the mitochondrial ancestor genome has shed light on the circumstances of the initial endosymbiosis event and led us to a better understanding of the mitochondrial origin and evolution.

References

Akhmanova A, Voncken F, van Alen T, van Hoek A, Boxma B, Vogels G, Veenhuiss M, Hackstein JHP. 1998. A hydrogenosome with a genome. Nature 396:527-528.

Andersson SG, Karlberg O, Canback B, Kurland CG. 2003. On the origin of mitochondria: a genomics perspective. Philos Trans R Soc Lond B Biol Sci 358:165-177; discussion 177-169.

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133-140.

Boussau B, Karlberg EO, Frank AC, Legault BA, Andersson SG. 2004. Computational inference of scenarios for alpha-proteobacterial genome evolution. Proc Natl Acad Sci U S A 101:9722-9727.

Bui ET, Bradley PJ, Johnson PJ. 1996. A common evolutionary origin for mitochondria and hydrogenosomes. Proc Natl Acad Sci U S A 93:9651-9656.

Cavalier-Smith T. 2009. Predation and eukaryote cell origins: a coevolutionary perspective. Int J Biochem Cell Biol 41:307-322.

Cavaliersmith T. 1987. Molecular Evolution - Eukaryotes with No Mitochondria. Nature 326:332-333.

Cavaliersmith T. 1989. Molecular Phylogeny - Archaebacteria and Archezoa. Nature 339:100-101.

Clark CG, Roger AJ. 1995. Direct evidence for secondary loss of mitochondria in Entamoeba histolytica. Proc Natl Acad Sci U S A 92:6518-6521.

Darwin C. 1859. On the origin of species by means of natural selection. London,: J. Murray.

Davidov Y, Huchon D, Koval SF, Jurkevitch E. 2006. A new alpha-proteobacterial clade of *Bdellovibrio*-like predators: implications for the mitochondrial endosymbiotic theory. Environmental Microbiology 8:2179-2188.

Davidov Y, Jurkevitch E. 2009. Predation between prokaryotes and the origin of eukaryotes. Bioessays 31:748-757.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361-375.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. Nature 440:623-630.

Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D, et al. 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol 21:1643-1660.

Ettema TJ, Andersson SG. 2009. The alpha-proteobacteria: the Darwin finches of the bacterial world. Biol Lett 5:429-432.

Finlay BJ, Span ASW, Harman JMP. 1983. Nitrate Respiration in Primitive Eukaryotes. Nature 303:333-336.

Fitzpatrick DA, Creevey CJ, McInerney JO. 2006. Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the *Rickettsiales.* Mol Biol Evol 23:74-85.

Galibert F, Finan TM, Long SR, Puhler A, Abola P, Ampe F, Barloy-Hubler F, Barnett MJ, Becker A, Boistard P, et al. 2001. The composite genome of the legume symbiont *Sinorhizobium meliloti.* Science 293:668-672.

Georgiades K, Madoui MA, Le P, Robert C, Raoult D. 2011. Phylogenomic analysis of *Odyssella thessalonicensis* fortifies the common origin of *Rickettsiales*, *Pelagibacter ubique* and *Reclimonas americana* mitochondrion. PLoS One 6:e24857.

Germot A, Philippe H, Le Guyader H. 1996. Presence of a mitochondrial-type 70-kDa heat shock protein in *Trichomonas vaginalis* suggests a very early mitochondrial endosymbiosis in eukaryotes. Proc Natl Acad Sci U S A 93:14614-14617.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. 1990. Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60-63.

Goldberg AV, Molik S, Tsaousis AD, Neumann K, Kuhnke G, Delbac F, Vivares CP, Hirt RP, Lill R, Embley TM. 2008. Localization and functionality of microsporidian iron-sulphur cluster assembly proteins. Nature 452:624-628.

Hackstein JHP, Akhmanova A, Boxma B, Harhangi HR, Voncken FGJ. 1999. Hydrogenosomes: eukaryotic adaptations to anaerobic environments. Trends in Microbiology 7:441-447.

Hirt RP, Logsdon JM, Healy B, Dorey MW, Doolittle WF, Embley TM. 1999. *Microsporidia* are related to Fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. Proceedings of the National Academy of Sciences of the United States of America 96:580-585.

Keeling PJ, Luker MA, Palmer JD. 2000. Evidence from beta-tubulin phylogeny that *microsporidia* evolved from within the fungi. Molecular Biology and Evolution 17:23-31.

Kobayashi M, Matsuo Y, Takimoto A, Suzuki S, Maruo F, Shoun H. 1996. Denitrification, a novel type of respiratory metabolism in fungal mitochondrion. Journal of Biological Chemistry 271:16263-16267.

Koonin EV. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. Genome Biol 11:209.

Kurland CG, Andersson SGE. 2000. Origin and evolution of the mitochondrial proteome. Microbiology and Molecular Biology Reviews 64:786-+.

Mai ZM, Ghosh S, Frisardi M, Rosenthal B, Rogers R, Samuelson J. 1999. Hsp60 is targeted to a cryptic mitochondrion-derived organelle ("crypton") in the microaerophilic protozoan parasite *Entamoeba histolytica*. Molecular and Cellular Biology 19:2198-2205.

Martin W, Muller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392:37-41.

Muller M. 1993. The Hydrogenosome. Journal of General Microbiology 139:2879-2889.

Peyretaillade E, Broussolle V, Peyret P, Metenier G, Gouy M, Vivares CP. 1998. *Microsporidia*, amitochondrial protists, possess a 70-kDa heat shock protein gene of mitochondrial evolutionary origin. Mol Biol Evol 15:683-689.

Raoult D, Roux V. 1997. Rickettsioses as paradigms of new or emerging infectious diseases. Clinical Microbiology Reviews 10:694-&.

Rodriguez-Ezpeleta N, Embley TM. 2012. The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. PLoS One 7:e30520.

Roger AJ, Clark CG, Doolittle WF. 1996. A possible mitochondrial gene in the early-branching amitochondriate protist *Trichomonas vaginalis*. Proc Natl Acad Sci U S A 93:14618-14622.

Roger AJ, Svard SG, Tovar J, Clark CG, Smith MW, Gillin FD, Sogin ML. 1998. A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. Proc Natl Acad Sci U S A 95:229-234.

Rosenthal B, Mai Z, Caplivski D, Ghosh S, de la Vega H, Graf T, Samuelson J. 1997. Evidence for the bacterial origin of genes encoding fermentation enzymes of the amitochondriate protozoan parasite *Entamoeba histolytica*. J Bacteriol 179:3736-3745.

Sagan L. 1967. On the origin of mitosing cells. J Theor Biol 14:255-274.

Sassera D, Lo N, Epis S, D'Auria G, Montagna M, Comandatore F, Horner D, Pereto J, Luciano AM, Franciosi F, et al. 2011. Phylogenomic evidence for the presence of a flagellum and cbb(3) oxidase in the free-living mitochondrial ancestor. Mol Biol Evol 28:3285-3296.

Stouthamer R, Breeuwer JA, Hurst GD. 1999. *Wolbachia pipientis*: microbial manipulator of arthropod reproduction. Annu Rev Microbiol 53:71-102.

Takaya N, Suzuki S, Kuwazaki S, Shoun H, Maruo F, Yamaguchi M, Takeo K. 1999. Cytochrome p450nor, a novel class of mitochondrial cytochrome P450 involved in nitrate respiration in the fungus *Fusarium oxysporum.* Arch Biochem Biophys 372:340-346.

Thrash JC, Boyd A, Huggett MJ, Grote J, Carini P, Yoder RJ, Robbertse B, Spatafora JW, Rappe MS, Giovannoni SJ. 2011. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Sci Rep 1:13.

Tielens AGM, Rotte C, van Hellemond JJ, Martin W. 2002. Mitochondria as we don't know them. Trends in Biochemical Sciences 27:564-572.

Tovar J, Fischer A, Clark CG. 1999. The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite *Entamoeba histolytica*. Molecular Microbiology 32:1013-1021.

Tovar J, Leon-Avila G, Sanchez LB, Sutak R, Tachezy J, van der Giezen M, Hernandez M, Muller M, Lucocq JM. 2003. Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. Nature 426:172-176.

Tsaousis AD, Kunji ER, Goldberg AV, Lucocq JM, Hirt RP, Embley TM. 2008. A novel route for ATP acquisition by the remnant mitochondria of *Encephalitozoon cuniculi*. Nature 453:553-556.

van der Giezen M, Rechinger KB, Svendsen I, Durand R, Hirt RP, Fevre M, Embley TM, Prins RA. 1997. A mitochondrial-like targeting signal on the hydrogenosomal malic enzyme

from the anaerobic fungus *Neocallimastix frontalis*: support for the hypothesis that hydrogenosomes are modified mitochondria. Mol Microbiol 23:11-21.

Vellai T, Takacs K, Vida G. 1998. A new aspect to the origin and evolution of eukaryotes. J Mol Evol 46:499-507.

Viklund J, Ettema TJ, Andersson SG. 2012. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. Mol Biol Evol 29:599-615.

Viklund J, Martijn J, Ettema TJ, Andersson SG. 2013. Comparative and Phylogenomic Evidence That the *Alphaproteobacterium HIMB59* Is Not a Member of the Oceanic SAR11 Clade. PLoS One 8:e78858.

Voncken F, Boxma B, Tjaden J, Akhmanova A, Huynen M, Verbeek F, Tielens AGM, Haferkamp I, Neuhaus HE, Vogels G, et al. 2002. Multiple origins of hydrogenosomes: functional and phylogenetic evidence from the ADP/ATP carrier of the anaerobic chytrid *Neocallimastix sp*. Molecular Microbiology 44:1441-1454.

Williams KP, Sobral BW, Dickerman AW. 2007. A robust species tree for the alphaproteobacteria. J Bacteriol 189:4578-4586.

Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, Brownlie JC, McGraw EA, Martin W, Esser C, Ahmadinejad N, et al. 2004. Phylogenomics of the reproductive parasite *Wolbachia pipientis wMel*: a streamlined genome overrun by mobile genetic elements. PLoS Biol 2:E69.

Zumft WG. 1997. Cell biology and molecular basis of denitrification. Microbiology and Molecular Biology Reviews 61:533-+.

Chapter 1. A phylum level bacterial phylogenetic marker database[1]

[1]Formatted as a co-authored manuscript and published as:

Wang Z, Wu M. Mol Biol Evol (2013) doi:10.1093/molbev/mst059

Referenced supplementary material is available online at:

http://mbe.oxfordjournals.org/content/early/2013/03/21/molbev.mst059.abstract

Abstract

Large-scale, genome-level molecular phylogenetic analyses present both opportunities and challenges for bacterial evolutionary and ecological studies. We constructed a phylum-level bacterial phylogenetic marker database by surveying all complete bacterial genomes and identifying single-copy genes that were widely distributed in each of the 20 bacterial phyla. We showed that phylum trees made using these markers were highly resolved and were more robust than the bacterial genome tree based on 31 universal bacterial marker genes. In addition, using the Global Ocean Sampling dataset as an example, we demonstrated that the expanded marker database greatly increased the power of metagenomic phylotyping. We incorporated the database into an automated phylogenomic inference application (Phyla-AMPHORA) and made it publicly available. We believe that this centralized resource should have broad applicability in bacterial systematics, phylogenetics and metagenomic studies.

Introduction

A robust phylogenetic framework provides the foundation for bacterial systematics, evolutionary and diversity studies. The small subunit ribosomal RNA (SSU rRNA or 16S rRNA) has long been the marker of choice in bacterial phylogenetics because it is ubiquitously distributed, easy to PCR and sequence, and shows little evidence of lateral gene transfer (LGT). However, the bacterial tree of life based on a single gene is usually not well resolved because a single gene does not contain sufficient phylogenetic signal to resolve either the ancient or very recent relationships. In addition, because a gene usually represents no more than 0.1% of an average bacterial genome, it has been questioned whether one gene can adequately represent the evolutionary history of a genome (Dagan and Martin 2006).

The explosion in the number of sequenced bacterial genomes brings the opportunity for using protein-coding genes for genome-level phylogenetic analysis, also known as phylogenomics (Eisen and Fraser 2003; Delsuc et al. 2005). It is expected that with many more genes, "genome trees" will be more robust than the individual gene trees because of the increased signal to stochastic noise ratio (Jeffroy et al. 2006). Studies attempting to reconstruct the bacterial tree of life have demonstrated the power of this approach (Brown et al. 2001; Brochier et al. 2002; Ciccarelli et al. 2006; Wu and Eisen 2008; Wu et al. 2009; Yutin et al. 2012) (for review see (Delsuc et al. 2005)). In these studies, generally several dozens of orthologous genes that are universally distributed in the bacteria domain were used.

Dense sampling of bacterial genomes made it possible to identify phylum-level phylogenetic marker genes and use them to reconstruct genome trees for several major bacterial groups such as α-proteobacteria (Williams et al. 2007), γ-proteobacteria (Lerat et al. 2003; Williams et al.

2010) and Cyanobacteria (Swingley et al. 2008; Criscuolo and Gribaldo 2011). Because typically several hundred marker genes were identified in each phylum, the phylum genome trees were found to be highly resolved and more robust than the bacterial genome trees based on several dozens of universal bacterial markers (hereafter referred as the universal genome trees). Although extremely high statistical support is common with long concatenated alignments and should be viewed with caution (Phillips et al. 2004), the phylum genome trees are in general congruent with the 16S rRNA tree, the individual marker gene trees and the universal genome trees, suggesting that they represent a central trend of the shared vertical inheritance of these genes. Not surprisingly, these phylum-specific, single-copy marker genes were found to be rarely laterally transferred (Lerat et al. 2003; Williams et al. 2007; Williams et al. 2010; Abby et al. 2012).

Identifying additional phylogenetic markers has great implications for metagenomic studies. One main goal of metagenomic studies is to determine what species are present in the community and their biological functions. Microbial species composition can be estimated by phylotyping single-copy marker genes. If a marker gene happens to be located in a sequence contig, then the entire contig can be anchored to a specific taxonomic clade, allowing us to determine which species is capable of performing what functions in the community. In theory, using more marker genes is always better because it means more sequences can be phylotyped and anchored. However, in light of potentially pervasive LGTs in bacteria (Ochman et al. 2000), only genes recalcitrant to LGT should be used as markers in this process. Previously we have demonstrated the power of phylotyping with 31 universal bacterial marker genes (Wu and Eisen 2008). The power of phylotyping can be increased by including phylum-level phylogenetic markers.

Although previous studies have identified phylum-level markers for several major bacterial groups, different procedures were used in each study. Currently there is no centralized resource of the phylum-level phylogenetic markers that researchers can readily use for large-scale phylogenomic analyses. Here we conducted a comprehensive survey of all complete bacterial genomes, identified phylum-specific marker genes in 20 bacterial phyla and incorporated them into an automated phylogenomic inference application for bacterial systematics, evolutionary and diversity studies.

New Approaches

Identifying phylum-level bacterial phylogenetic markers

The workflow of the marker gene identification and verification is shown in Figure 1. Since ubiquitous, single-copy genes show little evidence of LGT, we first searched each phylum for genes with these two attributes. We further excluded genes that showed signs of LGT as estimated by the Prunier program (Abby et al. 2010). The Proteobacteria phylum was split into five groups (α-, β-, γ-, δ- and ε-) and each group was treated as a 'phylum' in this study. In total, we identified 7542 marker genes from 1982 complete bacterial genomes belonging to 20 phyla (Table 1). The number of marker genes in each phylum obviously depends on the coding capacity of the member species of each phylum. For example, Tenericutes consists of exclusively intracellular bacteria with largely reduced genomes, therefore the number of Tenericutes markers is limited. The size of the marker gene pool also correlates negatively with the phylogenetic diversity of the genomes sequenced within each phylum (Supplementary Figure 1). The poorer the phylum is sampled, the larger the number of markers will be shared among the sequenced members. For example, the poorly sampled phyla Aquificae, Fusobacteria, Planctomycetes and Chlorobi all have relatively large sets of markers. As the sampling gaps are

filled by additional genome sequencing, we expect the number of marker genes in these phyla to drop substantially and level off at about 8% of the genomes (Supplementary Figure 1).

The phylum-level phylogenetic marker database

The phylum-level marker genes and their function descriptions are listed in the Supplementary File 1. To facilitate the use of the marker genes for phylogenetic analysis, we built a database in which each marker gene is associated with four files: a 'seed' sequence alignment, a profile Hidden Markov Model (profile HMM), a mask for the alignment and a gene tree. HMMs form the cornerstones of the database and offer four main advantages in large-scale phylogenetic analysis: 1. HMM based sequence similarity search is as fast as BLAST but is more sensitive (Eddy 2011). 2. HMM based sequence alignment is highly accurate and runs much faster than all *de novo* multiple sequence alignment programs such as MUSCLE, CLUSTAL, T-COFFEE and MAFFT. 3. HMM based alignment has a unique feature in that new sequences can be aligned to the HMM's 'seed' alignment, residue by residue. Therefore, the newly generated alignments can be automatically trimmed using pre-computed quality scores of each position of the 'seed' alignment (the mask), producing high-quality alignments without requiring manual curation. For each marker in the database, a mask file has been generated using the probabilistic masking program ZORRO (Wu et al. 2012). 4. HMM is the only variable in HMM-based alignment. This means that sequence alignments produced using the same HMM are always compatible, making comparisons between different phylogenetic studies simple and straightforward.

The phylum-level marker database has been incorporated into the Phyla-AMPHORA package for automated high-throughput, high-quality phylogenomic analysis. From a given set of genomic or metagenomic sequences, Phyla-AMPHORA can identify each marker gene in the database and align them to the orthologous sequences of the complete bacterial genomes. Users

can then proceed to either make "genome trees" or assign phylotypes to the newly identified metagenomic sequences.

Robust bacterial phylum-level genome trees

Using Phyla-AMPHORA, we reconstructed phylum genome trees for each of the 20 bacterial phyla (Supplementary File 2) and compared them to universal bacterial genome trees made using 31 universal markers from the same set of genomes (Wu and Eisen 2008). We used the Congruence Among Distance Matrices (CADM) (Campbell et al. 2011) to measure the congruence of the trees. The CADM test takes the tree branch length into account and its *W* statistic score ranges from 0 (no congruence) to 1 (complete congruence). Remarkably, the phylum trees and the universal bacterial genome trees were highly concordant (W = [0.961, 0.997]) (Table 1). The discordant lineages mostly consisted of very closely related taxa (e.g., different strains of the same species) that were simply unresolved in the universal trees. Our results support previous studies showing that despite LGT being an important force in bacterial evolution, a bacteria tree of life tracing the vertical inheritance history can be reconstructed if a set of carefully selected markers are used (Lerat et al. 2003; Williams et al. 2007; Wu and Eisen 2008; Williams et al. 2010; Abby et al. 2012). Consistently, the phylum trees of this study are also highly congruent with a bacterial genome tree made from 50 universal ribosomal genes (Yutin et al. 2012) (Supplementary Table 1) and the phylum genome trees published recently (Abby et al. 2012) (Supplementary Table 2).

Although congruent, the phylum trees are more robust than the universal tree. For 7 out of 20 phyla, the average bootstrap values of the phylum trees were significantly higher than that of the universal trees (t-test $p < 0.05$). The improvement was not only apparent among the much better resolved closely related lineages (Supplementary Figure 2), but also evident in the overall better

supported relationships throughout the trees (Supplementary Figure 3). Notably, the number of weakly supported nodes (bootstrap value < 80) also decreased substantially in the phylum trees (Table 1).

The reasons for the increased robustness of the phylum trees were at least two folds. Firstly, the number of markers used in the genome tree reconstruction expanded ~12 times from 31 to 377 on average (Table 1), thus greatly increasing the number of informative sites for phylogenetic inference. Secondly, the phylum-specific marker genes evolve more rapidly than the 31 universal genes, as evidenced by the longer branch lengths of the phylum trees (Figure 2). On average, the amino acid substitution rates of the phylum-specific markers were 2.04 times of those of the 31 universal markers, thus increasing the amount of phylogenetic signal per site.

Phylum-level phylogenetic markers increase the power of phylotyping

Using Phyla-AMPHORA, we reanalyzed the environmental shotgun sequences collected from the Global Ocean Sampling (GOS). The GOS collection contains 6,115,812 peptides predicted from the assembled sequence reads (Rusch et al. 2007). AMPHORA was able to phylotype 1.4% of the peptides using the 31 universal bacterial markers. In comparison, Phyla-AMPHORA identified 814,916 phylum-specific marker sequences that corresponded to 13.3% of the whole dataset, thus increasing the number of sequences that could be phylotyped and anchored by ~10 times.

Conclusion

We believe that the phylogenetic resource reported in this study will facilitate large-scale bacterial phylogenetic analyses. Therefore, it has a great potential to be used in many areas of microbial evolutionary and ecological studies.

Material and Methods

Identifying universal phylogenetic markers within each phylum

To reduce the computational cost, we used a two-phased approach. In phase I, representative genomes were selected from each phylum to maximize the phylogenetic diversity using the greedy algorithm described in (Steel 2005). An all-against-all BLASTP search with an e-value cutoff of 1e-7 was performed among the representatives of the same phylum. Proteins were then clustered into families using the Markov Cluster Algorithm (MCL) using an e-value cutoff of 1e-15 (Enright et al. 2002). Families with an average of 1.00 ± 0.20 genes per organism and present in at least 80% of the representative organisms were chosen for further analysis. In phase II, HMMs were built for each protein family and were then used to identify homologs in the full set of genomes using HMMer3 with an evalue cutoff of 1e-15 (Eddy 2011). A second round all-against-all BLASTP search and MCL clustering were performed. Families with an average of 1.00 ± 0.06 genes per organism and present in at least 88% of the all phylum members were selected as marker genes for each phylum. The distribution parameter cutoff values were calibrated using the 31 universal bacterial genes (Wu and Eisen 2008). Next we screened and removed marker genes that might have undergone LGT using the Prunier program (Abby et al. 2010). We also removed genes that belonged to the large families of ABC transporter ATP-binding proteins, GTP-binding proteins and histidine kinases. Prunier analysis, phylum genome tree reconstructions and comparison, and GOS phylotyping were described in the Supplementary

Materials. The marker database can be downloaded as part of the Phyla-AMPHORA package from http://wolbachia.biology.virginia.edu/WuLab/Software.html.

Acknowledgements

We would like to thank Alexandra J. Scott for her help on the initial sequence analyses.

References

Abby SS, Tannier E, Gouy M, Daubin V. 2010. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics* 11:324.

Abby SS, Tannier E, Gouy M, Daubin V. 2012. Lateral gene transfer as a support for the tree of life. *Proc Natl Acad Sci U S A.* 109:4962-4967.

Brochier C, Bapteste E, Moreira D, Philippe H. 2002. Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet.* 18:1-5.

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. 2001. Universal trees based on large combined protein sequence data sets. *Nat Genet.* 28:281-285.

Campbell V, Legendre P, Lapointe FJ. 2011. The performance of the congruence among distance matrices (cadm) test in phylogenetic analysis. *BMC Evol Biol.* 11:64.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283-1287.

Criscuolo A, Gribaldo S. 2011. Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. *Mol Biol Evol.* 28:3019-3032.

Dagan T, Martin W. 2006. The tree of one percent. *Genome Biol.* 7:118.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet.* 6:361-375.

Eddy SR. 2011. Accelerated profile hmm searches. *PLoS Comput Biol.* 7:e1002195.

Eisen JA, Fraser CM. 2003. Phylogenomics: Intersection of evolution and genomics. *Science* 300:1706-1707.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30:1575-1584.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: The beginning of incongruence? *Trends Genet.* 22:225-231.

Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes:The case of the γ-proteobacteria. *PLoS Biol.* 1:e9.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299-304.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol.* 21:1455-1458.

Rusch DB, Halpern AL, Sutton G, et al. 2007. The sorcerer II global ocean sampling expedition: Northwest atlantic through eastern tropical pacific. *PLoS Biol.* 5:e77.

Steel M. 2005. Phylogenetic diversity and the greedy algorithm. *Syst Biol.* 54:527-529.

Swingley WD, Blankenship RE, Raymond J. 2008. Integrating markov clustering and molecular phylogenetics to reconstruct the cyanobacterial species tree from conserved protein families. *Mol Biol Evol.* 25:643-654.

Williams KP, Gillespie JJ, Sobral BW, Nordberg EK, Snyder EE, Shallom JM, Dickerman AW. 2010. Phylogeny of gammaproteobacteria. *J Bacteriol.* 192:2305-2314.

Williams KP, Sobral BW, Dickerman AW. 2007. A robust species tree for the alphaproteobacteria. *J Bacteriol.* 189:4578-4586.

Wu D, Hugenholtz P, Mavromatis K, et al. 2009. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature* 462:1056-1060.

Wu M, Chatterji S, Eisen JA. 2012. Accounting for alignment uncertainty in phylogenomics. *PLoS ONE* 7(1): e30288.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* 9:R151.

Yutin N, Puigbo P, Koonin EV, Wolf YI. 2012. Phylogenomics of prokaryotic ribosomal proteins. *PLoS ONE* 7:e36972.

Figures

Figure 1. The workflow of phylum-level phylogenetic marker gene identification and verification processes. The steps are described in details in the Material and Methods section.

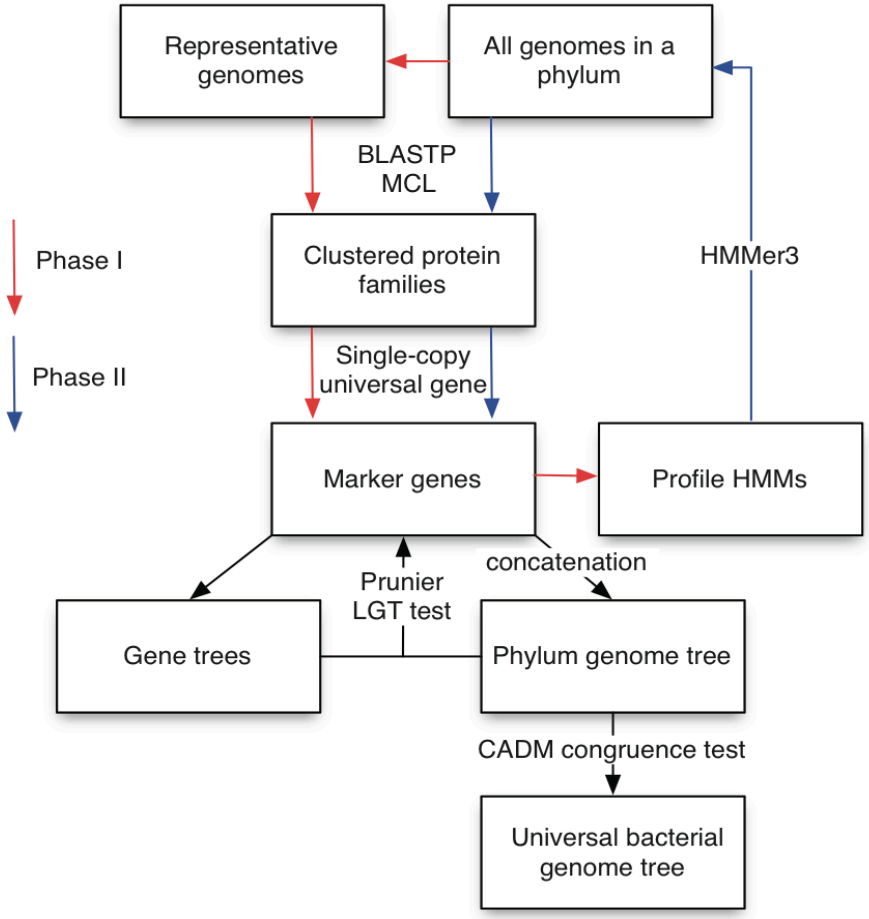
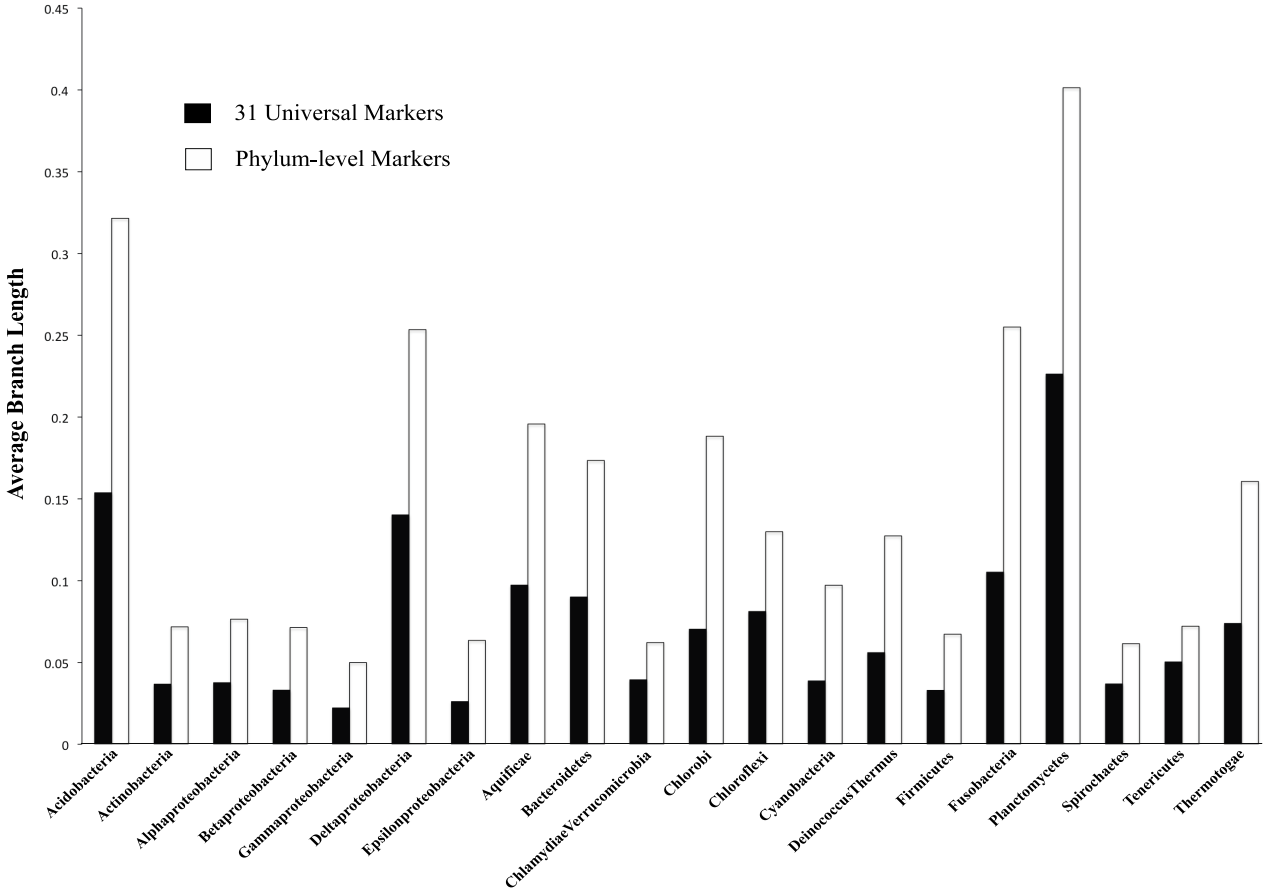


Figure 2. Phylum-specific marker genes evolve faster than the universal bacterial marker genes. For each phylum, the average branch lengths of the universal tree and the phylum-level tree are shown. The unit of the branch length is the number of amino acid substitutions per site.



Tables

Table 1. Overview of the phylum-level phylogenetic markers.

Phylum	Genomes	Markers	CADM *W* score	Average Bootstrap		Weakly Supported Nodes	
				Universal	Phylum	Universal	Phylum
Acidobacteria	9	559	0.997	100.0	100.0	0	0
Actinobacteria	222	218	0.996	88.4	95.9*	12	3
Alphaproteobacteria	214	200	0.976	93.5	97.8*	5	3
Betaproteobacteria	133	303	0.983	93.2	97.7*	9	2
Gammaproteobacteria	447	295	0.992	88.5	97.3*	11	2
Deltaproteobacteria	50	174	0.986	95.9	96.7	3	3
Epsilonproteobacteria	78	454	0.961	75.3	86.6	14	8
Aquificae	9	562	0.994	99.8	100.0	0	0
Bacteroidetes	82	215	0.984	89.6	97.8*	10	2
Chlamy/Verru [a]	54	248	0.988	93.9	95.0	3	2
Chlorobi	13	808	0.993	98.9	100.0	0	0
Chloroflexi	16	198	0.991	92.6	100.0	2	0
Cyanobacteria	43	499	0.973	93.0	99.9*	6	0
Deinococcus/Thermus	18	517	0.988	95.6	100.0	1	0
Firmicutes	455	168	0.993	84.5	90.0	19	9
Fusobacteria	5	470	0.982	91.5	89.0	0	1
Planctomycetes	6	849	0.989	100.0	100.0	0	0
Spirochaetes	50	160	0.992	99.3	99.8	0	0
Tenericutes	63	114	0.979	92.2	98.5*	6	1
Thermotogae	15	531	0.988	99.7	100.0	0	0

* Phyla with significantly increased average bootstrap support in phylum trees compared to the universal tree (t-test $p < 0.05$).

a. Chlamy/Verru: Chlamydiae/Verrucomicrobia

Supplementary Materials

Supplementary Notes

Supplementary Note 1. Material and Methods

LGT detection by Prunier

Prunier detects LGT by reconciling gene trees with a reference tree (e.g., species tree). For each protein family, a maximum likelihood tree was made using RAxML with the best model selected by RAxML (Stamatakis 2006) and was bootstrapped with 100 replicates. Gene trees were then compared to their corresponding phylum genome tree. Only highly supported branches (bootstrap support >=90) were used for LGT detection in the Prunier analysis. Protein families estimated to have at least 2 LGT events at the leaf branch or 1 LGT event at the internal branch by Prunier were purged from the marker database.

Phylum genome tree reconstruction and comparison

Phylum genome trees were reconstructed using RAxML with the best models selected by RAxML (Stamatakis 2006). Congruence Among Distance Matrices (CADM) was used to compare the phylum genome trees to 1) genome trees of the same genomes reconstructed using the 31 universal bacterial marker genes (Wu, Eisen 2008), 2) a bacterial genome tree of 996 genomes reconstructed using 50 universal ribosomal proteins (Yutin et al. 2012) and 3) the phylum trees reported in (Abby et al. 2012). The CADM global test was performed using the Analyses of Phylogenetics and Evolution (APE) package in R (Paradis, Claude, Strimmer 2004). TOPD/FMTS was used to pinpoint the specific nodes that were different between two trees (Puigbo, Garcia-Vallve, McInerney 2007).

GOS Phylotyping

Peptide sequences predicted from all assembled sequences of the GOS study were downloaded from the Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis (CAMERA, http://camera.calit2.net). Marker identification and phylotyping were carried out using the Phyla-AMPHORA package with the default parameters running on a Linux server with 8 Intel Xeon 2.67 GHz processors and 32 GB of memory. It took 10 days to phylotype the GOS dataset.

References

Abby, SS, E Tannier, M Gouy, V Daubin. 2012. Lateral gene transfer as a support for the tree of life. Proceedings of the National Academy of Sciences of the United States of America 109:4962-4967.

Paradis, E, J Claude, K Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289-290.

Puigbo, P, S Garcia-Vallve, JO McInerney. 2007. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics 23:1556-1558.

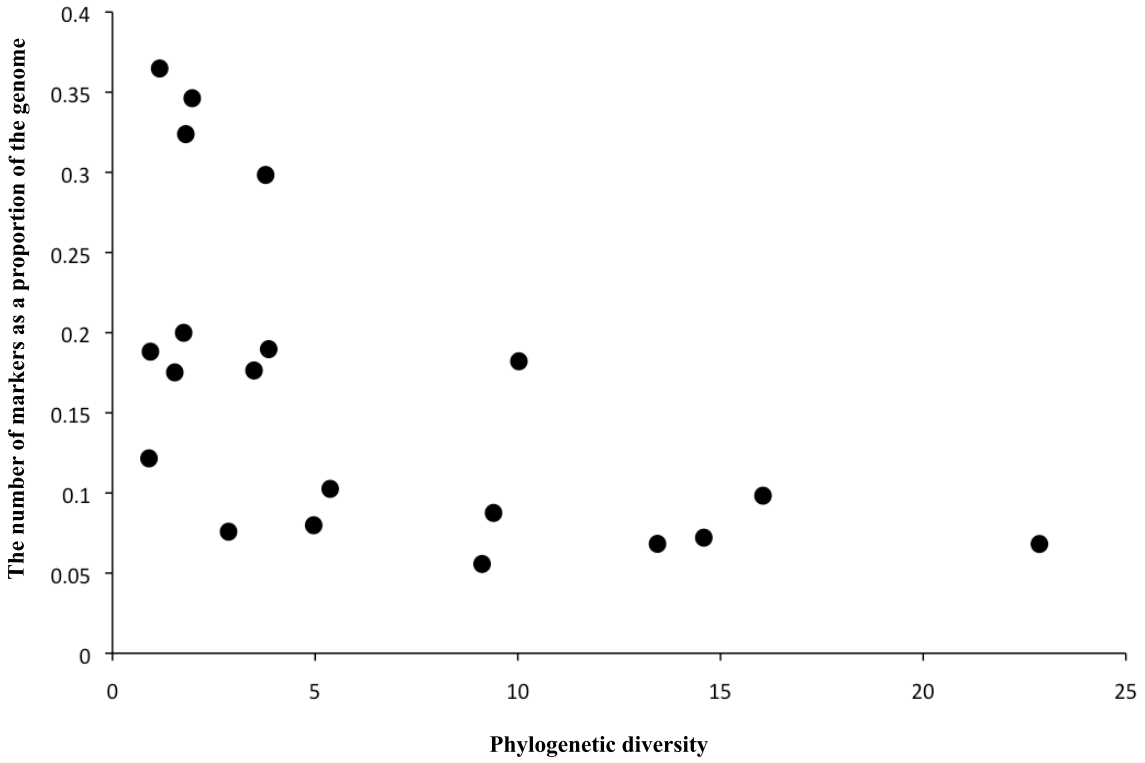
Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Wu, M, JA Eisen. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biology 9:R151.

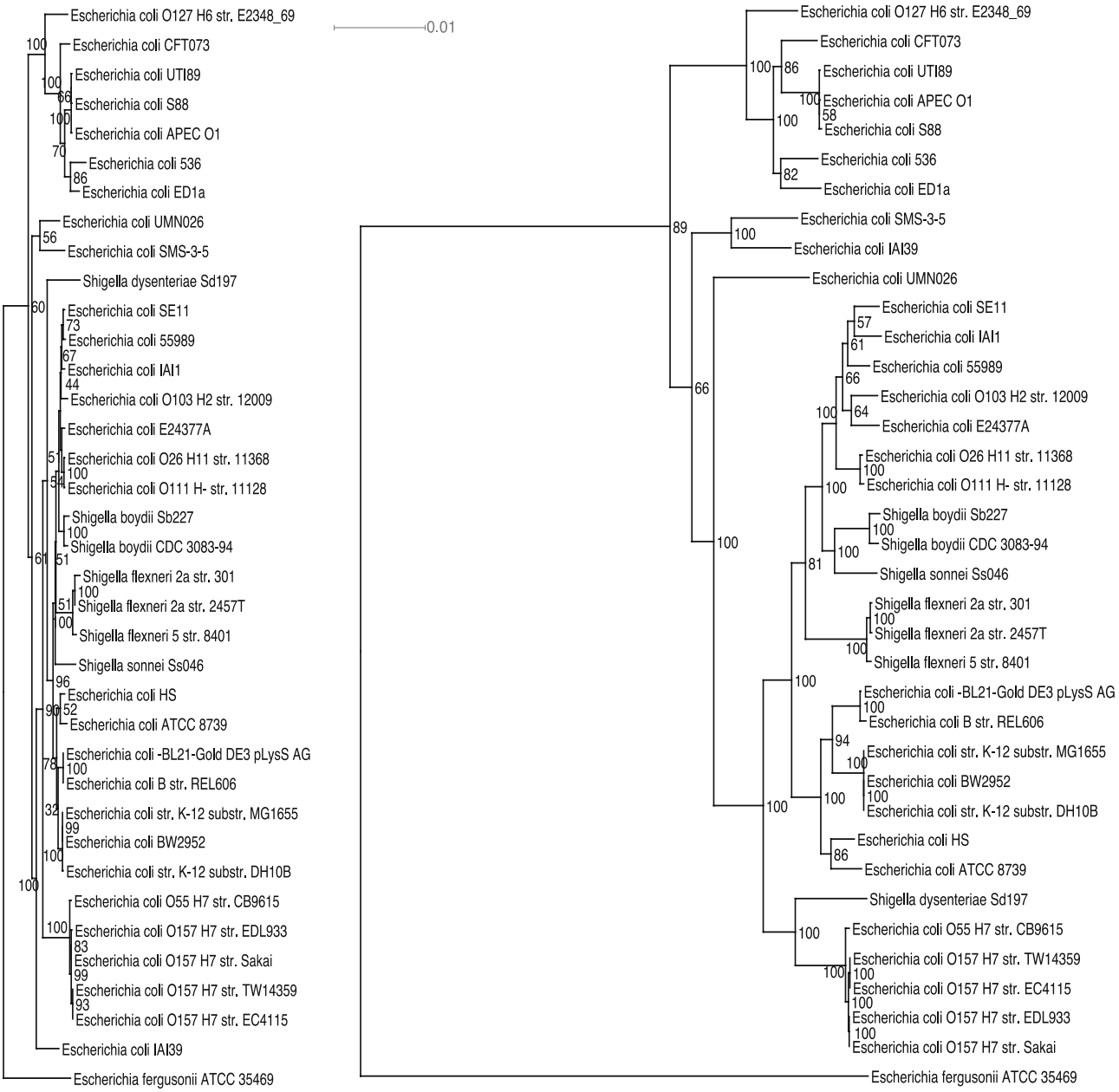
Yutin, N, P Puigbo, EV Koonin, YI Wolf. 2012. Phylogenomics of prokaryotic ribosomal proteins. PLoS ONE 7:e36972.

Supplementary Figures

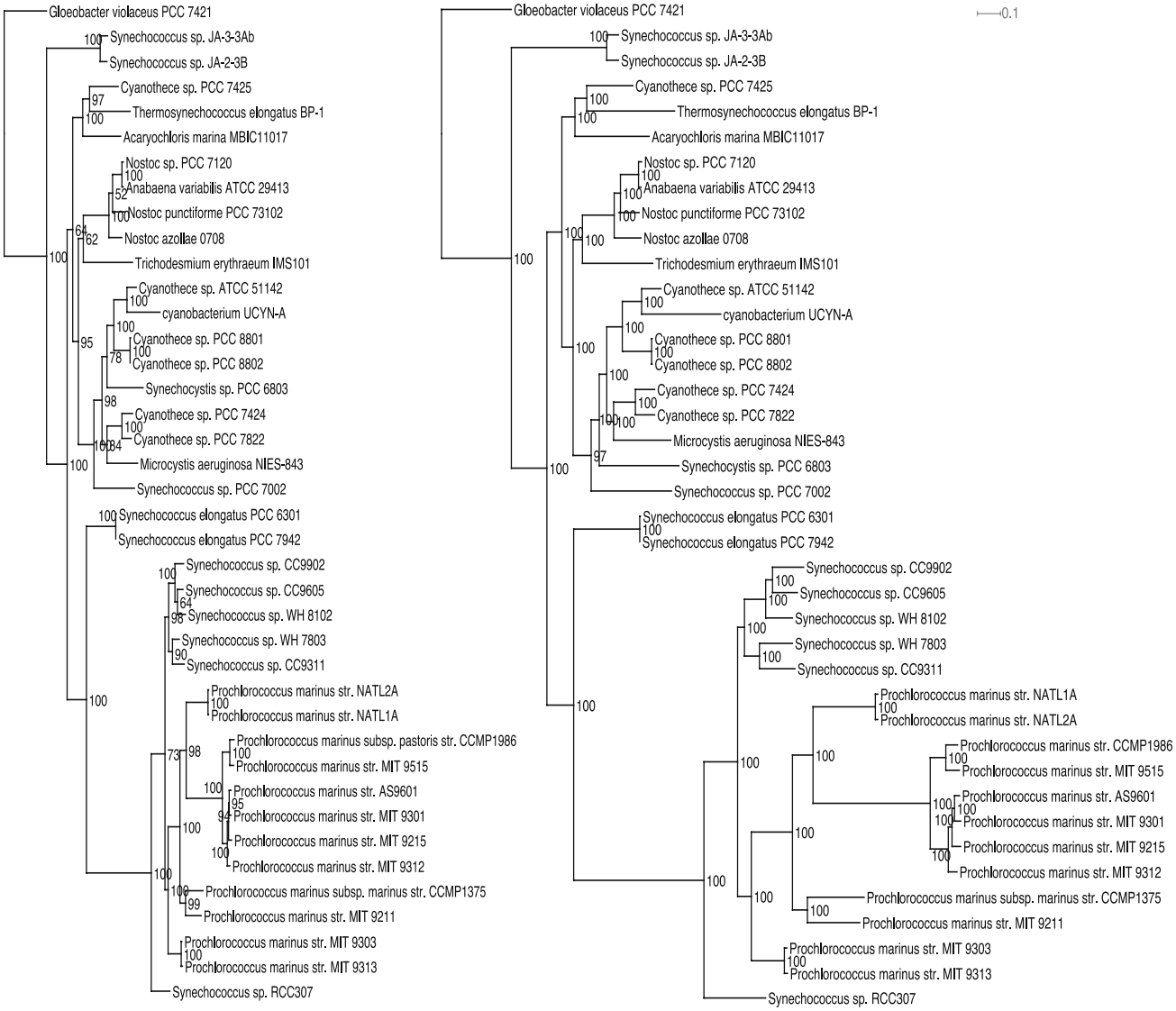
Supplementary Figure 1. The proportion of the genes identified as phylum-specific markers correlates negatively with the phylogenetic diversity of the genomes sequenced within each phylum. Each data point represents a phylum. The phylogenetic diversity was measured using the total branch lengths of the phylum trees.



Supplementary Figure 2. The genome trees of Escherichia coli strains reconstructed using both 31 universal markers (left) and 295 phylum-level markers (right) with 100 bootstrap replicates. The phylum-level tree is better resolved with significantly increased branch lengths and bootstrap supporting values.



Supplementary Figure 3. The Cyanobacteria genome trees reconstructed using both 31 universal markers (left) and 499 phylum-level markers (right) with 100 bootstrap replicates. The phylum-level tree is highly congruent with the universal tree but has significantly improved bootstrap support.



Supplementary Tables

Supplementary Table 1. Results of CADM congruence test between the phylum genome trees of this study and the bacterial genome tree made from 50 universal ribosomal genes (Yutin et al. 2012).

Major bacterial groups	CADM score
Acidobacteria	0.971
Actinobacteria	0.985
Alphaproteobacteria	0.977
Betaproteobacteria	0.962
Gammaproteobacteria	0.977
Deltaproteobacteria	0.958
Epsilonproteobacteria	0.960
Aquificae	0.985
Bacteroidetes	0.957
Chlamydiae/Verrucomicrobia	0.972
Chlorobi	0.984
Chloroflexi	0.984
Cyanobacteria	0.965
Deinococcus/Thermus	0.920
Firmicutes	0.990
Fusobacteria	0.943
Planctomycetes	1.000
Spirochaetes	0.961
Tenericutes	0.985
Thermotogae	0.995

Supplementary Table 2. Results of CADM congruence test between the phylum genome trees of this study and those of the Abby et al. study (Abby et al. 2012).

Major bacterial groups	**CADM score**
Alphaproteobacteria	0.989
Betaproteobacteria	0.980
Gammaproteobacteria	0.990
Deltaproteobacteria	0.979
Epsilonproteobacteria	0.992
Bacillales	0.980
Lactobacillales	0.992
Clostridia	0.973
Mollicutes	0.982
Bacteroidetes	0.984
Chlamydiae/Verrucomicrobia	0.998
Chlorobi	0.914
Actinobacteria	0.994
Cyanobacteria	0.986
Spirochaetes	1.000

Supplementary File:

Supplementary File 1. List of phylum-level marker genes with function descriptions in each of the 20 bacterial phyla.

Supplementary File 2. Newick-formatted genome trees of 20 bacterial phyla reconstructed using the phylum-level markers.

Supplementary File 1 and Supplementary File 2 are available at:

http://mbe.oxfordjournals.org/content/suppl/2013/03/21/mst059.DC1

Chapter 2. An integrated phylogenomic approach toward pinpointing the origin of mitochondria[1]

[1]Formatted as a co-authored manuscript (Zhang Wang and Martin Wu) in review at *Molecular Biology and Evolution*

Abstract

Overwhelming evidence supports the endosymbiosis theory that mitochondria originated once from the α-proteobacteria. However, its exact position in the tree of life remains highly debated. This is because systematic errors including sparse taxonomic sampling, sequence composition bias and high evolutionary rates have long plagued the mitochondrial phylogenetics. In this study, we address this issue by 1) increasing the taxonomic representation of α-proteobacterial genomes by sequencing 18 phylogenetically novel species. They include 5 *Rickettsiales* and 4 *Rhodospirillales*, two orders that have shown close affiliations with mitochondria previously, 2) using a set of 29 slowly evolving mitochondria-derived nuclear genes that are less biased than mitochondria-encoded genes as the alternative "well behaved" phylogenetic markers, 3) applying site heterogeneous mixture models that account for the sequence composition bias. With the integrated phylogenomic approach, we are able to for the first time place mitochondria unequivocally within the *Rickettsiales* order, as a sister clade to the *Rickettsiaceae* and *Anaplasmataceae* families, all subtended by the *Holosporaceae* family. Our results suggest that mitochondria most likely originated as an endosymbiont in the *Rickettsiales* lineage, but not from the distantly related free-living *Pelagibacter* and *Rhodospirillales*. In addition, the multiple diverse *Holosporaceae* genomes sequenced in this study will provide novel insights into the genetic complement of mitochondrial ancestor.

Introduction

The origin of mitochondria was a seminal event in the history of life. It is now widely accepted that mitochondria evolved only once from bacteria living within their host cells, probably two billion years ago (known as the endosymbiosis theory). Specifically, phylogenetic analyses have indicated that mitochondria originated from α-proteobacteria, a subgroup of the purple non-sulfur bacteria (Lang et al. 1999). However, exactly when it happened remains highly debated and this key piece of puzzle is still missing in our current assembly of the tree of life.

Defining precisely the α-proteobacterial ancestry of the mitochondria has important implications. It is a prerequisite for elucidating the origin and early evolution of mitochondria and eukaryotic cells. Placing mitochondria firmly within the tree of life will allow us to use comparative methods to gain insights into the biology of the last common ancestor of mitochondria and α-proteobacteria — Was it a free-living bacterium or an endosymbiont? What was its genetic makeup (Kurland and Andersson 2000; Gabaldon and Huynen 2003)? Did the mitochondrion arise at the same time as, or subsequent to, the appearance of the eukaryotic nucleus (Martin and Muller 1998)? Did it originate under initially anaerobic or aerobic conditions (Gray et al. 2001)? What was the driving force behind the initial symbiosis (Martin and Muller 1998; Kurland and Andersson 2000)?

Pinpointing the origin of mitochondria is inherently difficult, however, due to the compounding effects of at least three factors: 1) Weak phylogenetic signal. Most informative sites in the molecular sequence that allow us to resolve the deep evolutionary relationships have been erased by saturated mutations accumulated over a long period of time. As a result, individual genes such as the small subunit ribosomal RNA (SSU rRNA or 16S rRNA) usually do not contain sufficient

phylogenetic signals to resolve this deep relationship. 2) Long-branch attraction (LBA). Mitochondria and the obligate intracellular α-proteobacteria have highly accelerated rates of evolution than the free-living bacteria. Therefore, molecular phylogenetic inference of the origin of the mitochondria is prone to the well-known LBA artifact, when fast-evolving but distantly related lineages are erroneously grouped together as sister nodes in the tree (Felsenstein 1978; Hillis et al. 1994). 3) Extreme sequence composition bias. Mitochondria and the obligate intracellular α-proteobacteria are in general extremely AT rich in their genome sequences. It is well established that sequence composition bias could adversely affect the phylogenetic reconstruction and lead to statistically robust but misleading conclusions (Woese et al. 1991; Hasegawa and Hashimoto 1993; Foster and Hickey 1999).

Due to these reasons, results from early studies based on the sequences of a few genes were often inconclusive. Mitochondria have been placed near the *Rickettsiales* order, a subgroup of α-proteobacteria that contains obligate intracellular bacterial parasites such as *Rickettsia, Ehrlichia*, and *Anaplasma* (Viale and Arakaki 1994; Gupta 1995). And often, the *Rickettsia* genus was asserted to be the closest modern relative of mitochondria (Karlin and Brocchieri 2000; Emelyanov 2003). Phylogenomic analysis using 32 genes shared by mitochondria and bacteria called into question the conjecture that *Rickettsia* genus is the closest relative of mitochondria (Wu et al. 2004). Later it was suggested that *Rhodospirillum rubrum* within the *Rhodospirillales* order came as close to mitochondria as any α-proteobacteria investigated (Esser et al. 2004). Recent genome-level phylogenetic analyses with increasingly more bacterial species showed an emerging trend that places mitochondria basal to the *Rickettsiales* order with very high statistical support (Fitzpatrick et al. 2006; Williams et al. 2007; Georgiades et al. 2011; Thrash et al. 2011; Rodriguez-Ezpeleta and Embley 2012). However, who is the closest contemporary relative of mitochondria remains highly debated. Studies have suggested that a group of free-living bacteria

known as the SAR11 group form the sister clade to mitochondria (Georgiades et al. 2011; Thrash et al. 2011). Members of SAR11 dominate in the ocean surface water and have the smallest cells and genomes of any free-living organisms. A sister-clade relationship with the SAR11 group would suggest that mitochondria originated from free-living marine bacteria and the endosymbiosis events of mitochondria and intracellular *Rickettsiales* were independent. However, this hypothesis has been convincingly refuted by more recent studies demonstrating that this sister-clade relationship is a tree reconstruction artifact resulted from sequence composition bias (Brindefalk et al. 2011; Rodriguez-Ezpeleta and Embley 2012; Viklund et al. 2012).

Intriguingly, the conflicting sister-clade relationships of mitochondria all received high statistical support (Williams et al. 2007; Georgiades et al. 2011; Thrash et al. 2011; Rodriguez-Ezpeleta and Embley 2012). Obtaining a highly supported genome tree does not necessarily guarantee an accurate evolution reconstruction. It has been shown that highly supported branching patterns in a genome tree could be wrong because of unrealistic evolutionary models, composition biases in the sequence data, or the LBA (Jeffroy et al. 2006). Unlike the stochastic noise, systematic errors such as composition bias and LBA will not diminish but rather strengthen when more data of the same kind are added, ultimately leading the trees to converge toward the wrong tree with extremely high support (hence, be positively misleading) (Felsenstein 1978). It has been demonstrated by many studies that genome trees with high bootstrap, jackknife or posterior probability support should be treated with greater caution than single-gene trees for possible misleading tree reconstruction artifacts (Foster and Hickey 1999; Phillips et al. 2004; Soltis et al. 2004; Stefanovic et al. 2004; Delsuc et al. 2005; Lockhart and Penny 2005).

In this study, we first show that systematic errors in the current genome sequence dataset still present serious problems for precisely placing mitochondria in the tree of life. We then address the LBA and composition bias problems by 1) sequencing 18 strategically selected α-proteobacterial isolates to substantially increase the taxonomic representation of the α-proteobacterial genomes, 2) using a set of slowly evolving and less compositionally biased mitochondria-derived nuclear genes (compared to mitochondria-encoded genes) for phylogenetic reconstruction, 3) applying site heterogeneous mixture models that account for composition bias. With the integrated phylogenomic approach, we are able to place mitochondria firmly within the *Rickettsiales* order, as a sister clade to the *Rickettsiaceae*/*Anaplasmataceae* families, all subtended by the free-living α-proteobacterium HIMB59 and the *Holosporaceae* family.

Results

Substantial systematic errors are present in the current genomic sequence dataset

Because LBA and composition bias produce conflicting signals competing against the true phylogenetic signal, they can be detected using split-based methods (Bandelt and Dress 1992; Waddell et al. 1999; Lockhart and Cameron 2001; Clements et al. 2003). Split decomposition analysis produces a "neighbor net" where conflicting phylogenies are displayed as box-like structures. The more tree-like parts of the graph show where there is little conflict, and thus, little evidence of systematic errors. To determine whether there are significant systematic errors in the current genomic dataset, we performed a NeighborNet analysis on a concatenated protein sequence alignment of 26 mitochondria-encoded genes from genomes of 54 α-proteobacterial and 6 mitochondrial representatives. Figure 1 shows that α-proteobacteria can be divided into at least 7 major groups (*Rickettsiales, Rhodospirillales, Sphingomonadales, Rhodobacterales, Caulobacterales, Magnetococcales* and *Rhizobiales*), by and large consistent with the taxonomic

classification based on the SSU rRNA gene. Nevertheless, it also shows a large amount of networking or phylogenetic uncertainty around the base of mitochondria as observed previously (Esser et al. 2004; Fitzpatrick et al. 2006), indicating that the precise position of mitochondria within the α-proteobacteria is highly uncertain.

To further investigate the source of the systematic errors, we carried out spectral analysis. Spectral analysis is an extremely useful tool that can be used to pinpoint and quantify the source of errors independently of any one particular tree (Hendy and Penny 1993). If LBA is a problem, spectral analysis should indicate that there is support for two or more conflicting (i.e., mutually exclusive) splits, one of which grouping long-branch lineages together. Spectral analysis has been successfully applied to detect LBA in many datasets including mitochondrial genes (Lento et al. 1995; Kennedy et al. 1999; Mallatt and Winchell 2002; Kennedy et al. 2005; Wagele and Mayer 2007).

Figure 2 shows the split support spectrum of the same concatenated alignment used in the NeighborNet analysis. The strongest four splits are all compatible with the major groups shown in Figure 1, indicating that there is strong phylogenetic signal in the dataset. However, there are also substantial numbers of conflicting splits, many of them mutually incompatible. It is striking that incompatible splits in the top 50 splits are all associated with long-branch lineages (Supplementary Table 1). For example, most of these incompatible splits place a single mitochondrial species with long-branch lineages such as *Rickettsiales, Pelagibacter* and the outgroup (indicated by asterisks in Figure 2), but never with the "normal length" lineages. Conflicting splits placing a single species of *Rickettsiales* and *Pelagibacter* within other long-branch groups were also observed. The number of conflicting splits associated with each major group is shown in Figure 1. There is a strong correlation between the conflicting splits and the

long-branch lineages, indicating that LBA is a major source of errors in the current genomic dataset.

Increasing the phylogenetic diversity of α-proteobacterial genomes

Recent empirical phylogenomic studies have demonstrated that increasing taxon representation is very effective in mitigating LBA and improving the phylogenies (Philippe 1997; Stefanovic et al. 2004; Brinkmann et al. 2005; Delsuc et al. 2005; Leebens-Mack et al. 2005; Philippe et al. 2005; Yoon et al. 2008). At the beginning of this study, 425 α-proteobacterial genomes had been sequenced according to the GenomeOnline database (Pagani et al. 2012). However, most of them were selected from an anthropocentric point of view and did not take the phylogeny into consideration. As a result, many sequenced species were closely related and the taxonomic representation was extremely biased. For example, 220 or 52% of the sequenced α-proteobacterial genomes came from one single order (*Rhizobiales*). 123 of them were actually from one single genus (*Brucella*). On the other hand, for the *Rickettsiales* order that has shown close phylogenetic relationship to mitochondria, two families (*Holosporaceae* and *Incertae sedis 4*) were completely missing. Consequently, many gaps remain in the α-proteobacterial branch of the tree of life.

To fill the gaps in the tree, we selected α-proteobacterial species for sequencing by maximizing the total amount of phylogenetic diversity they represented. We estimated the phylogenetic diversity based on the SSU rRNA tree. Although not perfect, SSU rRNA has been shown to be a sound predictor of an organism's position in the genome tree (Wu et al. 2009). We downloaded the aligned SSU rRNA sequences of 9,817 α-proteobacterial isolates from the Ribosomal Database Project (Cole et al. 2009) and used them to construct a maximum likelihood tree. We

then used a tree-based greedy algorithm described in (Steel 2005) to rank isolates by their phylogenetic novelty. Species that had been sequenced were removed from the list. The availability of an isolate's genomic DNA was also an important factor in our selection process. In total, 18 species from six orders (*Rickettsiales, Rhodospirillales, Kordiimonadales, Magnetococcales, Rhizobiales* and *Rhodobacterales*) were selected for sequencing (Table 1, also highlighted in Figure 3). Together, they represented 18.5% of the phylogenetic diversity of the α-proteobacteria in the tree (Figure 3) and increased the phylogenetic diversity significantly compared to a random set of 18 genomes (1.7 - 3.0 times, p = 7e-65). We note that 9 of 18 selected species belong to the *Rickettsiales* and *Rhodospirillales* orders, which have shown close affiliation with mitochondria previously.

The 18 α-proteobacterial genomes were sequenced by whole-genome shotgun sequencing using a combination of 454 pyrosequencing and Illumina. The status and characteristics of the genomes are listed in Table 1.

Increasing the phylogenetic diversity reduced the systematic errors

We asked whether adding the 18 newly sequenced genomes reduced the systematic errors in the dataset. As shown in Figure 2, adding the 18 genomes visibly reduced the level of conflict in the split spectrum. Both the number of conflicting splits and their overall ranks decreased. Accordingly, the systematic errors in the dataset, calculated as the proportion of incompatible splits weighted by the supporting values, decreased from 0.377 to 0.266. The support for incompatible splits that grouped a single mitochondrial species within the *Rickettsiales* order also decreased. As a result, their ranks in the top 50 splits dropped. The improvement shows that the increased taxon sampling clearly has a positive effect on mitigating LBA.

Use of mitochondria-derived nuclear genes as alternative phylogenetic markers

As a consequence of their endosymbiotic lifestyle, mitochondria have gone through extensive genome reduction (Burger et al. 2003). For example, the 16 Kbp human mitochondrial genome only encodes 13 proteins (Anderson et al. 1981). A large fraction of mitochondrial genes have simply been lost, while many others have been transferred into the nucleus at the early stage (Kurland and Andersson 2000). Once in the nucleus, these genes would be no longer subject to the same evolutionary forces that have driven mitochondria evolution to an extreme. Consequently, these nuclear genes will be less derived and will not have evolution rates and GC biases as extreme as the mitochondria-encoded genes. In theory, trees made from these nuclear genes will be more recalcitrant to the LBA and composition bias that have plagued the phylogenetic analysis of mitochondria. In some sense, these genes could act as natural "time capsules" that when uncovered, will reveal cues about their distant past.

Mitochondria-to-nuclei gene transfers can be identified using a phylogenetic approach (Karlberg et al. 2000; Wu et al. 2004). Unlike many other lateral gene transfer events, here we have the rare benefit of knowing the donor and the acceptor in advance. Therefore, mitochondria-derived nuclear genes can be identified by looking for a seemingly anomaly in the gene trees — the placement of eukaryotic nuclear genes within the α-proteobacteria. Here we leveraged the large number of bacterial, eukaryotic and mitochondrial genomes that are now available to systematically identify mitochondria-derived nuclear genes.

The mitochondria-to-nuclei gene transfer is an ongoing process (Nugent and Palmer 1991; Covello and Gray 1992; Adams et al. 1999). Although there were parallel transfers, in general genes transferred at earlier stages should be found in a broader taxonomic range of eukaryotic nuclear genomes than these transferred at later stages. Therefore, genomes of phylogenetically

diverse eukaryotes, especially those from deep-branching eukaryotes, would be very useful for identifying the early transferred genes. We limit our phylogenomic analyses to these early-transferred genes as they are expected to be less derived than those transferred at a later stage. It will also be much easier to distinguish them from the spurious transfers that happened more recently (e.g., direct transfers from α-proteobacteria to the nucleus (Dunning Hotopp et al. 2007)).

We selected a set of 30 eukaryotic genomes that represented a broad range of taxonomic groups (Supplementary Table 2). From 2,527 eukaryotic protein families whose top BLAST hits included α-proteobacteria, our phylogenetic analysis identified 29 nuclear genes that were most likely transferred from the mitochondria early on (Table 2), as they were present in at least 8 diverse eukaryotic lineages.

Evaluation of phylogenetic marker genes and tree reconstruction methods

We compared the mitochondria-derived nuclear genes and mitochondria-encoded genes in terms of their sequence composition biases and substitution rates. To quantify the GC bias in the data, first we calculated aminoGC, the frequencies of amino acids (Gly, Ala and Pro) that are encoded by GC rich codons (Viklund et al. 2012). AminoGC essentially measures the effect of GC bias on the protein sequences. The mitochondrial proteins have significantly more extreme aminoGC than the nuclear proteins ($p<0.001$, Table 2), indicating that nuclear proteins are less biased than the mitochondrial proteins. We then measured the composition bias of the nuclear and mitochondrial sequences in the context of their α-proteobacterial homologs using chi-square scores. The larger the chi-square score, the stronger the composition bias. Table 2 shows that the composition bias of the nuclear sequences is substantially smaller than that of the mitochondrial sequences ($p<0.01$).

Next we compared the substitution rates of the nuclear and mitochondrial genes. In the RAxML genome tree made with mitochondrial markers, the average branch length from the root to mitochondria is 1.713 substitutions/site (stdev 0.225). In comparison, the average branch length from the root to eukaryotes is 1.273 substitutions/site (stdev 0.088) in the genome tree made with nuclear markers. Therefore, the nuclear genes evolved significantly slower than the mitochondrial genes ($p<0.01$). A similar result was observed when comparing the PhyloBayes trees. Taken all these together, it suggests that mitochondria-derived nuclear genes could be used as a set of alternative "well-behaved" markers to improve the mitochondrial phylogeny.

We carried out phylogenetic analyses using the concatenated protein sequences of the nuclear and mitochondrial marker genes respectively. As a reference, the analyses also included the phylum-level markers, a set of 200 single-copy marker genes that were shared by the α-proteobacteria (Wang and Wu 2013). We used both maximum likelihood and Bayesian methods to infer the phylogeny. To evaluate the effect of composition bias on the phylogeny, we applied the CAT mixture model in PhyloBayes to account for compositional heterogeneity. In contrast, the evolutionary models used to make RAxML maximum likelihood trees did not take compositional heterogeneity into account. Six unique combinations of datasets and methods yielded three different topologies (Figure 4 and Supplementary Figures 1-6). They differ primarily in the positions of the *Pelagibacter* and the *Holosporaceae* family, a group of mostly obligate endosymbionts in the protist acanthamoeba.

In all the RAxML trees, *Pelagibacter* forms a sister clade relationship with the *Rickettsiales*. It has been well demonstrated that this is a tree reconstruction artifact caused by sequence composition bias (Georgiades et al. 2011; Viklund et al. 2012; Viklund et al. 2013). Accordingly,

the PhyloBayes trees of different markers are in agreement with each other in that they all group *Pelagibacter* with the free-living α-proteobacteria. However, they differ in terms of the position of the *Holosporaceae*. Trees based on the 200 phylum-level markers and the nuclear markers are congruent and both place *Holosporaceae* within the *Rickettsiales*. The tree based on the mitochondrial markers, on the other hand, places *Holosporaceae* next to the free-living *Rhodospirillales*.

We then used gene order as an independent source of evidence to resolve the conflicting evolutionary relationships between *Holosporaceae*, *Pelagibacter* and *Rickettsiales*. In particular, we identified unique genome rearrangement events shared by *Holosporaceae* and other *Rickettsiales* in a number of gene clusters, which are otherwise highly syntenic between *Pelagibacter* and free-living α-proteobacteria. Figure 5 shows one such gene cluster encoding 12 proteins, most of which are involved in the TCA cycle and ATP synthesis. The 12 genes form a highly conserved cluster in *Pelagibacter* and free-living α-proteobacteria, with one deletion event occurred in *Pelagibacter* between genes *priA* and *pdhD*. However in *Holosporaceae* and *Rickettsia*, the gene cluster has been broken apart at several "hot spots". For example, the cluster was split on both sides of the *priA* gene in *Holosporaceae* and *Rickettsia*, and it was further split on both sides of the *sucCD* genes in *Rickettsia*. The similar gene order patterns in *Holosporaceae* and *Rickettsia* suggest that they are closely related and the genome rearrangement events likely occurred in their last common ancestor. Therefore, the independent gene order information supports placing the *Holosporaceae* with *Rickettsiales*, and *Pelagibacter* with the free-living α-proteobacteria. Based on the gene order information, we believe that the PhyloBayes trees of the phylum-level markers and the nuclear markers make more sense than the tree of the mitochondrial markers.

Assembly of the α-proteobacterial and mitochondrial branch of tree of life

Since mitochondria-derived nuclear genes have less composition bias, lower substitution rates and produce a phylogenetic tree that is consistent with the gene order patterns, we chose to use mitochondria-derived nuclear genes as the marker genes in our final phylogenomic analysis to infer the origin of mitochondria. We assembled a dataset of 29 genes from 72 diverse α-proteobacterial (including 18 sequenced in this study) and 6 eukaryotic representative genomes. The final concatenated protein sequence alignment consisted of 6,201 amino acids after the ambiguous alignment regions were removed using the program ZORRO (Wu et al. 2012). We used the CAT+GTR model in PhyloBayes to account for the compositional heterogeneity. Our genome tree divides the α-proteobacteria into at least 7 major groups, corresponding to 7 orders. It places mitochondria within *Rickettsiales* as a sister clade to the *Anaplasmataceae*/*Rickettsiaceae* families, all subtended by the free-living α-proteobacterium HIMB59 and the *Holosporaceae* family (Figure 6).

As a comparison, we also reconstructed genome trees with different combinations of datasets (the nuclear or mitochondrial markers), tree methods (RAxML or PhyloBayes) and data types (original or recoded). The trees are shown in Supplementary Figures S7-13.

Discussion

Placing mitochondria precisely in the tree of life has been problematic. Sparse taxonomic sampling, sequence composition biases, high evolutionary rates have all plagued the molecular phylogenetic inference of the origin of mitochondria. Here we address this issue with an integrated phylogenomic approach by using a broad taxonomic sampling, better-behaved marker genes and sophisticated models of sequence evolution.

Using NeighborNet and spectral analyses, we first demonstrated that there were significant systematic errors in the current genomic dataset. Of particular concern was the potential LBA problem. We alleviated this problem by filling the gaps in the tree with 18 genomes of novel phylogenetic lineages that had not been sequenced before. In particular, we sequenced five *Rickettsiales* and four *Rhodospirillales*, two orders that had shown close affiliations with mitochondria previously. We showed that with the broad taxonomic sampling we were able to reduce the systematic errors, evident by the less prominent incompatible splits observed in the spectral analysis after adding the novel lineages.

One big hurdle in mitochondrial phylogenetic analysis is the extreme composition biases and high evolutionary rates of the mitochondria-encoded genes. To address this issue, we resorted to well-behaved nuclear genes. We showed that mitochondria-derived nuclear genes have significantly less composition biases and lower rates of evolution than mitochondria-encoded genes. As expected, the tree topologies were sensitive to both the marker datasets and methods used to infer the phylogeny. Because the tree made from the nuclear dataset with the CAT site heterogeneous mixture model was congruent with the tree based on the 200 phylum-level marker genes and was most consistent with the gene order patterns, we chose to make the final tree using this setting.

Placing mitochondria firmly within α-proteobacteria depends on a robust α-proteobacterial phylogeny. Overall our final tree using the nuclear dataset is similar to the previously published α-proteobacterial species trees based on either mitochondrial or phylum-level marker genes (Fitzpatrick et al. 2006; Williams et al. 2007; Georgiades et al. 2011; Thrash et al. 2011; Rodriguez-Ezpeleta and Embley 2012; Viklund et al. 2012) in that they all recover the major α-

proteobacterial groups. However, our genome tree does present novel and interesting branching patterns of α-proteobacterial species that are particularly relevant to the placement of mitochondria. We discuss these new patterns first.

The *Holosporaceae* family consists of mostly obligate endosymbionts from acanthamoeba. Traditionally it has been assigned to the *Rickettsiales* order based on the SSU rRNA phylogeny (Garrity et al. 2004). With only one draft genome (*Odyssella thessalonicensis*) sequenced recently, this family was either absent or very poorly represented in all the previous published genome trees (Esser et al. 2004; Wu et al. 2004; Fitzpatrick et al. 2006; Williams et al. 2007; Wu and Eisen 2008; Wu et al. 2009; Georgiades et al. 2011; Thrash et al. 2011; Rodriguez-Ezpeleta and Embley 2012; Viklund et al. 2012; Viklund et al. 2013). In a recent study with *O. thessalonicensis* as the sole representative, *Holosporaceae* was placed outside of the *Rickettsiales* order and close to the *Rhodospirillales* (Georgiades et al. 2011). With a much broader taxonomic representation of this family, we placed *Holosporaceae* as a deep lineage within *Rickettsiales*, which is consistent with the traditional taxonomy (Figure 6). We think the topology of Georgiades' study is most likely an artifact of sequence composition bias in the data because when we used mitochondria-encoded genes or did not apply the CAT mixture model to account for compositional heterogeneity, we observed topologies similar to that of Georgiades' study as well (Figure S1-3, S5). In addition, our topology is supported by the gene order patterns and is congruent with the SSU rRNA tree and the genome tree based on 200 phylum-level marker genes.

While traditionally SAR11 has been placed within the *Rickettsiales* clade (Williams et al. 2007), and as a sister clade to mitochondria (Georgiades et al. 2011; Thrash et al. 2011), recent studies have conclusively shown that this placement is a tree artifact caused by composition bias, as

mitochondria, *Rickettsiales* and SAR11 all have AT rich genomes (Brindefalk et al. 2011; Rodriguez-Ezpeleta and Embley 2012; Viklund et al. 2012). Indeed, when we used methods that did not account for composition bias, we observed the traditional topology (Figure S1, S3, S5). However, when we applied models that accounted for compositional heterogeneity, only HIMB59 was mostly placed within the *Rickettsiales*, while all the other SAR11 members clustered with the free-living bacteria (Figure S2, S4, S6). The paraphyletic nature of the SAR11 group has been well documented previously (Rodriguez-Ezpeleta and Embley 2012; Viklund et al. 2013), but there is still uncertainty about the exact position of HIMB59 (Viklund et al. 2013). In the Viklund study, HIMB59 has been positioned either within the *Rickettsiales* or the *Rhodospirillales* order depending on the marker datasets used. In our analyses, HIMB59 is almost always positioned within the *Rickettsiales* regardless of the markers (mitochondrial, nuclear or phylum-level markers) or the methods used (RAxML or PhyloBayes). The only exception is in the PhyloBayes tree of the mitochondrial dataset, where HIMB59 and other SAR11 species together group with free-living bacteria (Supplementary Figures 1-6). The placement of HIMB59 within *Rickettsiales* is unlikely caused by the composition bias because the other SAR11 members with more biased AT rich genomes have been separated from the *Rickettsiales*. We note however that the branch leading to HIMB59 is not completely resolved from other *Rickettsiales* (Figure 6), indicating that the position of HIMB59 is unstable. Therefore, we consider the position of HIMB59 tentative and sampling of additional taxa close to HIMB59 should help resolve this issue.

Recent phylogenomic studies have supported two alternative topologies regarding the position of mitochondria: 1) grouping with the free-living *Rhodospirillales* order (Esser et al. 2004), 2) grouping with the *Rickettsiales* order (Wu et al. 2004; Fitzpatrick et al. 2006; Williams et al. 2007; Georgiades et al. 2011; Thrash et al. 2011; Rodriguez-Ezpeleta and Embley 2012).

Resolving this conflict has clear bearing on our understanding of the driving force behind the initial endosymbiosis event. For example, the "hydrogen hypothesis" proposes the metabolic syntrophy between a H_2-producing α-proteobacterial symbiont and a H_2-dependant archaeon as the driving force behind the endosymbiosis (Martin and Muller 1998). The "oxygen scavenger" hypothesis, on the other hand, proposes that the removal of the toxic oxygen by the α-proteobacterium from the anaerobic host has driven the initial symbiosis (Andersson et al. 2003). A key piece of support for the "hydrogen hypothesis" necessitates that the α-proteobacterial ancestor of mitochondria possessed a H_2-producing machinery. Members of the *Rhodospirillales* order are capable of producing H_2 by fermentation while *Rickettsiales* species are not. Grouping mitochondria with *Rhodospirillales* certainly lends stronger support to the "hydrogen hypothesis". With a much broader taxon sampling of both *Rickettsiales* and *Rhodospirillales*, our phylogenomic analyses have almost always placed mitochondria with *Rickettsiales* and never with *Rhodospirillales*, regardless of the marker datasets and phylogenetic methods used (Figures 6, S7-13). Using the same dataset in Esser et al. study but a more sophisticated trimming method to remove fast-evolving sites, Fizpatrick et al. have shown that mitochondria are grouped with *Rickettsiales* and not with *Rhodospirillales* (Fitzpatrick et al. 2006). Taking our and Fizpatrick et al.'s results together, we suspect the topology observed by Esser et al. might be a phylogenetic tree reconstruction artifact.

Our genome tree shows that the *Rickettsiaceae*/*Anaplasmataceae* families are the closest relatives of mitochondria (posterior probability 1.0, Figure 6). This suggests that the ancestor of mitochondria was most likely an endosymbiont that had been already living inside the host cells. For the first time, we are able to place mitochondria firmly within the *Rickettsiales* order. Previous studies have all placed mitochondria as a sister clade to *Rickettsiales* but never unequivocally within *Rickettsiales* (if we discount the sister clade relationship of *Pelagibacter*

and mitochondria). In our genome tree, *Holosporaceae* forms the deepest branch within the *Rickettsiales*. Mitochondria originated sometime after the divergence of *Holosporaceae* from the rest of the *Rickettsiales*. The *Rickettsiales*/mitochondria clade has a very strong posterior probability support value of 0.97. Therefore, we conclude that mitochondria evolved as a derived lineage from within the *Rickettsiales* order.

The multiple novel *Holosporaceae* genomes will be extremely valuable in providing insights into the genetic complement of mitochondrial ancestor. Because they are the immediate outgroup of the mitochondria/*Rickettsiaceae*/*Anaplasmataceae* clade, they have great potentials to improve the accuracy of the mitochondrial ancestral reconstruction. For example, based on the genome sequence of Candidatus *Midichloria mitochondrii*, a novel phylogenetic lineage within *Rickettsiales,* it has been recently predicted that mitochondrial ancestor possessed flagella and could undergo oxidative phosphorylation under both aerobic and microoxic conditions (Sassera et al. 2011).

In conclusion, using an integrated phylogenomic approach, we placed mitochondria firmly within the tree of life and moved a step closer toward pinpointing the origin of mitochondria. Our results suggest that mitochondria most likely originated as an endosymbiont in the *Rickettsiales* lineage*,* but not from the distantly related free-living *Pelagibacter* and *Rhodospirillales*.

Material and Methods

NeighborNet and spectral analyses

The 26 mitochondria-encoded genes (Table 2) from 54 α-proteobacterial genomes and 6 mitochondria representatives were identified, aligned, trimmed using AMPHORA2 (Wu and Scott 2012). NeighborNet analysis was performed using the SplitsTree program (Huson and Bryant 2006) on the concatenated alignment of the 26 mitochondria-encoded proteins with the default parameters. The spectral analysis was performed using the Split Analyses Methods (SAMS) (Wagele and Mayer 2007) with the same dataset after recoding amino acids into 4 categories according to their physicochemical properties. In the spectral analysis, the support for each split was calculated as the number of sites in the alignment supporting that split. The splits were then ranked by their supporting values. To evaluate the systematic errors in the dataset, each of the 50 top-ranked splits was manually evaluated to determine whether it was compatible with well established phylogenetic relationships such as the monophyly of mitochondria or *Rickettsiales*. The systematic errors in the dataset were quantified as the proportion of incompatible splits normalized by their supporting values.

Selection of novel α-proteobacterial species for sequencing

The aligned SSU rRNA gene sequences of 9,817 α-proteobacterial isolates were retrieved from the Ribosomal Database Project (Cole et al. 2009) and were used to construct a maximum likelihood tree using FastTree (Price et al. 2010). A tree-based greedy algorithm was then used to rank isolates by their phylogenetic novelty (Steel 2005), taking into consideration at the same time whether genome sequences of closely related species were available. The availability of an isolate's genomic DNA was also considered in the selection process. In total, 18 isolates were selected for genome sequencing. A SSU rRNA maximum likelihood tree of 70 α-proteobacterial representatives including the 18 targeted species was then made by RAxML (Stamatakis 2006) using the GTR+Gamma model.

Genome sequencing, assembly and annotation

Genomes of the 18 bacterial strains were sequenced by 454 and Illumina sequencing. 7 bacterial strains (*Micavibrio aeruginosavorus, endosymbiont of acanthamoeba UWC8, Candidatus Caedibacter acanthamoebae, Candidatus Paracaedibacter acanthamoebae, Candidatus Paracaedibacter symbiosus, Stella vacuolata, Magnetococcus yuandaducum)* were sequenced by 454 using a combination of indexed shotgun and 3kb paired-end libraries, and assembled using Newbler 2.5.3. The rest 11 strains were sequenced by the Illumina paired-end sequencing using HiSeq 2000, and assembled using the CLCGenomicWorkbench 6.0.1. PCR and Sanger sequencing were used to close the gaps between contigs when necessary. Protein-coding genes of all 18 genomes were predicted using the GLIMMER software package (Delcher et al. 2007). The genome sequence of *M. aeruginosavorus* has already been reported previously (Wang et al. 2011).

Systematic identification of mitochondria-derived nuclear genes

The phylogenetic distribution of all sequenced eukaryotic genomes was retrieved from the GenomeOnline database (Pagani et al. 2012). A total of 30 eukaryotic genomes, representing a broad range of phylogenetic diversity, were selected for identifying the mitochondria-derived nuclear genes (Supplementary Table 2). For every single protein in the 30 eukaryotic nuclear genomes, an initial BLASTP search was performed against a local database containing all complete bacterial, archaeal and mitochondrial genomes. A eukaryotic gene was retained for further analysis if its top 5 hits contained an α-proteobacterial or mitochondrial sequence (e-value cutoff 1e-4). The eukaryotic genes passing the initial BLASTP screening were clustered into protein families using the Markov Cluster Algorithm (Enright et al. 2002) and only families that were present in at least 8 eukaryotic species were selected for phylogenetic analysis. For each of retained protein families, its homologs from all complete bacterial genomes were

retrieved by BLASTP search (e-value cutoff 1e-15). Protein sequences of each family were aligned by MAFFT (Katoh et al. 2002) and trimmed by ZORRO (Wu et al. 2012). Phylogenetic trees constructed using FastTree were subject to manual inspection. Paralogs, if existed in a family, were separated and each was treated as a new family so that only orthologous genes were used for inferring phylogeny. We looked for a specific branching pattern in the trees where eukaryotic sequences clustered with α-proteobacteria and/or mitochondria. Families with less than 8 eukaryotic species, or few α-proteobacterial species, or a complex evolutionary history (e.g., α-, β- and γ-proteobacterial lineages were not clustered together) were removed. In the end, 29 mitochondria-derived nuclear genes were identified as the marker genes for phylogenomic analysis (Table 2).

Assembly of mitochondrial, nuclear and phylum-level marker datasets

For each of 26 mitochondrial and 29 nuclear marker genes, its homologs in 192 α-proteobacterial genomes (Supplementary Table 3) and mitochondrial/eukaryotic representatives were identified, aligned and trimmed using the program AMPHORA2 (Wu and Scott 2012). With very few exceptions, the marker genes were single-copy genes in all of the bacterial, mitochondrial and nuclear genomes analyzed. In those rare cases in which two or more homologs were identified within a single genome, a tree-guided approach was used to resolve the redundancy as described in (Wu and Eisen 2008). If the redundancy was caused by a species-specific duplication event, then one homolog was randomly chosen as the representative. Otherwise, to avoid potential complications in interpreting the phylogeny, we treated the marker as 'missing' in that particular genome. We also identified 200 single-copy marker genes that were present in all the α-proteobacterial genomes using Phyla-AMPHORA (Wang and Wu 2013) and we called them the phylum-level marker dataset. Aligned and trimmed protein sequences within each dataset were concatenated by species and were used as the master datasets for the downstream analyses. The

final mitochondrial, nuclear and phylum-level marker alignments contain 5,790, 6,201 and 54,006 amino acids respectively.

Evaluation of marker datasets and phylogenetic methods

We selected 47 representatives of α-proteobacterial genomes using the tree-based greedy algorithm described above (Steel 2005) and used this set of taxa as a benchmark to evaluate the different datasets (mitochondrial, nuclear and phylum-level markers) and tree construction methods (RAxML and PhyloBayes). We limited this analysis to 47 α-proteobacterial genomes to reduce the computational cost associated with reconstructing the PhyloBayes tree from the phylum-level marker alignment, which contained 54,006 amino acids. For each concatenated dataset, a maximum likelihood (ML) tree and a Bayesian tree were made. ML trees were reconstructed using RAxML (Stamatakis 2006) with the best model selected by the program, and was bootstrapped with 100 replicates. Bayesian consensus trees were reconstructed using PhyloBayes (Lartillot and Philippe 2004) with the -CAT -GTR options, as recommended in the manual. Two independent MCMC chains were run and the chains were considered converged when the maxdiff dropped below 0.3, as suggested in the manual. The trees were sampled every 10 cycles and the beginning one fifth of the trees from each chain were discarded as burn-in.

Estimation of the composition biases and evolutionary rates of the mitochondrial and nuclear marker genes

To estimate the composition biases and evolutionary rates of the mitochondrial and nuclear marker genes, we selected a larger set of 72 α-proteobacteria representatives (including 18 genomes sequenced in this study). For the mitochondrial marker dataset, we added 6 mitochondrial representatives (*Reclinomonas americana, Marchantia polymorpha, Hemiselmis andersenii, Mesostigma viride, Rhodomonas salina* and *Phytophthora infestans*). For the nuclear

marker dataset, we added 6 eukaryotic representatives (*Cryptococcus neoformans, Arabidopsis thaliana, Nematostella vectensis, Spizellomyces punctatus, Monosiga brevicollis, Phytophthora infestans*). The composition bias of each taxon was calculated as a chi-square score using a scheme described in (Viklund et al. 2012). To better account for the missing data in the alignment, we modified the scheme and used the normalized frequency of each amino acid instead of the absolute count. RAxML and PhyloBayes trees were reconstructed using the mitochondrial and nuclear marker alignments. The overall mitochondria/eukaryotes evolutionary rate was estimated as the average branch length from the root of the tree to all the mitochondrial/eukaryotic lineages.

Reconstruction of final genome tree

For the final genome tree reconstruction, we used the nuclear dataset of 72 α-proteobacteria representatives, 6 eukaryotic representatives and 8 outgroups (*Nitrosomonas sp. Is79A3, Ralstonia solanacearum GMI1000, Dechloromonas aromatica RCB, Chromobacterium violaceum ATCC 12472, Francisella tularensis subsp. tularensis FSC198, Legionella pneumophila str. Lens, Escherichia coli str. K-12 substr. MG1655* and *Pseudomonas aeruginosa PA7*). A Bayesian consensus tree was made using PhyloBayes as described above.

As a comparison, we also reconstructed both RAxML and PhyloBayes trees from mitochondrial and nuclear markers with and without amino acid recoding. For the Bayesian analysis, amino acids were recoded to 6 Dayhoff categories. Bayesian consensus trees were made using PhyloBayes as described above plus the '–recode dayhoff6' option. For the RAxML analysis, amino acids were recoded to 4 Dayhoff categories. ML trees were made using the GTR+Gamma model.

Acknowledgements

We would like to thank Drs. Wei Lin and Yongxin Pan at the Institute of Geology and Geophysics, Chinese Academy of Sciences for providing the genomic DNA of *Candidatus Magnetococcus yuandaducum*, and Dr. John Dustin Loy at University of Nebraska for providing the shrimp tissues infected with the NHP bacterium.

References

Adams KL, Song K, Roessler PG, Nugent JM, Doyle JL, Doyle JJ, Palmer JD. 1999. Intracellular gene transfer in action: Dual transcription and multiple silencings of nuclear and mitochondrial cox2 genes in legumes. *Proc Natl Acad Sci U S A* 96:13863-13868.

Anderson S, Bankier AT, Barrell BG, et al. 1981. Sequence and organization of the human mitochondrial genome. *Nature* 290:457-465.

Andersson SG, Karlberg O, Canback B, Kurland CG. 2003. On the origin of mitochondria: A genomics perspective. *Philos Trans R Soc Lond B Biol Sci* 358:165-177; discussion 177-169.

Bandelt HJ, Dress AWM. 1992. A canonical decomposition theory for metrics on a finite set. *Adv Math* 92:47-105.

Brindefalk B, Ettema TJ, Viklund J, Thollesson M, Andersson SG. 2011. A phylometagenomic exploration of oceanic alphaproteobacteria reveals mitochondrial relatives unrelated to the SAR11 clade. *PLoS ONE* 6:e24457.

Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol* 54:743-757.

Burger G, Gray MW, Lang BF. 2003. Mitochondrial genomes: Anything goes. *Trends Genet* 19:709-716.

Clements KD, Gray RD, Howard Choat J. 2003. Rapid evolutionary divergences in reef fishes of the family Acanthuridae (perciformes: Teleostei). *Mol Phylogenet Evol* 26:190-201.

Cole JR, Wang Q, Cardenas E, et al. 2009. The ribosomal database project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 37:D141-145.

Covello PS, Gray MW. 1992. Silent mitochondrial and active nuclear genes for subunit 2 of cytochrome c oxidase (cox2) in soybean: Evidence for rna-mediated gene transfer. *EMBO J* 11:3815-3820.

Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with glimmer. *Bioinformatics* 23:673-679.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6:361-375.

Dunning Hotopp JC, Clark ME, Oliveira DC, et al. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 317:1753-1756.

Emelyanov VV. 2003. Mitochondrial connection to the origin of the eukaryotic cell. *Eur J Biochem* 270:1599-1618.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575-1584.

Esser C, Ahmadinejad N, Wiegand C, et al. 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol* 21:1643-1660.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 27:401-410.

Fitzpatrick DA, Creevey CJ, McInerney JO. 2006. Genome phylogenies indicate a meaningful alphaproteobacterial phylogeny and support a grouping of the mitochondria with the *Rickettsiales. Mol Biol Evol* 23:74-85.

Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol* 48:284-290.

Gabaldon T, Huynen MA. 2003. Reconstruction of the proto-mitochondrial metabolism. *Science* 301:609.

Garrity GM, Bell JA, Lilburn TG. 2004. Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology, second edition. New York: Springer-Verlag.

Georgiades K, Madoui MA, Le P, Robert C, Raoult D. 2011. Phylogenomic analysis of *Odyssella thessalonicensis* fortifies the common origin of *Rickettsiales*, *Pelagibacter ubique* and *Reclimonas americana* mitochondrion. *PLoS ONE* 6:e24857.

Gray MW, Burger G, Lang BF. 2001. The origin and early evolution of mitochondria. *Genome Biol* 2:REVIEWS1018.

Gupta RS. 1995. Evolution of the chaperonin families (hsp60, hsp10 and tcp-1) of proteins and the origin of eukaryotic cells. *Mol Microbiol* 15:1-11.

Hasegawa M, Hashimoto T. 1993. Ribosomal RNA trees misleading? *Nature* 361:23.

Hendy MD, Penny D. 1993. Spectral analysis of phylogenetic data. *J Classif* 10:5-24.

Hillis DM, Huelsenbeck JP, Swofford DL. 1994. Hobgoblin of phylogenetics? *Nature* 369:363-364.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254-267.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: The beginning of incongruence? *Trends Genet* 22:225-231.

Karlberg O, Canback B, Kurland CG, Andersson SG. 2000. The dual origin of the yeast mitochondrial proteome. *Yeast* 17:170-187.

Karlin S, Brocchieri L. 2000. Heat shock protein 60 sequence comparisons: Duplications, lateral transfer, and mitochondrial evolution. *Proc Natl Acad Sci U S A* 97:11348-11353.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. Mafft: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res* 30:3059-3066.

Kennedy M, Holland BR, Gray RD, Spencer HG. 2005. Untangling long branches: Identifying conflicting phylogenetic signals using spectral analysis, neighbor-net, and consensus networks. *Syst Biol* 54:620-633.

Kennedy M, Paterson AM, Morales JC, Parsons S, Winnington AP, Spencer HG. 1999. The long and short of it: Branch lengths and the problem of placing the new zealand short-tailed bat, Mystacina. *Mol Phylogenet Evol* 13:405-416.

Kurland CG, Andersson SG. 2000. Origin and evolution of the mitochondrial proteome. *Microbiol Mol Biol Rev* 64:786-820.

Lang BF, Seif E, Gray MW, O'Kelly CJ, Burger G. 1999. A comparative genomics approach to the evolution of eukaryotes and their mitochondria. *J Eukaryot Microbiol* 46:320-326.

Lartillot N, Philippe H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21:1095-1109.

Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, depamphilis CW. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the felsenstein zone. *Mol Biol Evol* 22:1948-1963.

Lento GM, Hickson RE, Chambers GK, Penny D. 1995. Use of spectral analysis to test hypotheses on the origin of pinnipeds. *Mol Biol Evol* 12:28-52.

Lockhart PJ, Cameron SA. 2001. Trees for bees. *Trends Ecol Evol* 16:84-88.

Lockhart PJ, Penny D. 2005. The place of amborella within the radiation of angiosperms. *Trends Plant Sci* 10:201-202.

Mallatt J, Winchell CJ. 2002. Testing the new animal phylogeny: First use of combined large-subunit and small-subunit rrna gene sequences to classify the protostomes. *Mol Biol Evol* 19:289-301.

Martin W, Muller M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature* 392:37-41.

Nugent JM, Palmer JD. 1991. RNA-mediated transfer of the gene coxii from the mitochondrion to the nucleus during flowering plant evolution. *Cell* 66:473-481.

Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The genomes online database (gold) v.4: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 40:D571-579.

Philippe H. 1997. Rodent monophyly: Pitfalls of molecular phylogenies. *J Mol Evol* 45:712-715.

Philippe H, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol* 22:1246-1253.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol* 21:1455-1458.

Price MN, Dehal PS, Arkin AP. 2010. Fasttree 2--approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490.

Rodriguez-Ezpeleta N, Embley TM. 2012. The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. *PLoS ONE* 7:e30520.

Sassera D, Lo N, Epis S, et al. 2011. Phylogenomic evidence for the presence of a flagellum and cbb(3) oxidase in the free-living mitochondrial ancestor. *Mol Biol Evol* 28:3285-3296.

Soltis DE, Albert VA, Savolainen V, et al. 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": A cautionary tale in phylogenetics. *Trends Plant Sci* 9:477-483.

Stamatakis A. 2006. Raxml-vi-hpc: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688-2690.

Steel M. 2005. Phylogenetic diversity and the greedy algorithm. *Syst Biol* 54:527-529.

Stefanovic S, Rice DW, Palmer JD. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? *BMC Evol Biol* 4:35.

Thrash JC, Boyd A, Huggett MJ, Grote J, Carini P, Yoder RJ, Robbertse B, Spatafora JW, Rappé MS, Giovannoni SJ. 2011. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Scientific Reports* 1.

Viale AM, Arakaki AK. 1994. The chaperone connection to the origins of the eukaryotic organelles. *FEBS Lett* 341:146-151.

Viklund J, Ettema TJ, Andersson SG. 2012. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol* 29:599-615.

Viklund J, Martijn J, Ettema TJ, Andersson SG. 2013. Comparative and phylogenomic evidence that the *alphaproteobacterium HIMB59* is not a member of the oceanic sar11 clade. *PLoS ONE* 8:e78858.

Waddell PJ, Cao Y, Hauf J, Hasegawa M. 1999. Using novel phylogenetic methods to evaluate mammalian mtdna, including amino acid-invariant sites-logdet plus site stripping, to detect internal conflicts in the data, with special reference to the positions of hedgehog, armadillo, and elephant. *Syst Biol* 48:31-53.

Wagele JW, Mayer C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. *BMC Evol Biol* 7:147.

Wang Z, Kadouri DE, Wu M. 2011. Genomic insights into an obligate epibiotic bacterial predator: *Micavibrio aeruginosavorus* ARL-13. *BMC Genomics* 12:453.

Wang Z, Wu M. 2013. A phylum-level bacterial phylogenetic marker database. *Mol Biol Evol* 30:1258-1262.

Williams KP, Sobral BW, Dickerman AW. 2007. A robust species tree for the alphaproteobacteria. *J Bacteriol* 189:4578-4586.

Woese CR, Achenbach L, Rouviere P, Mandelco L. 1991. Archaeal phylogeny: Reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. *Syst Appl Microbiol* 14:364-371.

Wu D, Hugenholtz P, Mavromatis K, et al. 2009. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature* 462:1056-1060.

Wu M, Chatterji S, Eisen JA. 2012. Accounting for alignment uncertainty in phylogenomics. *PLoS One* 7:e30288.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* 9:R151.

Wu M, Scott AJ. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* 28:1033-1034.

Wu M, Sun LV, Vamathevan J, et al. 2004. Phylogenomics of the reproductive parasite *Wolbachia pipientis* wmel: A streamlined genome overrun by mobile genetic elements. *PLoS Biol* 2:E69.

Yoon HS, Grant J, Tekle YI, Wu M, Chaon BC, Cole JC, Logsdon JM, Jr., Patterson DJ, Bhattacharya D, Katz LA. 2008. Broadly sampled multigene trees of eukaryotes. *BMC Evol Biol* 8:14.

Figures

Figure 1. A rooted genome tree of 54 α-proteobacteria and 6 mitochondria represented by a NeighborNet graph. Conflicting signals are apparent in the form of large amount of networking at the base of the tree (enlarged at top-left). The tree is rooted using β- and γ-proteobacteria as the outgroup. The number of conflicting splits associated with each major group by spectral analysis is listed at bottom-left. Only 75 top splits were counted here.

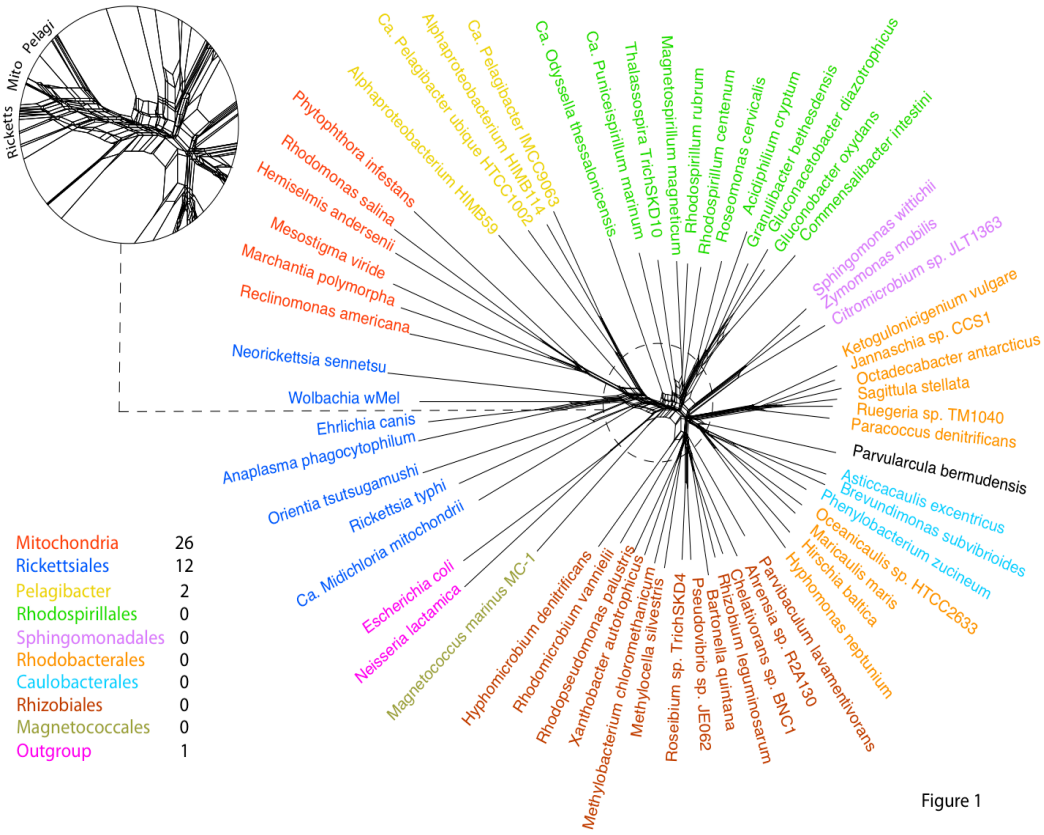


Figure 1

Figure 2. Split spectrum of the concatenated alignment of 26 mitochondria-encoded genes for A) the original dataset, B) the original dataset plus 18 genomes sequenced in this study. Each bar represents a split and the height of bar (Y-axis) is the number of sites in the alignment supporting the split. The splits were ranked by their support and only the top 50 splits are shown. The splits were considered as compatible or incompatible by reconciling with well established phylogenetic relationships such as the monophyly of mitochondria or *Rickettsiales*. Compatible splits are in white and incompatible splits are in black. Asterisks indicate conflicting splits where a single mitochondrial species is placed within the *Rickettsiales* order.

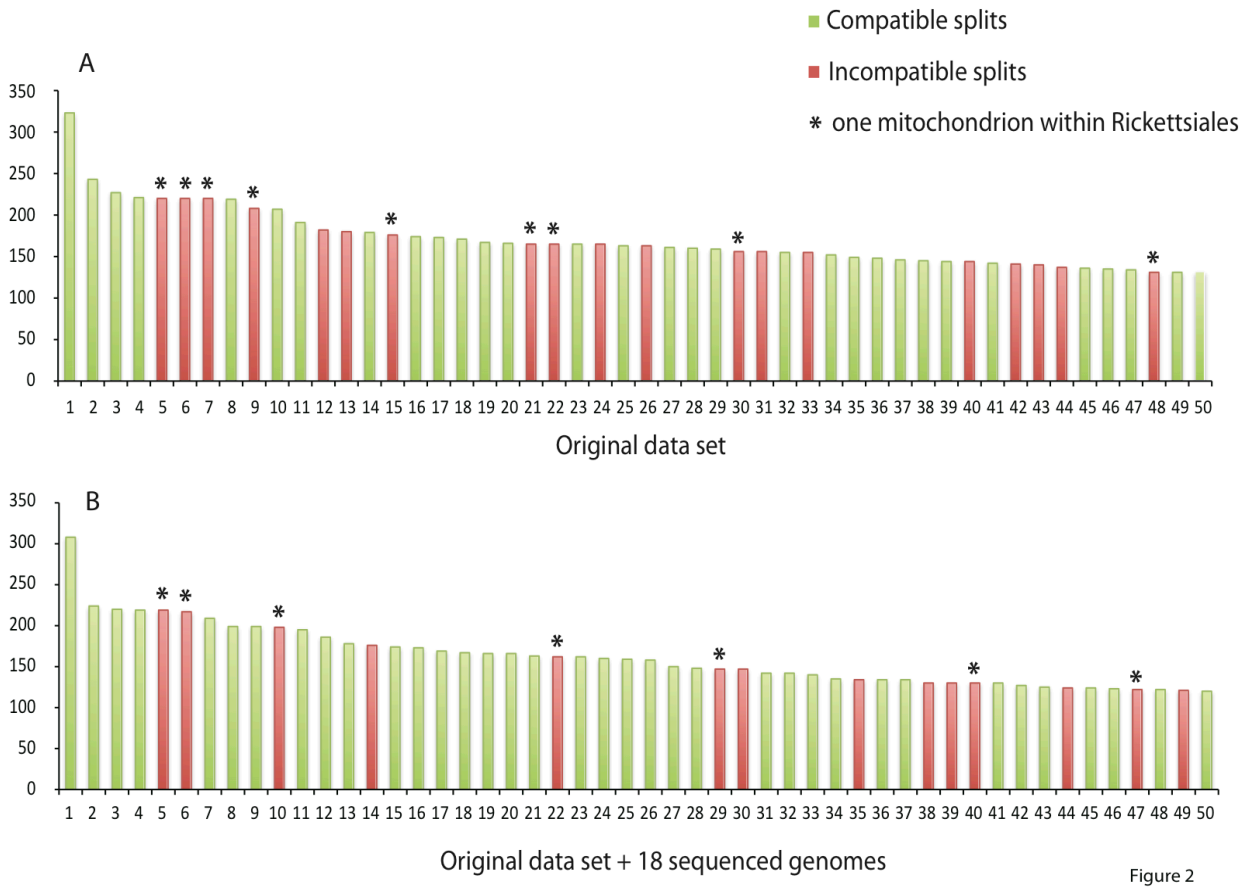


Figure 2

Figure 3. A rooted SSU rRNA maximum likelihood tree of α-proteobacterial representatives using RAxML. Highlighted with asterisks are the 18 isolates selected for sequencing in this study. The tree was rooted using β- and γ-proteobacteria as the outgroup. Bootstrap values (out of 100 replicates) are shown.

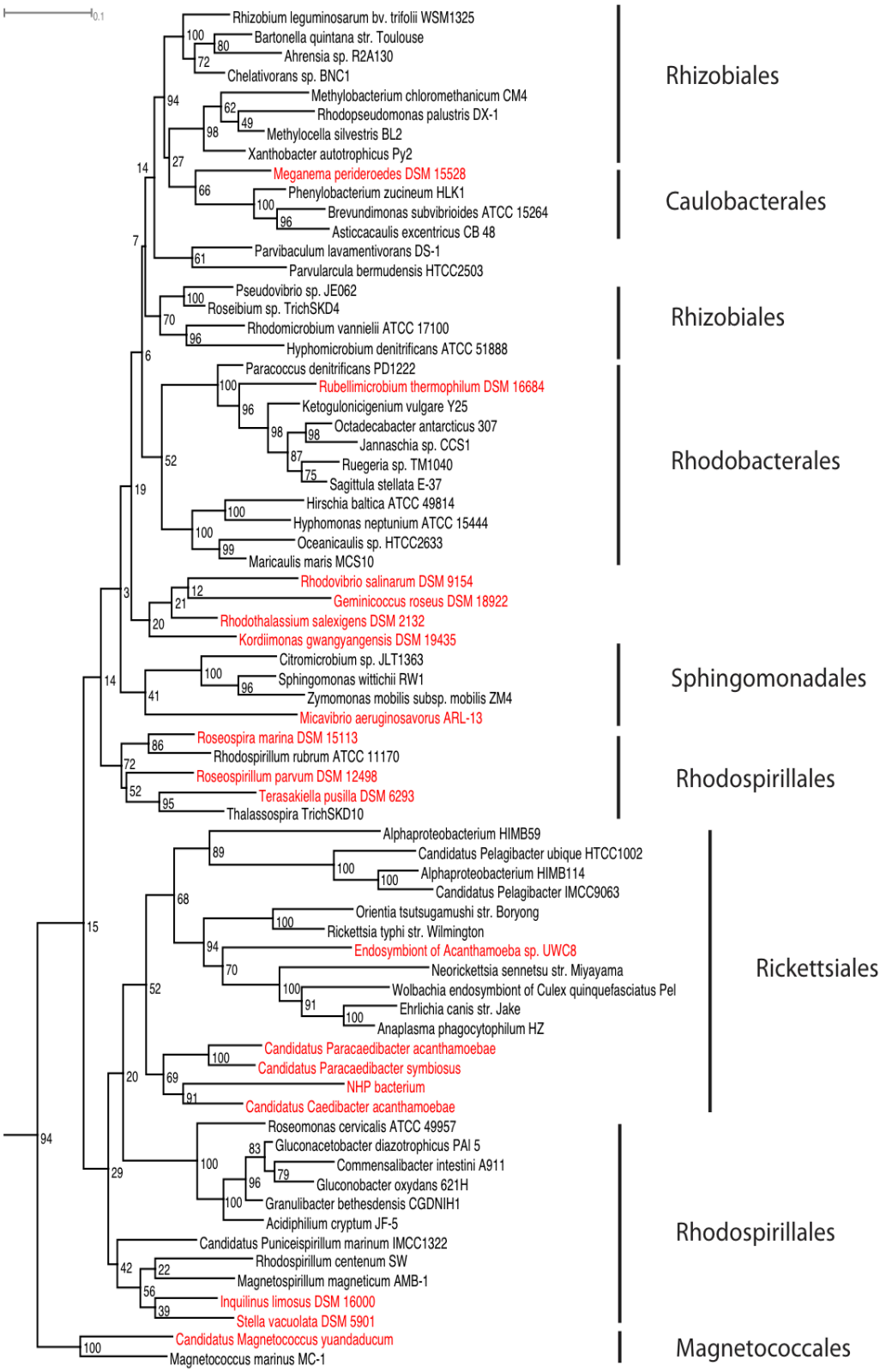

0.1
Rhizobium leguminosarum bv. trifolii WSM1325
Bartonella quintana str. Toulouse
Ahrensia sp. R2A130
Chelativorans sp. BNC1
Methylobacterium chloromethanicum CM4
Rhodopseudomonas palustris DX-1
Methylocella silvestris BL2
Xanthobacter autotrophicus Py2
Meganema perideroedes DSM 15528
Phenylobacterium zucineum HLK1
Brevundimonas subvibrioides ATCC 15264
Asticcacaulis excentricus CB 48
Parvibaculum lavamentivorans DS-1
Parvularcula bermudensis HTCC2503
Pseudovibrio sp. JE062
Roseibium sp. TrichSKD4
Rhodomicrobium vannielii ATCC 17100
Hyphomicrobium denitrificans ATCC 51888
Paracoccus denitrificans PD1222
Rubellimicrobium thermophilum DSM 16684
Ketogulonicigenium vulgare Y25
Octadecabacter antarcticus 307
Jannaschia sp. CCS1
Ruegeria sp. TM1040
Sagittula stellata E-37
Hirschia baltica ATCC 49814
Hyphomonas neptunium ATCC 15444
Oceanicaulis sp. HTCC2633
Maricaulis maris MCS10
Rhodovibrio salinarum DSM 9154
Geminicoccus roseus DSM 18922
Rhodothalassium salexigens DSM 2132
Kordiimonas gwangyangensis DSM 19435
Citromicrobium sp. JLT1363
Sphingomonas wittichii RW1
Zymomonas mobilis subsp. mobilis ZM4
Micavibrio aeruginosavorus ARL-13
Roseospira marina DSM 15113
Rhodospirillum rubrum ATCC 11170
Roseospirillum parvum DSM 12498
Terasakiella pusilla DSM 6293
Thalassospira TrichSKD10
Alphaproteobacterium HIMB59
Candidatus Pelagibacter ubique HTCC1002
Alphaproteobacterium HIMB114
Candidatus Pelagibacter IMCC9063
Orientia tsutsugamushi str. Boryong
Rickettsia typhi str. Wilmington
Endosymbiont of Acanthamoeba sp. UWC8
Neorickettsia sennetsu str. Miyayama
Wolbachia endosymbiont of Culex quinquefasciatus Pel
Ehrlichia canis str. Jake
Anaplasma phagocytophilum HZ
Candidatus Paracaedibacter acanthamoebae
Candidatus Paracaedibacter symbiosus
NHP bacterium
Candidatus Caedibacter acanthamoebae
Roseomonas cervicalis ATCC 49957
Gluconacetobacter diazotrophicus PAl 5
Commensalibacter intestini A911
Gluconobacter oxydans 621H
Granulibacter bethesdensis CGDNIH1
Acidiphilium cryptum JF-5
Candidatus Puniceispirillum marinum IMCC1322
Rhodospirillum centenum SW
Magnetospirillum magneticum AMB-1
Inquilinus limosus DSM 16000
Stella vacuolata DSM 5901
Candidatus Magnetococcus yuandaducum
Magnetococcus marinus MC-1
Rhizobiales
Caulobacterales
Rhizobiales
Rhodobacterales
Sphingomonadales
Rhodospirillales
Rickettsiales
Rhodospirillales
Magnetococcales

Figure 3

Figure 4. Schematic phylogenetic trees based on the mitochondrial, nuclear and phylum-level marker datasets and reconstructed using RAxML and PhyloBayes. Bootstrap values (for RAxML trees) and posterior probability values (for PhyloBayes trees) for internal nodes are shown beside them.

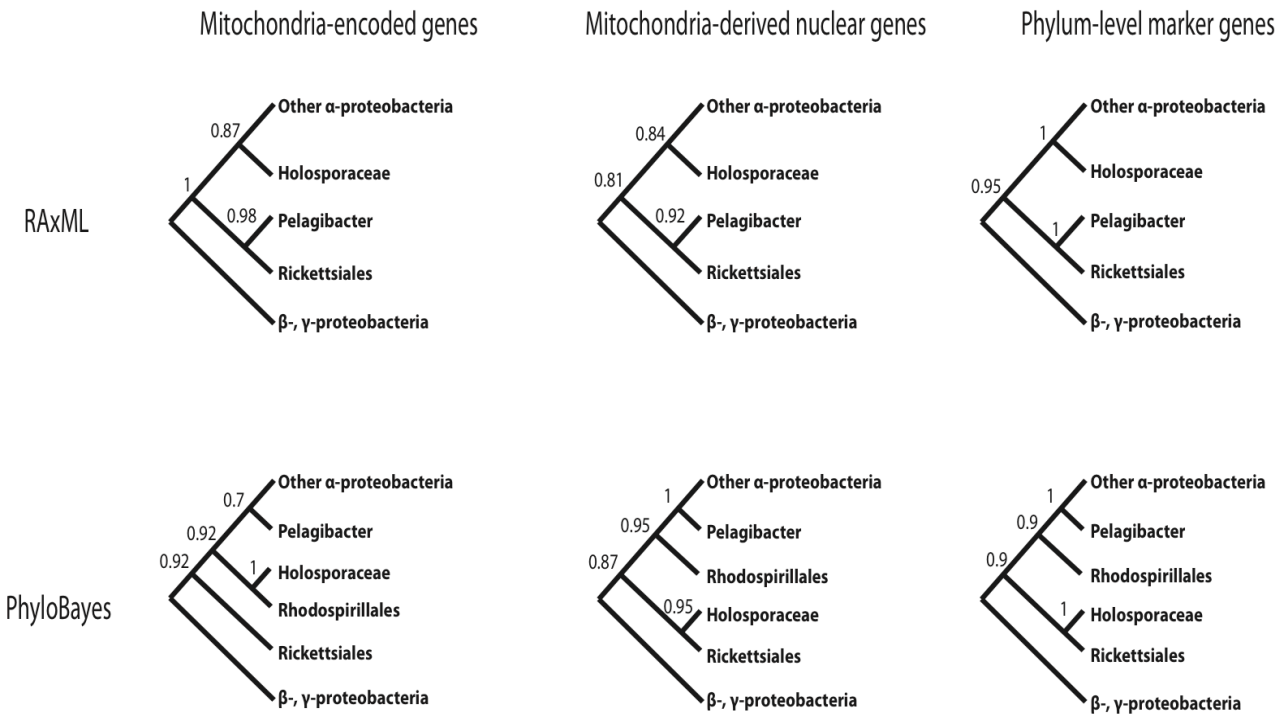


Figure 4

Figure 5. The gene orders of a gene cluster of 12 protein-coding genes in *Rickettsiales* (red), *Holosporaceae* (green), the SAR11 group (purple) and the free-living *Rhodospirillum rubrum* (black). Each arrow represents a gene in the cluster. Arrows with dotted lines represent a missing gene. Genome rearrangements are shown as dotted lines between two genes, with the distance between them shown above the lines. Because of the incomplete nature of some genome assemblies, the exact distance between two genes could not be determined. In this case, a minimum distance was estimated as the sum of distances of each gene to the end of the contig it was located on. For the same reason, the orientation of some genes could not be determined (indicated by asterisks below the genes).

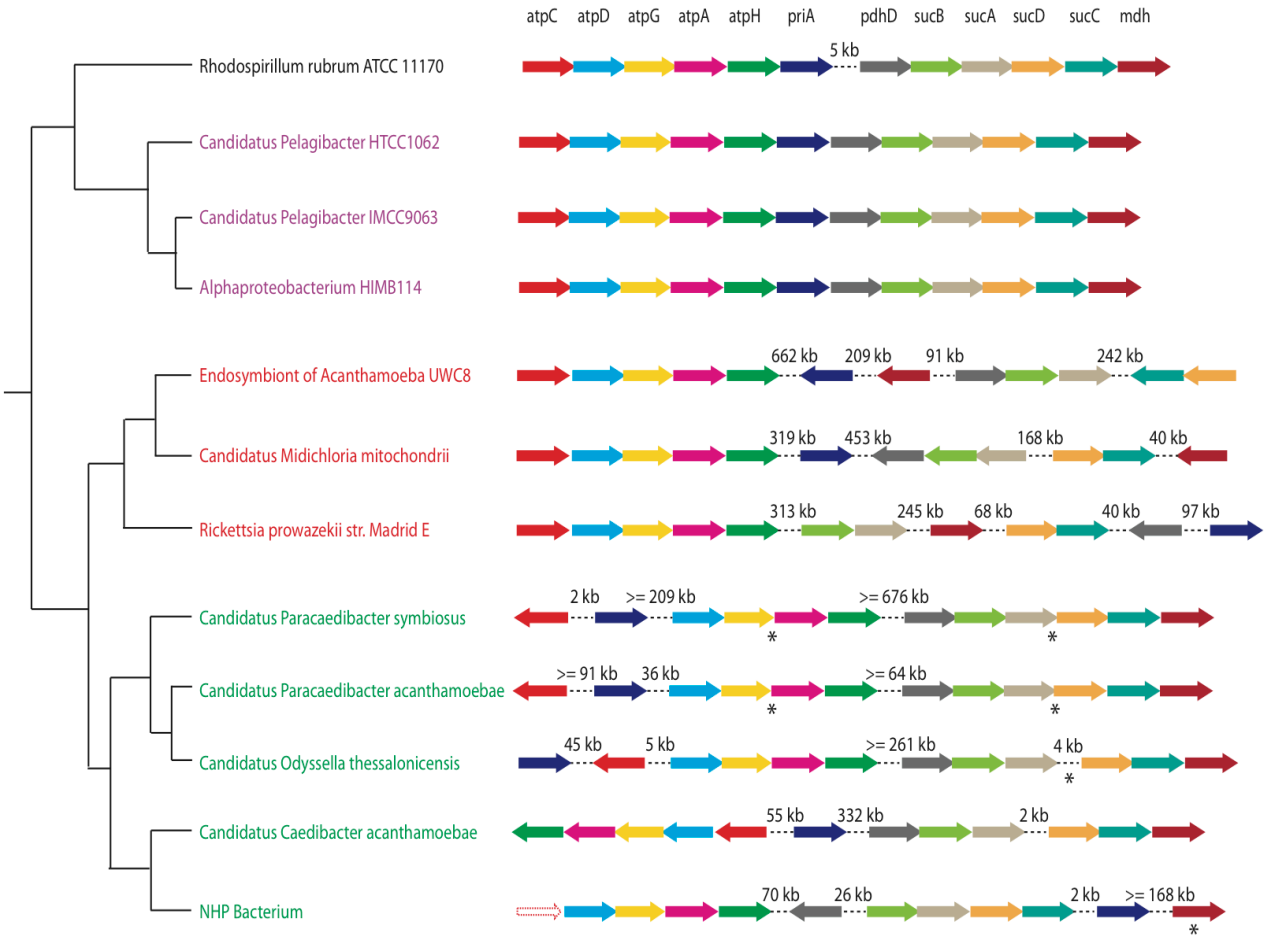


Figure 5

Figure 6. A rooted Bayesian consensus tree made with the nuclear dataset of 72 α-proteobacteria and 6 eukaryotes. Asterisks indicate the 18 genomes sequenced in this study. The tree was rooted using β- and γ-proteobacteria as the outgroup. The posterior probability support values of the internal nodes are 1.0 unless as indicated in the tree.


Sphingomonadales
Citromicrobium sp. JLT1363
Sphingomonas wittichii RW1
Zymomonas mobilis subsp. mobilis ZM4
Rhodothalassium salexigens DSM 2132 *
Kordiimonas gwangyangensis DSM 19435 *
SAR11
Candidatus Pelagibacter ubique HTCC1002
Candidatus Pelagibacter IMCC9063
Alphaproteobacterium HIMB114
Rhodobacterales
Meganema perideroedes DSM 15528 *
Ketogulonicigenium vulgare Y25
0.99
Paracoccus denitrificans PD1222
Sagittula stellata E-37
Ruegeria sp. TM1040
Octadecabacter antarcticus 307
Jannaschia sp. CCS1
Rubellimicrobium thermophilum DSM 16684 *
Maricaulis maris MCS10
Oceanicaulis sp. HTCC2633
Hirschia baltica ATCC 49814
Hyphomonas neptunium ATCC 15444
Caulobacterales
Phenylobacterium zucineum HLK1
Asticcacaulis excentricus CB 48
Brevundimonas subvibrioides ATCC 15264
Parvularcula bermudensis HTCC2503
Rhizobiales
Rhizobium leguminosarum bv. trifolii WSM1325
Chelativorans sp. BNC1
Bartonella quintana str. Toulouse
Ahrensia sp. R2A130
Roseibium sp. TrichSKD4
Pseudovibrio sp. JE062
Methylocella silvestris BL2
0.98
Methylobacterium chloromethanicum CM4
Rhodopseudomonas palustris DX-1
Xanthobacter autotrophicus Py2
Hyphomicrobium denitrificans ATCC 51888
Rhodomicrobium vannielii ATCC 17100
Parvibaculum lavamentivorans DS-1
Candidatus Puniceispirillum marinum IMCC1322
Rhodospirillales
Granulibacter bethesdensis CGDNIH1
0.87
Acidiphilium cryptum JF-5
Gluconobacter oxydans 621H
Gluconacetobacter diazotrophicus PAI 5
0.92
Commensalibacter intestini A911
Roseomonas cervicalis ATCC 49957
Magnetospirillum magneticum AMB-1
Roseospirillum parvum DSM 12498 *
Rhodospirillum rubrum ATCC 11170
0.88
Roseospira marina DSM 15113 *
Thalassospira TrichSKD10
Terasakiella pusilla DSM 6293 *
Micavibrio aeruginosavorus ARL-13 *
Rhodospirillum centenum SW
Inquilinus limosus DSM 16000 *
Rhodovibrio salinarum DSM 9154 *
Stella Vacuolata DSM 5901 *
Mitochondria
Cryptococcus neoformans
Spizellomyces punctatus
Nematostella vectensis
0.97
Monosiga brevicollis
Phytophthora infestans
0.84
Arabidopsis thaliana
Rickettsiales
Orientia tsutsugamushi str. Boryong
Rickettsia typhi str. Wilmington
Neorickettsia sennetsu str. Miyayama
Wolbachia endosymbiont of Culex quinquefasciatus Pel
Ehrlichia canis str. Jake
Anaplasma phagocytophilum HZ
0.97
Candidatus Midichloria mitochondrii IricVA
Endosymbiont of Acanthamoeba sp. UWC8 *
Holosporaceae
Candidatus Paracaedibacter symbiosus *
Candidatus Odyssella thessalonicensis L13
Candidatus Paracaedibacter acanthamoebae *
NHP bacterium *
0.99
Candidatus Caedibacter acanthamoebae *
Alphaproteobacterium HIMB59
Geminicoccus roseus DSM 18922 *
Magnetococcales
Candidatus Magnetococcus yuandaducum *
Magnetococcus marinus MC-1
Nitrosomonas sp. Is79A3
Ralstonia solanacearum GMI1000
Dechloromonas aromatica RCB
Chromobacterium violaceum ATCC 12472
Francisella tularensis subsp. tularensis FSC198
Legionella pneumophila str. Lens
Escherichia coli str. K-12 substr. MG1655
Pseudomonas aeruginosa PA7

Figure 6

Tables

Table 1. Overview of the 18 α-proteobacterial genomes sequenced in this study.

Genomes	Order	Draft genome size	No. of contigs	Coverage	GC content (%)	Protein coding genes	Mito markers	Nuclear markers	Phylum markers
Kordiimonas gwangyangensis DSM 19435	*Kordiimonadales*	4149991	272	320x	57.6	3970	25	28	198
Candidatus Magnetococcus yuandaducum	*Magnetococcales*	2228395	649	23x	58.9	2699	23	15	131
Meganema perideroedes DSM 15528	*Rhizobiales*	3464569	324	209x	67.1	3494	24	26	197
Roseospirillum parvum DSM 12498	*Rhizobiales*	3436975	3024	323x	69.6	4127	22	20	187
Terasakiella pusilla DSM 6293	*Rhizobiales*	4067442	259	150x	50.1	4098	24	27	200
Rhodothalassium salexigens DSM 2132	*Rhodobacterales*	3156491	3163	294x	68.0	4058	26	23	193
Rubellimicrobium thermophilum DSM 16684	*Rhodobacterales*	3328337	361	99x	69.2	3381	25	27	197
Inquilinus limosus DSM 16000	*Rhodospirillales*	6772298	4283	83x	69.3	8184	25	24	190
Rhodovibrio salinarum DSM 9154	*Rhodospirillales*	4170570	258	117x	65.9	4040	25	27	199
Roseospira marina DSM 15113	*Rhodospirillales*	3635965	8906	91x	67.0	6978	22	20	175
Stella vacuolata DSM 5901	*Rhodospirillales*	4353044	1038	7x	70.2	4337	20	22	145
Candidatus Caedibacter acanthamoebae	*Rickettsiales*	2175773	5	50x	37.9	2332	26	26	193
Candidatus Paracaedibacter acanthamoebae	*Rickettsiales*	2454690	55	67x	41.0	2535	26	26	197
Candidatus Paracaedibacter symbiosus	*Rickettsiales*	2668935	299	15x	41.2	2967	23	26	195
Endosymbiont of Acanthamoeba sp. UWC8	*Rickettsiales*	1615277	1	20x	34.8	1608	24	26	196
NHP bacterium	*Rickettsiales*	1115609	15	927x	49.8	1309	23	21	171
Geminicoccus roseus DSM 18922	unclassified	5676036	1169	109x	68.4	5909	24	27	191
Micavibrio aeruginosavorus ARL-13	unclassified	2481983	1	60x	54.7	2432	26	27	198

Table 2. Comparison between mitochondria-encoded genes and mitochondria-derived nuclear genes in terms of the evolutionary rate and composition bias.

		Mitochondria-encoded genes	Mitochondria-derived nuclear genes
Functional categories	Energy production and conversion	*cob, cox2, cox3, nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad9*	*cox11, sdhB, sucD, petA, erpA, hesB, ybjS, nuoC, nuoD, nuoF, nuoG, nuoI*
	Translation and posttranslational modification	*rpl2, rpl5, rpl6, rpl16, rps1, rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps13, rps14, rps19*	*rpl13, grpE, groEL, dnaK, clpB, clpP, hslV, engA, gidA, trmE*
	Others		*AFG1, apaG, bioC, hemN, ksgA, mraW, hypothetical*
Mitochondrial/Nuclear average evolutionary rate (substitution/site) *		1.713 (stdev 0.225)	1.273 (stdev 0.088)
Mitochondrial/Nuclear average aminoGC content **		0.152 (stdev 0.017)	0.215 (stdev 0.004)
Mitochondrial/Nuclear average compositional chi-square scores *		662.4 (stdev 394.3)	89.6 (stdev 41.4)

* T-test $P < 0.01$ ** T-test $P < 0.001$

Supplementary Figures

Figure S1. A rooted RAxML ML tree made with the mitochondrial marker dataset of 47 α-proteobacteria representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Bootstrap values for internal nodes are shown beside them.

Figure S1

Figure S2. A rooted Bayesian tree made with the mitochondrial marker dataset of 47 α-proteobacteria representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Posterior probability values for internal nodes are shown beside them.

Figure S2

Figure S3. A rooted RAxML ML tree made with the nuclear marker dataset of 47 α-proteobacteria representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Bootstrap values for internal nodes are shown beside them.

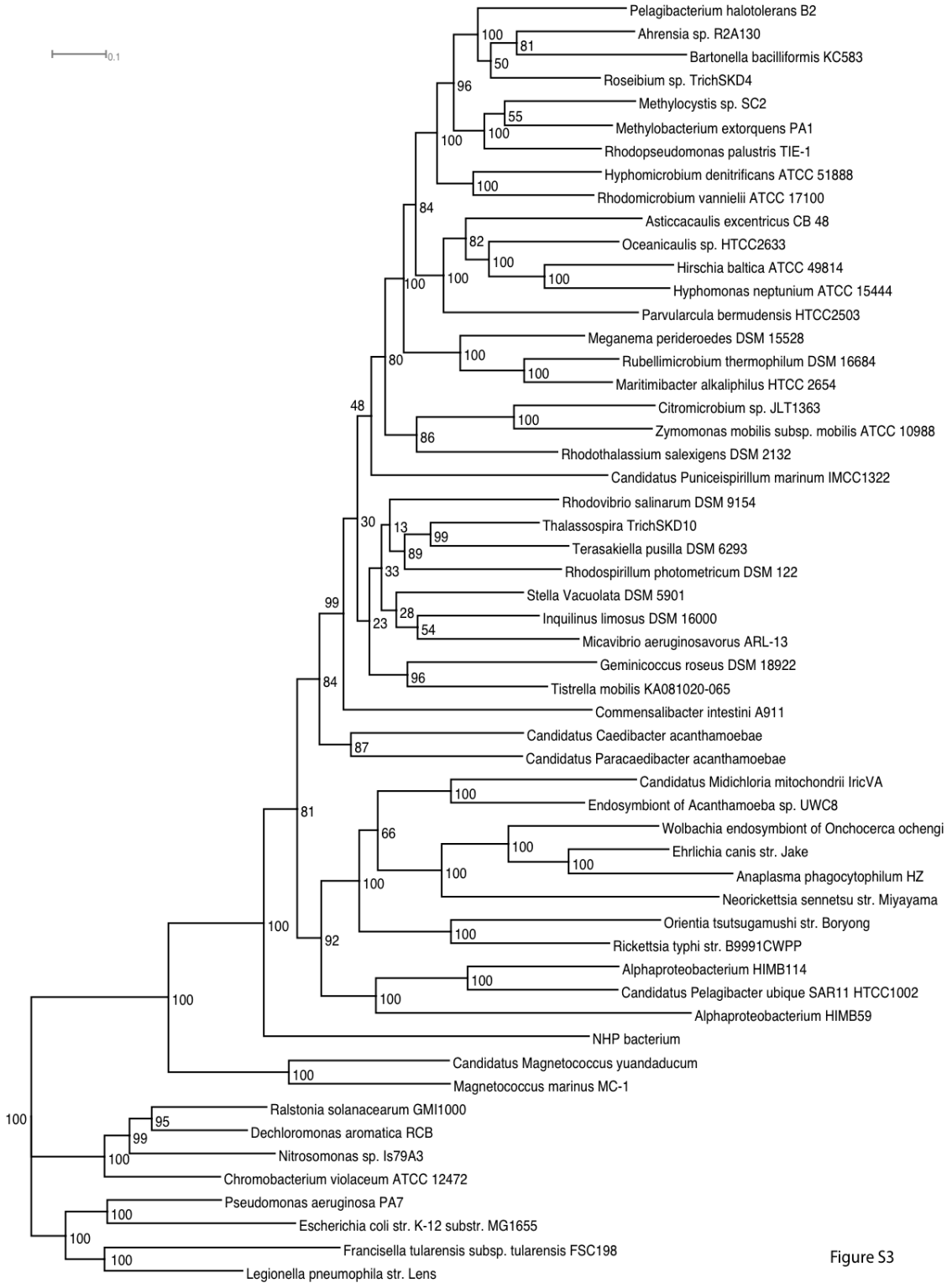


Figure S3

Figure S4. A rooted Bayesian tree made with the nuclear marker dataset of 47 α-proteobacteria representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Posterior probability values for internal nodes are shown beside them.

Figure S4

Figure S5. A rooted RAxML ML tree made with the phylum-level marker dataset of 47 α-proteobacteria representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Bootstrap values for internal nodes are shown beside them.

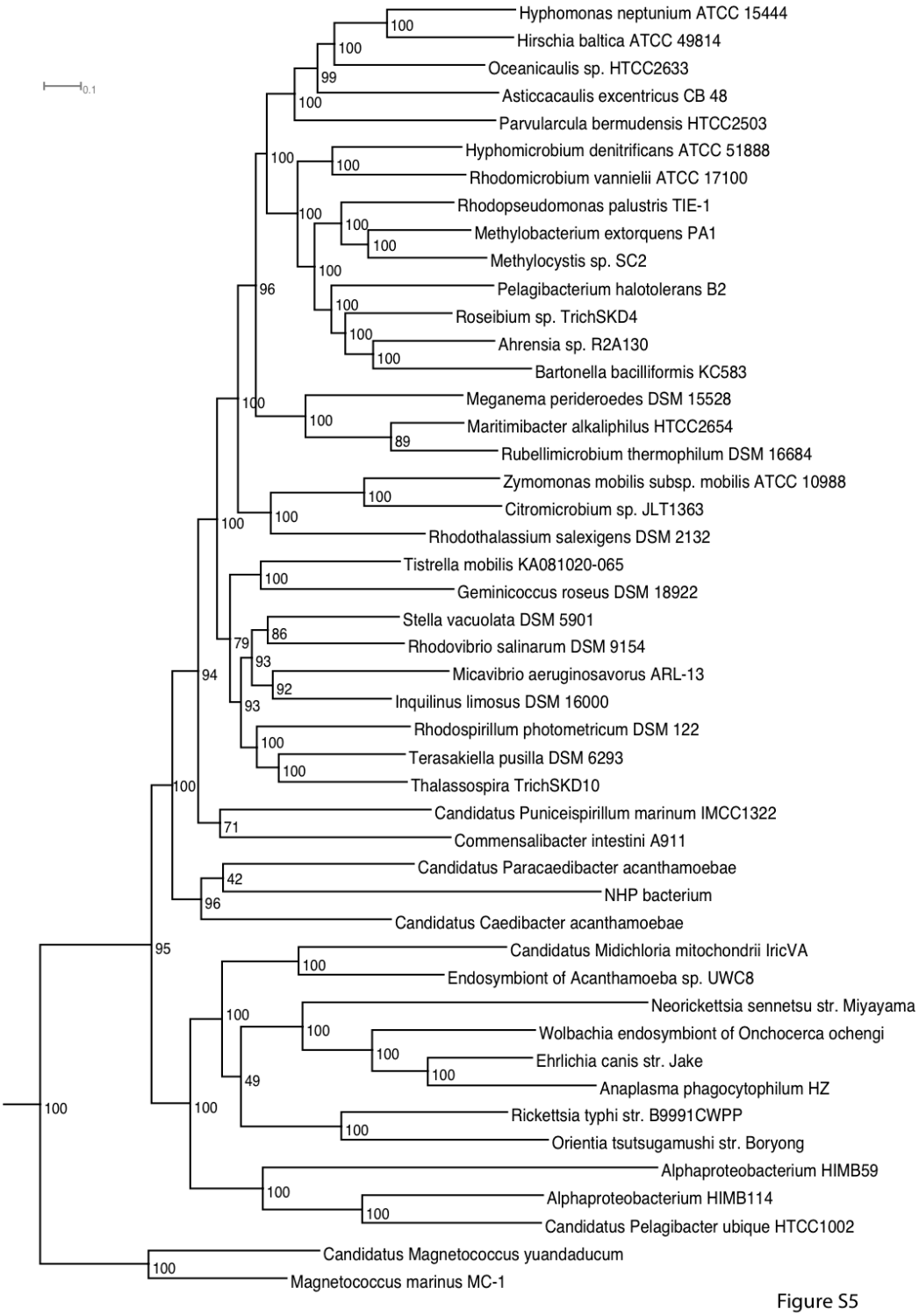


Figure S5

Figure S6. A rooted Bayesian tree made with the phylum-level marker dataset of 47 α-proteobacteria representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Posterior probability values for internal nodes are shown beside them.

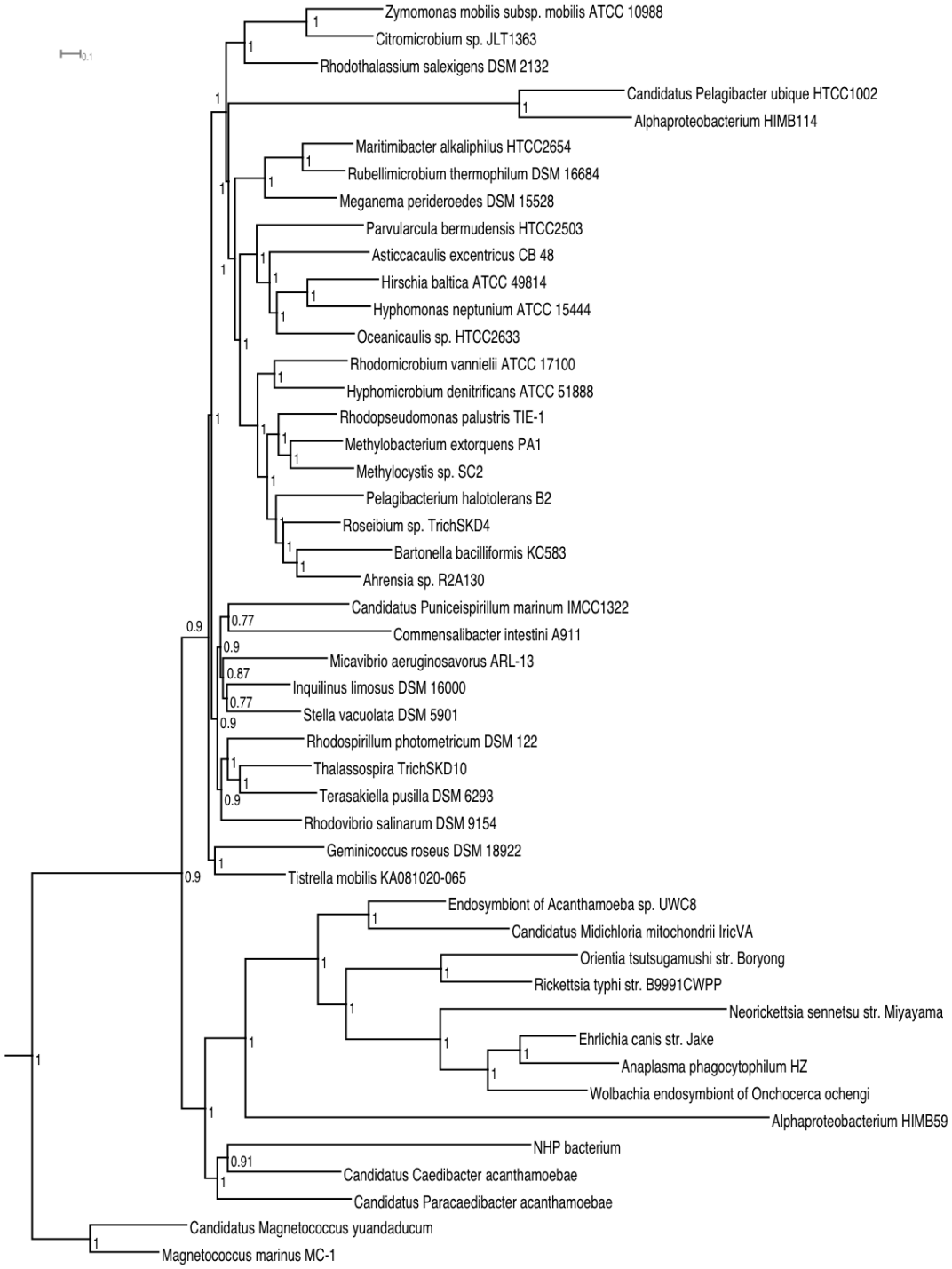


Figure S6

Figure S7. A rooted RAxML ML tree made with the mitochondrial marker dataset of 72 α-proteobacteria and 6 mitochondria representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Bootstrap values for internal nodes are shown beside them.

Figure S7

Figure S8. A rooted Bayesian tree made with the mitochondrial marker dataset of 72 α-proteobacteria and 6 mitochondria representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Posterior probability values for internal nodes are shown beside them.

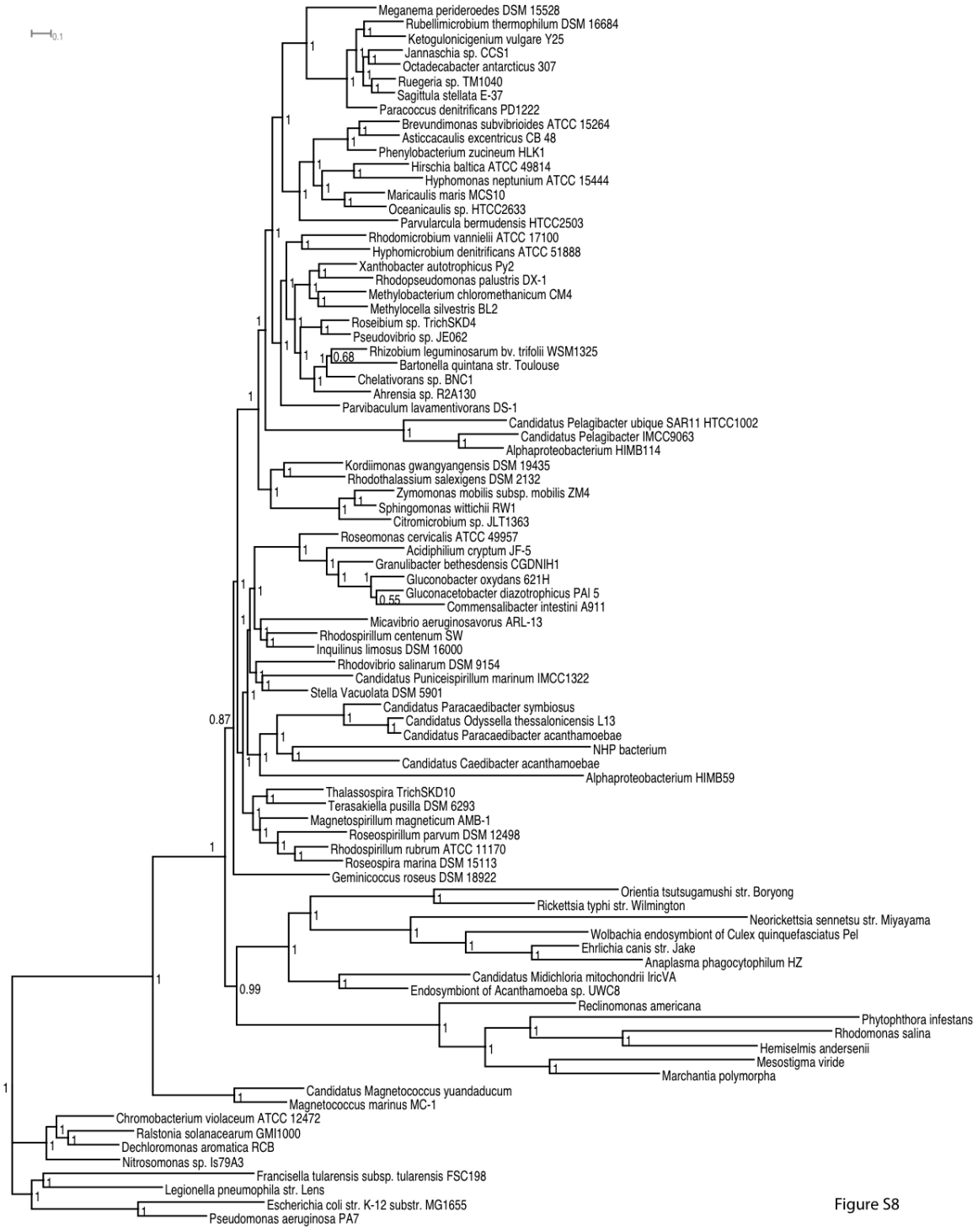


Figure S8

Figure S9. A rooted RAxML ML tree made with the nuclear marker dataset of 72 α-proteobacteria and 6 eukaryote representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Bootstrap values for internal nodes are shown beside them.

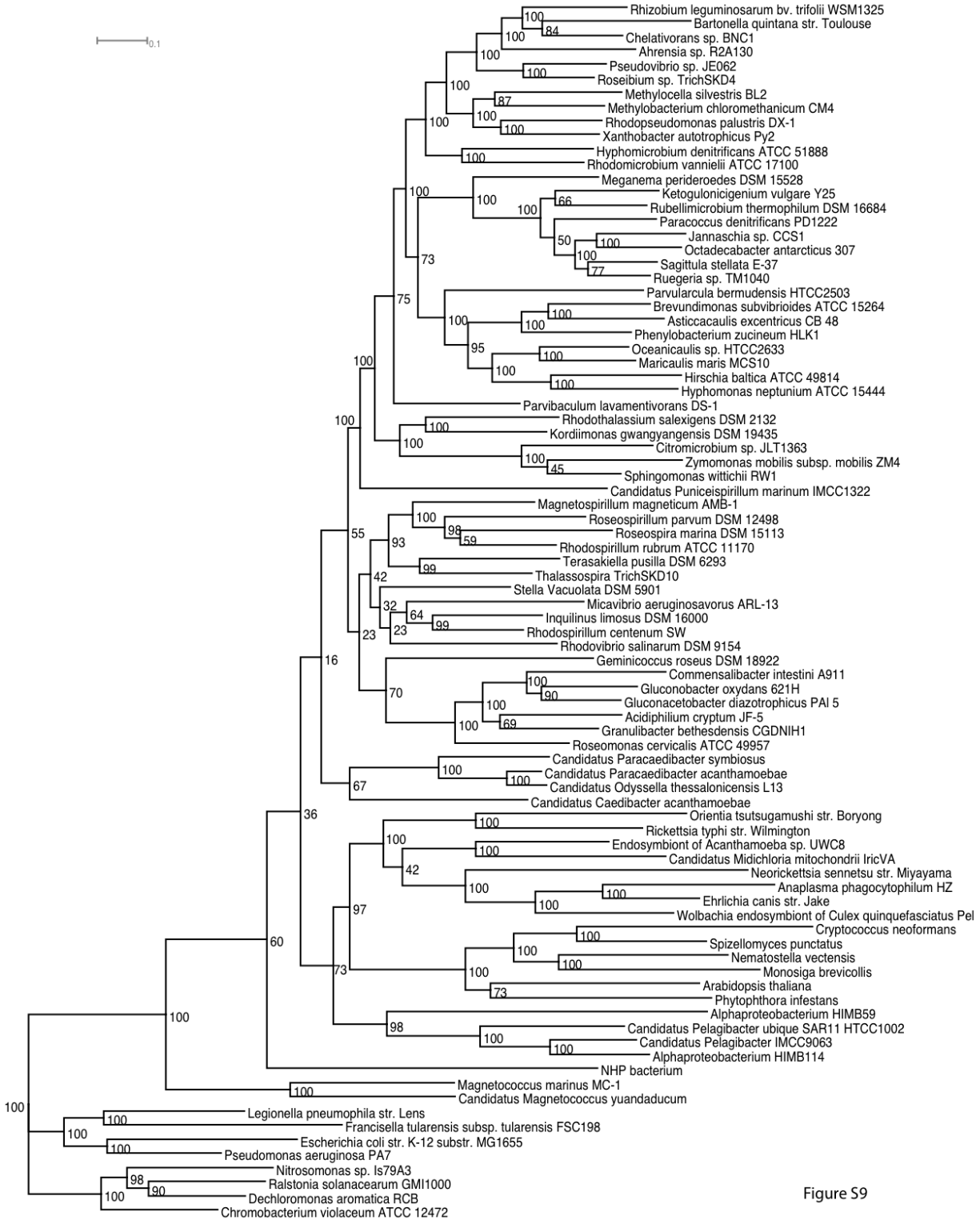


Figure S9

Figure S10. A rooted Bayesian tree made with the Dayhoff6 recoded mitochondria marker dataset of 72 α-proteobacteria and 6 mitochondria representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Posterior probability values for internal nodes are shown beside them.

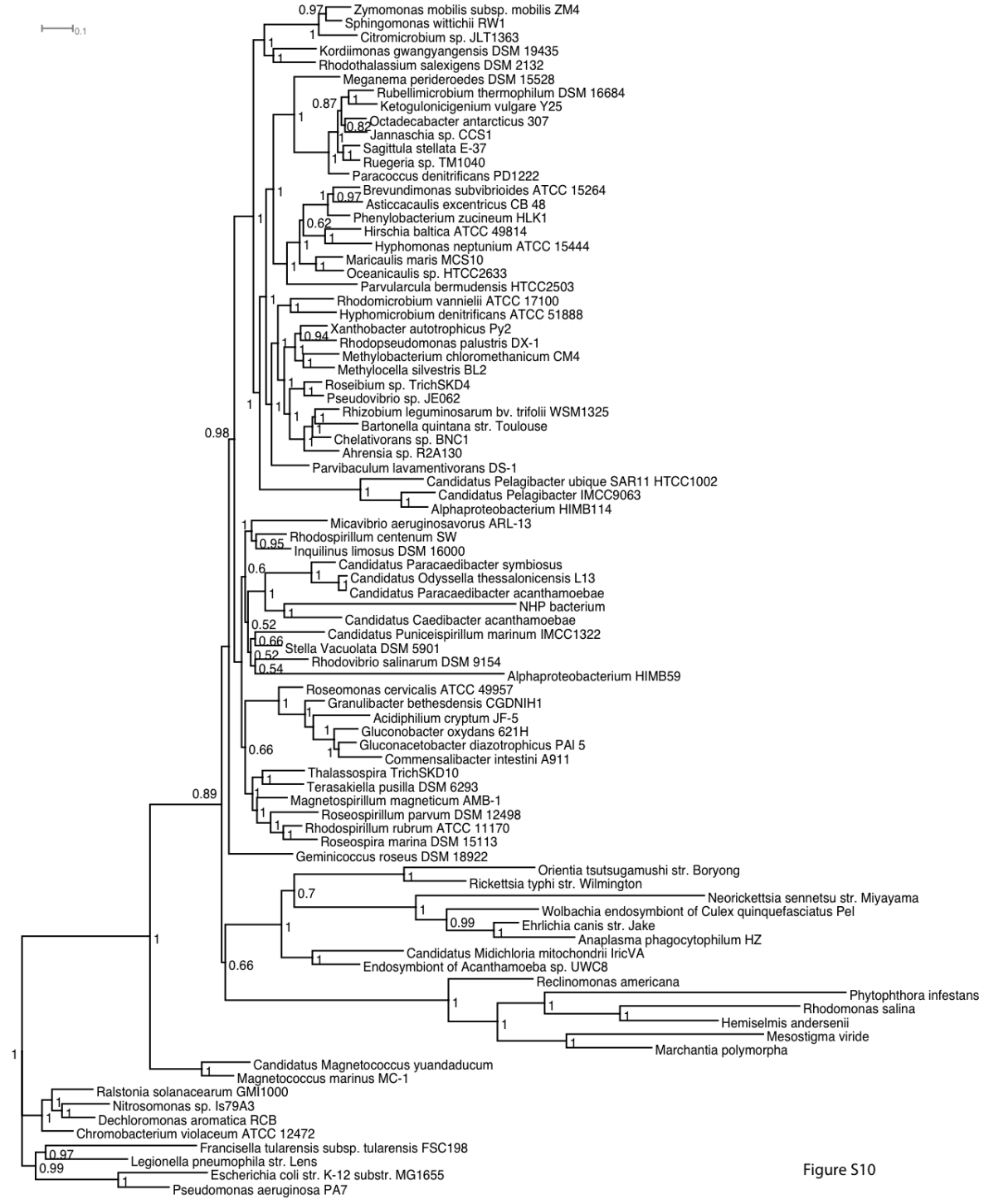


Figure S10

Figure S11. A rooted Bayesian tree made with the Dayhoff6 recoded nuclear marker dataset of 72 α-proteobacteria and 6 eukaryote representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Posterior probability values for internal nodes are shown beside them.

Figure S11

Figure S12. A rooted RAxML ML tree made with the Dayhoff4 recoded mitochondria marker dataset of 72 α-proteobacteria and 6 mitochondria representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Bootstrap values for internal nodes are shown beside them.

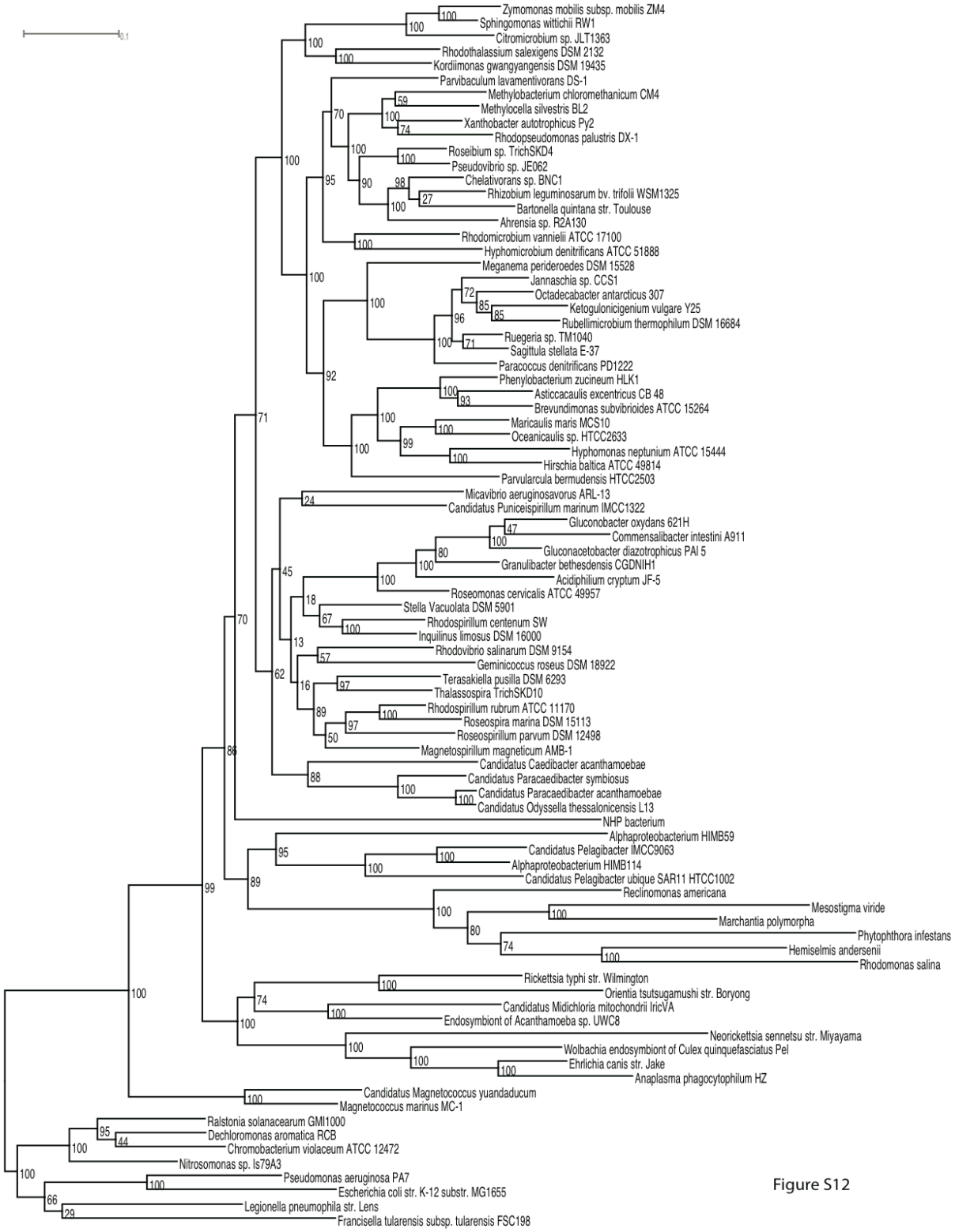


Figure S12

Figure S13. A rooted RAxML ML tree made with the Dayhoff4 recoded nuclear marker dataset of 72 α-proteobacteria and 6 eukaryote representatives. The tree was rooted using β- and γ-proteobacteria as the outgroup. Bootstrap values for internal nodes are shown beside them.

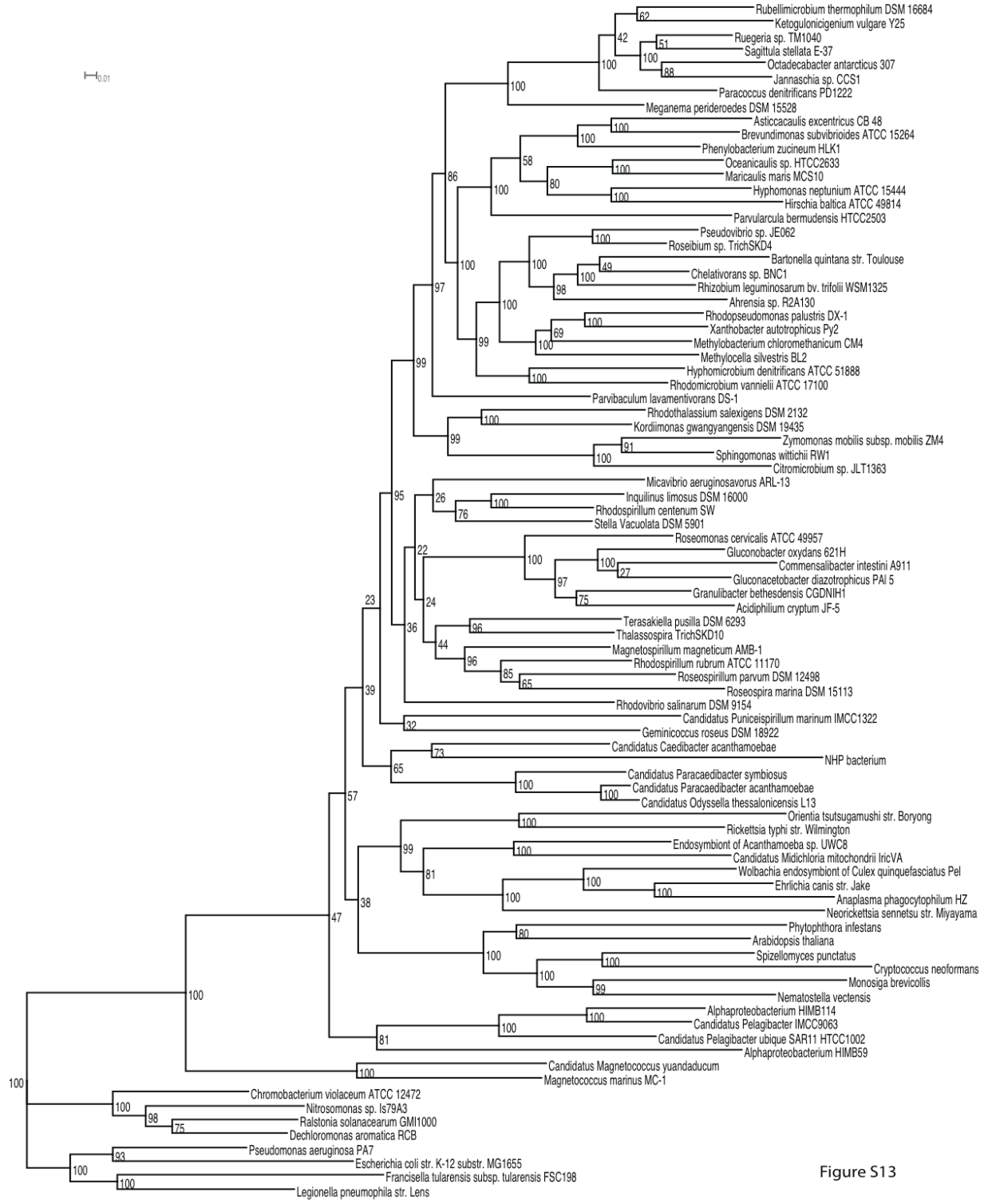


Figure S13

Supplementary Tables

Supplementary Table 1. List of 50 top-ranked splits in the spectral analysis of the original dataset. Long-branch lineages that violate the well established phylogenetic relationships (e.g., the monophyly of mitochondria or Rickettsiales) are highlighted in bold.

Rank	Split	Support	Compatibility
1	(Wolbachia_endosymbiont_of_Culex_quinquefasciatus_Pel,Neorickettsia_sennetsu_str._Miyayama,Anaplasma_phagocytophilum_HZ,Ehrlichia_canis_str._Jake), (others)	323	Compatible
2	(Hemiselmis_andersenii,Rhodomonas_salina), (others)	243	Compatible
3	(Mesostigma_viride,Marchantia_polymorpha), (others)	227	Compatible
4	(Ruegeria_sp._TM1040,Octadecabacter,Sagittula,Paracoccus_denitrificans_PD1222,Ketogulonicigenium_vulgare_Y25,Jannaschia_sp._CCS1), (others)	221	Compatible
5	(**Hemiselmis_andersenii**,Candidatus_pelagibacter_IMCC9063,Alphaproteobacterium_HIMB114,Candidatus_Pelagibacter_ubique_SAR11_HTCC1002), (others)	220	Incompatible
6	(**Phytophthora_infestans**,Candidatus_pelagibacter_IMCC9063,Alphaproteobacterium_HIMB114,Candidatus_Pelagibacter_ubique_SAR11_HTCC1002), (others)	220	Incompatible
7	(Wolbachia_endosymbiont_of_Culex_quinquefasciatus_Pel,Neorickettsia_sennetsu_str._Miyayama,Anaplasma_phagocytophilum_HZ,Ehrlichia_canis_str._Jake,**Rhodomonas_salina**), (others)	220	Incompatible
8	(Orientia_tsutsugamushi_str._Boryong,Rickettsia_typhi_str._Wilmington), (others)	219	Compatible
9	(Candidatus_pelagibacter_IMCC9063,Alphaproteobacterium_HIMB114,Candidatus_Pelagibacter_ubique_SAR11_HTCC1002,**Rhodomonas_salina**), (others)	208	Incompatible
10	(Candidatus_pelagibacter_IMCC9063,Alphaproteobacterium_HI	207	Compatible

	MB114), (others)		
11	(Escherichia_coli_str._K-12_substr._MG1655,Neisseria_lactamica_020-06), (others)	191	Compatible
12	(Citromicrobium_sp._JLT1363,Zymomonas_mobilis_subsp._mobilis_ZM4,Sphingomonas_wittichii_RW1,**Rhodomonas_salina**), (others)	182	Incompatible
13	(Citromicrobium_sp._JLT1363,Zymomonas_mobilis_subsp._mobilis_ZM4,**Phytophthora_infestans**,Sphingomonas_wittichii_RW1), (others)	180	Incompatible
14	(Hemiselmis_andersenii,Marchantia_polymorpha,Reclinomonas_americana,Rhodomonas_salina), (others)	179	Compatible
15	(Candidatus_pelagibacter_IMCC9063,Alphaproteobacterium_HIMB114,**Reclinomonas_americana**,Candidatus_Pelagibacter_ubique_SAR11_HTCC1002), (others)	176	Incompatible
16	(Hemiselmis_andersenii,Mesostigma_viride,Marchantia_polymorpha,Rhodomonas_salina), (others)	174	Compatible
17	(Hemiselmis_andersenii,Mesostigma_viride,Phytophthora_infestans,Rhodomonas_salina), (others)	173	Compatible
18	(Mesostigma_viride,Phytophthora_infestans), (others)	171	Compatible
19	(Hemiselmis_andersenii,Mesostigma_viride,Phytophthora_infestans,Marchantia_polymorpha,Reclinomonas_americana,Rhodomonas_salina), (others)	167	Compatible
20	(Hemiselmis_andersenii,Phytophthora_infestans,Marchantia_polymorpha,Rhodomonas_salina), (others)	166	Compatible
21	(**Mesostigma_viride**,Wolbachia_endosymbiont_of_Culex_quinquefasciatus_Pel,Anaplasma_phagocytophilum_HZ,Ehrlichia_canis_str._Jake), (others)	165	Incompatible
22	(Wolbachia_endosymbiont_of_Culex_quinquefasciatus_Pel,**Phytophthora_infestans**,Anaplasma_phagocytophilum_HZ,Ehrlichia	165	Incompatible

	_canis_str._Jake), (others)		
23	(Marchantia_polymorpha,Reclinomonas_americana), (others)	165	Compatible
24	(Neorickettsia_sennetsu_str._Miyayama,**Phytophthora_infestans**,Escherichia_coli_str._K-12_substr._MG1655,Neisseria_lactamica_020-06), (others)	165	Incompatible
25	(Hemiselmis_andersenii,Phytophthora_infestans,Marchantia_polymorpha,Reclinomonas_americana), (others)	163	Compatible
26	(**Hemiselmis_andersenii**,Neorickettsia_sennetsu_str._Miyayama,Escherichia_coli_str._K-12_substr._MG1655,Neisseria_lactamica_020-06), (others)	163	Incompatible
27	(Mesostigma_viride,Phytophthora_infestans,Marchantia_polymorpha,Reclinomonas_americana), (others)	161	Compatible
28	(Hemiselmis_andersenii,Mesostigma_viride,Phytophthora_infestans,Marchantia_polymorpha), (others)	160	Compatible
29	(Hemiselmis_andersenii,Phytophthora_infestans,Reclinomonas_americana,Rhodomonas_salina), (others)	159	Compatible
30	(**Hemiselmis_andersenii**,Wolbachia_endosymbiont_of_Culex_quinquefasciatus_Pel,Neorickettsia_sennetsu_str._Miyayama,Anaplasma_phagocytophilum_HZ,Ehrlichia_canis_str._Jake), (others)	156	Incompatible
31	(Magnetococcus_sp._MC-1,Escherichia_coli_str._K-12_substr._MG1655,**Rhodomonas_salina**,Neisseria_lactamica_020-06), (others)	156	Incompatible
32	(Hemiselmis_andersenii,Mesostigma_viride,Marchantia_polymorpha,Reclinomonas_americana), (others)	155	Compatible
33	(**Neorickettsia_sennetsu_str._Miyayama,**Candidatus_pelagibacter_IMCC9063,Alphaproteobacterium_HIMB114,Candidatus_Pelagibacter_ubique_SAR11_HTCC1002), (others)	155	Incompatible
34	(Hemiselmis_andersenii,Phytophthora_infestans), (others)	152	Compatible
35	(Hemiselmis_andersenii,Phytophthora_infestans,Marchantia_poly	149	Compatible

	morpha,Reclinomonas_americana,Rhodomonas_salina), (others)		
36	(Mesostigma_viride,Phytophthora_infestans,Marchantia_polymorpha,Rhodomonas_salina), (others)	148	Compatible
37	(Hemiselmis_andersenii,Mesostigma_viride,Phytophthora_infestans,Marchantia_polymorpha,Reclinomonas_americana), (others)	146	Compatible
38	(Phytophthora_infestans,Marchantia_polymorpha), (others)	145	Compatible
39	(Phytophthora_infestans,Marchantia_polymorpha,Reclinomonas_americana,Rhodomonas_salina), (others)	144	Compatible
40	(**Hemiselmis_andersenii**,Magnetococcus_sp._MC-1,Escherichia_coli_str._K-12_substr._MG1655,Neisseria_lactamica_020-06), (others)	144	Incompatible
41	(Candidatus_pelagibacter_IMCC9063,Alphaproteobacterium_HIMB114,Candidatus_Pelagibacter_ubique_SAR11_HTCC1002), (others)	142	Compatible
42	(**Orientia_tsutsugamushi_str._Boryong**,Candidatus_pelagibacter_IMCC9063,Alphaproteobacterium_HIMB114,Candidatus_Pelagibacter_ubique_SAR11_HTCC1002), (others)	141	Incompatible
43	(**Mesostigma_viride**,Magnetococcus_sp._MC-1,Escherichia_coli_str._K-12_substr._MG1655,Neisseria_lactamica_020-06), (others)	140	Incompatible
44	(**Neorickettsia_sennetsu_str._Miyayama**,Alphaproteobacterium_HIMB59), (others)	137	Incompatible
45	(Hemiselmis_andersenii,Mesostigma_viride,Marchantia_polymorpha,Reclinomonas_americana,Rhodomonas_salina), (others)	136	Compatible
46	(Hemiselmis_andersenii,Mesostigma_viride,Phytophthora_infestans,Marchantia_polymorpha,Rhodomonas_salina), (others)	135	Compatible
47	(Anaplasma_phagocytophilum_HZ,Ehrlichia_canis_str._Jake), (others)	134	Compatible
48	(Neorickettsia_sennetsu_str._Miyayama,**Marchantia_polymorp**	131	Incompatible

	ha), (others)		
49	(Candidatus_pelagibacter_IMCC9063,Alphaproteobacterium_HIMB114,Alphaproteobacterium_HIMB59,Candidatus_Pelagibacter_ubique_SAR11_HTCC1002), (others)	131	Compatible
50	(Hemiselmis_andersenii,Mesostigma_viride,Phytophthora_infestans,Reclinomonas_americana,Rhodomonas_salina), (others)	131	Compatible

Supplementary Table 2. Overview of 30 eukaryotic lineages selected for identifying the mitochondria-derived nuclear genes.

Eukaryotes	**Phylum**	**GC content (%)**	**Protein coding genes**
Allomyce macrogynus	Fungi	61.6	17600
Batrachochytrium dedrobatidis	Fungi	39.3	8818
Cryptococcus neoformans	Fungi	48.3	6967
Enterocytozoon bieneusi	Fungi	33.7	3632
Saccharomyce cerevisiae	Fungi	38.2	5886
Spizellomyces punctatus	Fungi	47.6	8804
Encephalitozoon intestinalis	Fungi	41.5	1833
Nectria haematococca	Fungi	50.8	15708
Nosema ceranae	Fungi	25.3	2060
Caenorhabditis elegans	Metazoa	35.4	23894
Homo sapiens	Metazoa	41.6	33610
Schistosoma mansoni	Metazoa	35.8	13191
Strongylocentrotus purpuratus	Metazoa	37.7	42420
Trichoplax adhaerens	Metazoa	32.7	11520
Drosophila melanogaster	Metazoa	42.3	22152
Nematostella vectensis	Metazoa	40.6	24780
Arabidopsis thaliana	Viridiplantae	36.1	33200
Chlamydomonas reinhardtii	Viridiplantae	63.8	14412
Micromonas pusilla	Viridiplantae	65.9	10109
Plasmodium falciparum	Alveolata	19.4	3180
Tetrahymena thermophila	Alveolata	22.3	24725
Dictyostelium discoideum	Amoebozoa	22.4	13267
Entamoeba histolytica	Amoebozoa	24.3	8163
Leishmania major	Euglenozoa	59.7	8335
Trypanosoma brucei	Euglenozoa	46.4	9079

Phytophthora infestans	Stramenopiles	51	17797
Thalassiosira pseudonana	Stramenopiles	46.9	10660
Monosiga brevicollis	Choanoflagellida	54.8	9171
Giardia lamblia	Diplomonadida	49.2	6502
Naegleria gruberi	Heterolobosea	33.1	15711

Supplementary Table 3. List of 192 α-proteobacterial genomes used in the phylogenomic analysis.

Species	**NCBI taxon ID**
Acetobacter aceti NBRC 14818	887700
Acetobacter pasteurianus IFO 3283-01	634452
Acidiphilium cryptum JF-5	349163
Agrobacterium radiobacter K84	311403
Agrobacterium tumefaciens str. C58	176299
Agrobacterium vitis S4	311402
Ahrensia sp. R2A130	744979
Alphaproteobacterium HIMB114	684719
Alphaproteobacterium HIMB5	859653
Alphaproteobacterium HIMB59	744985
Anaplasma centrale str. Israel	574556
Anaplasma marginale str. Florida	320483
Anaplasma marginale str. St. Maries	234826
Anaplasma phagocytophilum HZ	212042
Asticcacaulis excentricus CB 48	573065
Azorhizobium caulinodans ORS 571	438753
Azospirillum sp. B510	137722
Bartonella bacilliformis KC583	360095
Bartonella clarridgeiae 73	696125
Bartonella grahamii as4aup	634504
Bartonella henselae str. Houston-1	283166
Bartonella quintana str. Toulouse	283165
Bartonella tribocorum CIP 105476	382640
Beijerinckia indica subsp. indica ATCC 9039	395963
Bradyrhizobium japonicum USDA 110	224911
Bradyrhizobium sp. BTAi1	288000

Bradyrhizobium sp. ORS 278	114615
Brevundimonas subvibrioides ATCC 15264	633149
Brucella abortus bv. 1 str. 9-941	262698
Brucella abortus S19	430066
Brucella canis ATCC 23365	483179
Brucella melitensis ATCC 23457	546272
Brucella melitensis biovar Abortus 2308	359391
Brucella melitensis bv. 1 str. 16M	224914
Brucella microti CCM 4915	568815
Brucella ovis ATCC 25840	444178
Brucella suis 1330	204722
Brucella suis ATCC 23445	470137
Candidatus Caedibacter acanthamoebae	244581
Candidatus Hodgkinia cicadicola Dsem	573234
Candidatus Liberibacter asiaticus str. psy62	537021
Candidatus Liberibacter solanacearum CLso-ZC1	658172
Candidatus Magnetococcus yuandaducum	304587
Candidatus Midichloria mitochondrii IricVA	696127
Candidatus Odyssella thessalonicensis L13	985867
Candidatus Paracaedibacter acanthamoebae	91604
Candidatus Paracaedibacter symbiosus	244582
Candidatus Pelagibacter sp. HTCC7211	439493
Candidatus Pelagibacter sp. IMCC9063	1002672
Candidatus Pelagibacter ubique HTCC1002	314261
Candidatus Pelagibacter ubique HTCC1062	335992
Candidatus Puniceispirillum marinum IMCC1322	488538
Caulobacter crescentus CB15	190650
Caulobacter crescentus NA1000	565050
Caulobacter segnis ATCC 21756	509190

Caulobacter sp. K31	366602
Chelativorans sp. BNC1	266779
Citreicella sp. SE45	501479
Citromicrobium bathyomarinum JL354	685035
Citromicrobium sp. JLT1363	517722
Commensalibacter intestini A911	1088868
Dinoroseobacter shibae DFL 12	398580
Ehrlichia canis str. Jake	269484
Ehrlichia chaffeensis str. Arkansas	205920
Ehrlichia ruminantium str. Gardel	302409
Ehrlichia ruminantium str. Welgevonden	254945
Endosymbiont of Acanthamoeba UWC8	876852
Erythrobacter litoralis HTCC2594	314225
Geminicoccus roseus DSM 18922	1089551
Gluconacetobacter diazotrophicus PAl 5	272568
Gluconacetobacter hansenii ATCC 23769	714995
Gluconobacter oxydans 621H	290633
Granulibacter bethesdensis CGDNIH1	391165
Hirschia baltica ATCC 49814	582402
Hyphomicrobium denitrificans ATCC 51888	582899
Hyphomonas neptunium ATCC 15444	228405
Inquilinus limosus DSM 16000	1122125
Jannaschia sp. CCS1	290400
Ketogulonicigenium vulgare Y25	880591
Kordiimonas gwangyangensis DSM 19435	1122137
Labrenzia aggregata IAM 12614	384765
Loktanella vestfoldensis SKA53	314232
Magnetococcus marinus MC-1	156889
Magnetospirillum magneticum AMB-1	342108

Maricaulis maris MCS10	394221
Maritimibacter alkaliphilus HTCC2654	314271
Meganema perideroedes DSM 15528	1122218
Mesorhizobium ciceri biovar biserrulae WSM1271	765698
Mesorhizobium loti MAFF303099	266835
Methylobacterium chloromethanicum CM4	440085
Methylobacterium extorquens AM1	272630
Methylobacterium extorquens DM4	661410
Methylobacterium extorquens PA1	419610
Methylobacterium nodulans ORS 2060	460265
Methylobacterium populi BJ001	441620
Methylobacterium radiotolerans JCM 2831	426355
Methylobacterium sp. 4-46	426117
Methylocella silvestris BL2	395965
Micavibrio aeruginosavorus ARL-13	856793
Neorickettsia risticii str. Illinois	434131
Neorickettsia sennetsu str. Miyayama	222891
NHP bacterium	1274402
Nitrobacter hamburgensis X14	323097
Nitrobacter winogradskyi Nb-255	323098
Novosphingobium aromaticivorans DSM 12444	279238
Oceanibulbus indolifex HEL-45	391624
Oceanicaulis sp. HTCC2633	314254
Oceanicola batsensis HTCC2597	252305
Ochrobactrum anthropi ATCC 49188	439375
Octadecabacter antarcticus 307	391626
Oligotropha carboxidovorans OM5	504832
Orientia tsutsugamushi str. Boryong	357244
Orientia tsutsugamushi str. Ikeda	334380

Paracoccus denitrificans PD1222	318586
Paracoccus sp. TRP	412597
Parvibaculum lavamentivorans DS-1	402881
Parvularcula bermudensis HTCC2503	314260
Pelagibaca bermudensis HTCC2601	314265
Phaeobacter gallaeciensis ANG1	1002340
Phenylobacterium zucineum HLK1	450851
Pseudovibrio sp. JE062	439495
Rhizobium etli CFN 42	347834
Rhizobium etli CIAT 652	491916
Rhizobium leguminosarum bv. trifolii WSM1325	395491
Rhizobium leguminosarum bv. trifolii WSM2304	395492
Rhizobium leguminosarum bv. viciae 3841	216596
Rhodobacter capsulatus SB 1003	272942
Rhodobacter sphaeroides 2.4.1	272943
Rhodobacter sphaeroides ATCC 17025	349102
Rhodobacter sphaeroides ATCC 17029	349101
Rhodobacter sphaeroides KD131	557760
Rhodobacteraceae bacterium HTCC2083	314270
Rhodobacteraceae bacterium HTCC2150	388401
Rhodobacteraceae bacterium HTCC2255	367336
Rhodobacterales bacterium Y4I	439496
Rhodomicrobium vannielii ATCC 17100	648757
Rhodopseudomonas palustris BisA53	316055
Rhodopseudomonas palustris BisB18	316056
Rhodopseudomonas palustris BisB5	316057
Rhodopseudomonas palustris CGA009	258594
Rhodopseudomonas palustris DX-1	652103
Rhodopseudomonas palustris HaA2	316058

Rhodopseudomonas palustris TIE-1	395960
Rhodospirillum centenum SW	414684
Rhodospirillum rubrum ATCC 11170	269796
Rhodothalassium salexigens DSM 2132	1188247
Rhodovibrio salinarum DSM 9154	1089552
Rickettsia africae ESF-5	347255
Rickettsia akari str. Hartford	293614
Rickettsia bellii OSU 85-389	391896
Rickettsia bellii RML369-C	336407
Rickettsia canadensis str. McKiel	293613
Rickettsia conorii str. Malish 7	272944
Rickettsia endosymbiont of Ixodes scapularis	444612
Rickettsia felis URRWXCal2	315456
Rickettsia massiliae MTU5	416276
Rickettsia peacockii str. Rustic	562019
Rickettsia prowazekii str. Madrid E	272947
Rickettsia rickettsii str. 'Sheila Smith'	392021
Rickettsia rickettsii str. Iowa	452659
Rickettsia typhi str. Wilmington	257363
Roseibium sp. TrichSKD4	744980
Roseobacter denitrificans OCh 114	375451
Roseomonas cervicalis ATCC 49957	525371
Roseospira marina DSM 15113	140057
Roseospirillum parvum DSM 12498	83401
Roseovarius sp. TM1035	391613
Rubellimicrobium thermophilum DSM 16684	1123069
Ruegeria pomeroyi DSS-3	246200
Ruegeria sp. TM1040	292414
Sagittula stellata E-37	388399

Silicibacter sp. TrichCH4B	644706
Sinorhizobium fredii NGR234	394
Sinorhizobium medicae WSM419	366394
Sinorhizobium meliloti 1021	266834
Sphingobium chlorophenolicum L-1	690566
Sphingobium japonicum UT26S	452662
Sphingomonas wittichii RW1	392499
Sphingopyxis alaskensis RB2256	317655
Starkeya novella DSM 506	639283
Stella vacuolata DSM 5901	1123295
Sulfitobacter sp. EE-36	52598
Terasakiella pusilla DSM 6293	1123355
Thalassiobium sp. R2A62	633131
Thalassospira sp. TrichSKD10	744981
Wolbachia endosymbiont of Culex quinquefasciatus Pel	570417
Wolbachia endosymbiont of Drosophila melanogaster	163164
Wolbachia endosymbiont strain TRS of Brugia malayi	292805
Wolbachia sp. wRi	66084
Xanthobacter autotrophicus Py2	78245
Zymomonas mobilis subsp. mobilis NCIMB 11163	622759
Zymomonas mobilis subsp. mobilis ZM4	264203

Chapter 3. Phylogenomic reconstruction of the mitochondrial ancestors

Abstract

Reconstruction of mitochondrial ancestor has great impact on our understanding of mitochondrial evolution. Although many studies have been trying to infer the mitochondrial ancestor, the results were largely limited by the sparse genome sampling and the less informative ancestral time point they aimed for. In this study, we first reconstructed the metabolism of the last common ancestor of all mitochondria (proto-mitochondrion) by identifying the mitochondria-derived nuclear genes. Then we reconstructed the last common ancestor of mitochondria and α-proteobacteria (pre-mitochondrion) using a Bayesian character mapping method. In contrast with a diverse metabolism suggested by previous studies, our reconstruction shows that the function of proto-mitochondria was highly specialized as a primitive organelle. In addition, our phylogenomic reconstruction reveals several novel insights into the mitochondria-derived eukaryotic metabolism including the lipid metabolism. Reconstruction of the pre-mitochondrion suggests that it was most likely a parasitic bacterium. Intriguingly, it is predicted to possess a plastid/parasite type of ATP/ADP translocase that imported ATP from the host, which posits the pre-mitochondrion as an "energy scavenger" that directly contrasts with the current role of mitochondria as the cell's energy producer. In addition, it is predicted to encode a large number of flagella genes and several cytochrome oxidases functioning under the low oxygen level, providing strong evidence supporting the previous finding that the mitochondrial ancestor was likely motile and capable of oxidative phosphorylation under micro-oxic condition. Finally, our reconstruction finds a lack of evidence for the "hydrogen hypothesis" and instead supports the alternative "oxygen scavenger hypothesis" for the origin of mitochondria.

Introduction

Mitochondria are eukaryotic organelles with a bacterial origin. Known as the endosymbiotic theory, it is now widely accepted that mitochondria originated once from an α-proteobacterium probably two billion years ago (Lang et al. 1999). However, it remains unclear what constituted the initial endosymbiosis between the ancestral α-proteobacterium and its host (Andersson et al. 1998; Martin and Muller 1998; Gray et al. 1999). Specifically, what was the role played by the mitochondrial ancestor that initiated the endosymbiosis? Were mitochondria originated under oxic, microoxic, or anoxic condition? Did the mitochondria arise at the same time as, or subsequent to, the appearance of the nucleus? What is the driving force behind the initial symbiosis (Martin and Muller 1998; Andersson et al. 2003)? Several hypotheses have been proposed to account for the circumstances of the founding endosymbiotic events (Embley and Martin 2006; Koonin 2010). The "hydrogen hypothesis", proposed by Martin et al, hypothesizes that the metabolic syntrophy between a H_2-producing α-proteobacterium and a H_2-dependent archaeon as the driving force behind the endosymbiosis (Martin and Muller 1998). This hypothesis allows the possibility of a simultaneous origination of the mitochondrion and the nucleus, with the same α-proteobacterium also contributing to the rise of the nucleus by fusing its genome with the host genome. Therefore, the "hydrogen hypothesis" directly challenges the traditional serial endosymbiosis model in which the host is posited to be a full-fledged, nucleus-containing (but amitochondriate) eukaryote. In contrast, the "oxygen scavenger" hypothesis proposes that the removal of the toxic oxygen by the α-proteobacterium from the anaerobic host has driven the initial symbiosis (Andersson et al. 2003). The circumstances under which the founding events occurred remain highly debated.

Reconstructing the gene complement of the mitochondria ancestor can shine light on the origin of mitochondria. Estimation of its genome size will help us better understand the timing of its signature reductive evolution. More importantly, reconstruction of its metabolism will help elucidate the driving force of the endosymbiosis by testing the alternative hypotheses. For example, whether the mitochondrial ancestor was aerobic or anaerobic is a key yet debated point among different hypotheses (Martin and Muller 1998; Andersson et al. 2003). The hydrogen hypothesis supports anaerobic syntrophy whereas the oxygen scavenger hypothesis supports aerobic mutualism. In addition, the hydrogen hypothesis requires that the mitochondrial ancestor possessed a functional hydrogen producing machinery, which could be used to distinguish it from other hypotheses. A recent study predicted the presence of both flagella and a cytochrome cbb3 oxidase in the mitochondrial ancestor and suggested that the mitochondrial ancestor was motile and capable of oxidative phosphorylation under micro-oxic condition (Sassera et al. 2011). However, this prediction was largely based on the analysis of one bacterial genome and needs to be evaluated with additional genomic data.

Although reconstructing the mitochondrial ancestral state is the key to the understanding of the origin of mitochondria, it faces a multitude of problems. Firstly, there have been massive gene losses since its origination. For example, mitochondrion of *Reclinomonas americana*, the most primitive mitochondrion recognized so far, encodes only 67 proteins in its genome (Lang et al. 1997). Dramatic metabolic turnover occurred with mitonchodria's transformation from a bacterium to an organelle. Vast majority of mitochondrial genes have been either lost or transferred to the nucleus (Gray et al. 1999), resulting in the highly reduced genomes of modern mitochondria which only encode proteins functioning in translation and energy conversion. Therefore, identifying genes that were transferred from mitochondria to nucleus (hereafter

referred as mitochondria-derived nuclear genes) is a prerequisite to reconstructing the mitochondrial ancestor.

Secondly, a robust phylogenetic relationship is required for the ancestral reconstruction. However, the closest contemporary relatives of mitochondria remain elusive. Although mitochondria have been firmly placed within α-proteobacteria, their phylogenetic position within the group remains uncertain. Weak phylogenetic signal and serious systematic errors, such as long-branch attraction and sequence compositional bias, all hamper the effort to pinpoint the origin of the mitochondria. Nevertheless, recent phylogenomic studies with increasing genomic sampling have started to form a consensus by placing mitochondria in or near the *Rickettsiales* order (Andersson et al. 1998; Wu et al. 2004; Fitzpatrick et al. 2006; Williams et al. 2007), although its affiliation with the order of *Rhodospirillales* has also been suggested (Esser et al. 2004). The *Rickettsiales* order itself is a highly diversified group with at least two major lineages: a group of obligate intracellular bacterial parasites including *Rickettsia, Ehrlichia*, and *Anaplasma* (Viale and Arakaki 1994; Gupta 1995) and a group of marine bacteria *Pelagibacter* that are known to be the smallest free-living bacteria. Although there were debates on which lineage forms the sister clade of the mitochondria (Georgiades et al. 2011; Thrash et al. 2011), recent studies have consistently shown that the placement of *Pelagibacter* within *Rickettsiales* is likely a tree artifact caused by sequence compositional bias (Rodriguez-Ezpeleta and Embley 2012; Viklund et al. 2012; Viklund et al. 2013). With increased taxon sampling and better phylogenetic markers and methods, our recent phylogenomic study for the first time placed mitochondria unequivocally within the *Rickettsiales* order, as a sister clade to the *Rickettsiaceae* and *Anaplasmaceae* families, all subtended to the *Holosporaceae* family (Ref ##).

Previous studies bypassed the problem of uncertain phylogeny by reconstructing the last common ancestor of all mitochondria, hereafter referred as proto-mitochondrion. To address the problem of massive gene loss, mitochondria-derived nuclear genes were identified and added to the ancestral mitochondrial gene pool. For example, by looking for nuclear genes that cluster with α-proteobacterial homologs in the gene trees, Karlberg et al. identified at least 38 yeast nuclear genes that were putatively transferred from mitochondria (Karlberg et al. 2000). With a similar phylogenetic approach but a broader sampling of bacterial and eukaryotic genomes, Gabaldon et al. identified a total of 630 putative mitochondria-derived nuclear gene families (Gabaldon and Huynen 2003), and 842 gene families with an updated genome set (Gabaldon and Huynen 2007), substantially augmenting the mitochondrial gene pool. Although extremely insightful, results from these studies have their limitations when it comes to understanding the origin of mitochondria. Figure 1 illustrates our point. Point A represents the last common ancestor of mitochondria and α-proteobacteria (hereafter referred as pre-mitochondrion), while point B represents the last common ancestor of all mitochondria (proto-mitochondrion). Mitochondrion emerged somewhere between points A and B, i.e., after it split off from α-proteobacteria but before the divergence of eukaryotic lineages (point C). All previous studies essentially reconstructed the proto-mitochondria, because they simply pooled all known mitochondrial genes (including genes that have been transferred to the nucleus). Considering the dramatic transformation after the origin of mitochondria, and the massive gene loss associated with this transformation, reconstructing the proto-mitochondria would only reveal little of what it looked like at the origin of mitochondria and therefore provide limited insights on the initial endosymbiosis event.

In order to understand what was happening at the beginning of the endosymbiosis, ideally we should reconstruct the ancestral state at time point C. It is difficult to delineate point C in the

tree, however, because the endosymbiosis event is not associated with any lineage diversification. If point B represents an end point for studying the origin of mitochondria, then point A is a good starting point. Therefore, reconstructing the pre-mitochondria at point A would be the logical next step for us to gain better insights into the origin of mitochondria.

In this study, we set out to infer the ancestral gene complement of the pre-mitochondria. Using a phylogenomic approach and with a substantially increased eukaryotic and α-proteobacterial representation, we firstly revisited the mitochondria-derived nuclear genes and reconstructed the proto-mitochondria. We then used a Bayesian character mapping method to reconstruct the pre-mitochondria. The reconstructed pre-mitochondria possessed a diversified metabolism typical of an obligate intracellular bacterium. In comparison, the reconstructed proto-mitochondria had a substantially reduced metabolic capacity that was functionally very close to modern mitochondria. Finally, we evaluated the alternative hypotheses based on our ancestral state reconstruction.

Results and Discussion

Identifying mitochondria-derived nuclear genes

Using a phylogenomic approach, Gabaldon et al. identified a set of 842 mitochondria-derived nuclear gene families and reconstructed a diverse proto-mitochondrial metabolism typical of an aerobic endosymbiont catabolizing lipids, glycerol, and amino acids provided by the eukaryotic host (Gabaldon and Huynen 2007). However, their results were based on a rather limited availability of bacterial and eukaryotic genomes at the time of their study. Leveraging on a substantially larger number of eukaryotic and α-proteobacterial genomes, we performed a large-scale phylogenomic analysis to identify mitochondria-derived nuclear genes. Using SSU rRNA

phylogeny as a guide, we selected 30 eukaryotic genomes representing a broad range of phylogenetic diversity. Each of 427,186 genes within these genomes was subject to firstly a BLASTP screening. Sequences with α-proteobacterial homologs in the top five BLASTP hits were then clustered into gene families, followed by phylogenetic tree reconstruction for each family. In particular, we looked for a specific pattern in the tree where eukaryotic nuclear genes were clustered with α-proteobacterial or mitochondrial homologs. To eliminate recent, lineage-specific gene transfers between some endosymbionts and their eukaryotic hosts (e.g., between *Trichoplax adhaerens* and its *Rickettsial* endosymbiont (Driscoll et al. 2013)), we asked at least two α-proteobacteria and two eukaryotic lineages to be present in each α-proteobacteria/mitochondria/eukaryotes cluster. In total, 4,459 genes belonging to 394 families were identified as mitochondria-derived nuclear genes. The number of gene families varied markedly from 3 to 156 over the 30 eukaryotic representatives (Supplementary Figure 1). Notably, five amitochondriate eukaryotes all showed evidence of mitochondria-to-nucleus lateral gene transfers, supporting that mitochondria did once exist in these lineages (Peyretaillade et al. 1998; Roger et al. 1998; Mai et al. 1999).

We evaluated the specificity and sensitivity of our method by estimating the false positive and false negative rates respectively. To estimate the false positive rate, we benchmarked our procedure using the phylum of *Deinococcus/Thermus*, which shows no known close relationship with eukaryotic lineages. Of all the 427,186 eukaryotic sequences we screened, only 278 sequences in 44 families were clustered with *Deinococcus/Thermus* (false positive rate 0.07%), indicating our procedure had a very high specificity. To estimate the false negative rate, we used *Reclinomonas americana*, the hitherto most primitive mitochondrial genome as the positive control. Considering that some of the *R. americana* mitochondrial genes (e.g., 2 hypothetical proteins) have diverged too far to reliably identify their homologs, we used a subset of 50 *R.*

americana mitochondria genes that were present in at least 2 other mitochondrial genomes. 46 out of 50 genes were recovered by our procedure (false negative rate 8%), indicating that our procedure is also very sensitive.

To gain insight into the metabolic features of the mitochondria-derived nuclear genes, we assigned 394 families to Clusters of Orthologous Groups (COGs) (Tatusov et al. 2000) and mapped them onto the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto 2000). We note that a large number of the COGs (31.6%) belong to two categories: translation and energy production (Figure 2), which are also the two main functions of modern mitochondria-encoded genes. These COGs include genes involved in pyruvate metabolism, TCA cycle, electron transport and ribosomal biogenesis (Figure 3). The missing stretches in the TCA cycle (from citrate to 2-oxoglutarate) are consistent with previous phylogenetic studies, which showed complex evolutionary histories of these genes (Kurland and Andersson 2000; Schnarrenberger and Martin 2002; Gabaldon and Huynen 2003; Gabaldon and Huynen 2007). A large number of COGs are involved in other functions of modern mitochondria. For example, metabolic pathways were almost completely recovered for fatty acids biosynthesis, beta-oxidation, branched-chain amino acids degradation (Leucine, Valine, Isoleucine) and the biosynthesis of ubiquinone, biotin and one carbon unit pool, all of which are functional in the current organelle. Conversely, functional categories such as DNA replication and transcription are largely absent in our reconstructed metabolism, and the heterotrophic carbohydrate metabolisms such as glycolysis and pentose phosphate pathway are entirely missing. Therefore, our reconstruction suggests that proto-mitochondria have already substantially reduced its genome and is functionally very close to the modern mitochondria.

Overall, our result is similar to Gabaldon's reconstruction. Well-characterized mitochondrial functions such as TCA cycle, electron transfer chain, ATP synthesis and translation were recovered in both studies. As a result, there are many genes in common (Figure 2). However, our reconstructed proto-mitochondria are much leaner than Gabaldon's. We identified 394 gene families compared to Gabaldon's 842. Our reconstruction depicts streamlined proto-mitochondria highly similar to the modern mitochondria, while Gabaldon study suggests an ancestor with more diverse functions (Figure 2) (Gabaldon and Huynen 2007).

Compared to Gabaldon's study, our analysis is more sensitive. Despite identifying a smaller set of genes, we were able to fill many of the gaps in well-characterized mitochondrial pathways that were present in Gabaldon's reconstruction. For example, in the pyruvate metabolism and TCA cycle, we identified a pyruvate dehydrogenase E3 subunit (*lpdA/pdhD*) (COG1249), a succinyl-CoA synthetase (*sucD*, COG0074) and a succinate dehydrogenase (*sdh2*, COG0479), which are essential for the pathways but were all missing in Gabaldon's reconstruction. Similarly for ATP synthesis, we added one F0F1-type ATP synthase subunit (*atpC*, COG0355), three NADH dehydrogenases (*nuoB*, COG0377; *ND4*, COG1008; *ND5*, COG1009) and four cytochrome c components (*fbcC*, COG2857; *cyb561*, COG3038; COG3474; *cycM*, COG5274), completing a functional electron transport chain. For the assembly of iron-sulfur cluster, we added *iscU* (COG0822) and *iscA* (COG0820), two critical scaffold proteins upon which the cluster is assembled and transferred. In terms of the translation machinery, we added a total of 19 ribosomal proteins, along with 8 translation factors (*IF-2*, COG0532; *EF-P*, COG0231 and 6 GTPase (COG0012, COG0206, COG0050, COG2262, COG0218, COG1159)) and 3 aminoacyl-tRNA synthetases (COG0124, COG0162, COG0180). Also, we added a *COQ3* (COG2227) enzyme involved in the ubiquinone biosynthesis, essentially recovering a fully functional *de novo* ubiquinone biosynthesis pathway. In the biotin metabolism, we added a biotin-protein

ligase (*birA*, COG4285), the key enzyme that connects the biotin metabolism with fatty acid biosynthesis.

Our analysis also has a higher specificity. Of the 156 gene families we identified that have human nuclear gene as a member, 104 (66.7%) families are present in the human mitochondrial proteome, compared with 121 out of 355 (34.1%) families in Gabaldon et al. 2007. Similar results were also observed in Yeast and Arabidopsis (Table 1). In addition, 56.1% (221 out of 394) of the families identified in our study contain at least one gene with a N-terminal mitochondria targeting signal, compared to 30.7% (258 out of 842) in Gabaldon et al. 2007. Since mitochondria-derived nuclear genes are often recruited back to mitochondria, the higher percentage of mitochondria-localized nuclear gene families in our reconstruction indicates a higher specificity of our results.

The reasons for the increased sensitivity and specificity in our results could be at least two folds. First, our phylogenomic analysis used a much larger genome dataset representing a substantially broader range of taxon sampling (1,613 genomes in our study compared to 144 in Gabaldon's 2007 study) (Table 1). In particular, the number of genomes was considerably higher in both the *Rhodospirillales* and *Rickettsiales* orders that have shown close relationships to mitochondria (Table 1). The phylogenetic diversity of eukaryotic genomes was also greatly increased in our sampling, including 6 novel phyla that had not been sampled in previous studies (Supplementary Figure 1) (Gabaldon and Huynen 2003; Gabaldon and Huynen 2007). Having a broader taxon sampling improved the phylogenetic analysis, which enabled us to more reliably trace the evolutionary history of gene families. Second, in Gabaldon study, nuclear genes that clustered with β- and γ-proteobacteria were also identified as mitochondria-derived nuclear genes. The rationale for including β- and γ-proteobacteria in their analysis is to increase the recovery rate of

the *R. americana* mitochondrial genes. For example, most of the ribosomal proteins in *R. americana* mitochondrial genome were found to cluster with β- and γ-proteobacteria in their phylogenetic analyses (Gabaldon and Huynen 2007). However, with increased α-proteobacterial genome sampling we only identified two ribosomal protein genes with such a spurious pattern in our analysis. Therefore, we think the criterion used in Gabaldon study is unnecessarily relaxed and could increase the number of false positives. For example, genes involved in the pentose phosphate pathway all clustered with γ-proteobacteria or a mixture of γ- and α-proteobacteria and were identified as mitochondria-derived in Gabaldon study. Therefore, we think the number of mitochondria-derived nuclear genes was likely overestimated by the previous study (Gabaldon and Huynen 2007).

Novel insights into the mitochondria-derived eukaryotic metabolisms

Our reconstruction provides several novel insights regarding mitochondria-derived eukaryotic metabolisms. Of particular interest are a number of genes involved in the eukaryotic lipid metabolism (Table 2). For example, we identified three enzymes involved in the steroid biosynthesis, including a squalene/phytoene synthase (COG1562), a sterol-C5-desaturase (COG3000) and a 1-deoxy-D-xylulose-5-phosphate synthase (COG1154), suggesting that the mitochondrial ancestor also contributed to the eukaryotic steroid biosynthesis. The mitochondrial origin of these enzymes is supported by functional studies. Mitochondria are known to play an essential role in the biosynthesis of steroid by providing sites for the onset of the process (Duarte et al. 2012). In return, steroids are also critical in maintaining the mitochondrial morphology (Prince and Buttle 2004). Indeed, studies in *C. neoformans* and *T. brucei* indicated that mutants of squalene synthase and sterol desaturase were defective in mitochondrial membrane integrity (Ingavale et al. 2008; Perez-Moreno et al. 2012).

In addition, we identified a ceramide glycosyltransferase (COG1215) involved in the glycosphingolipids (GSL) biosynthesis, carrying out the ceramides glycosylation reactions. Interestingly, this enzyme is located at the "mitochondria-associated membrane (MAM)", a specific ER subdomain that bridges between the ER and mitochondria (Ardail et al. 2003). Both glycosphingolipids and ceramides are ubiquitously present as essential membrane components in almost all eukaryotic cells and mitochondria, but are rarely identified in bacteria. Accordingly, the substrates and glycolipid products of the bacterial and eukaryotic glycosyltransferases were suggested to be very different (Holzl et al. 2005). Therefore, the bacterial origin of this gene indicates an acquisition of novel function by eukaryotes for synthesizing its own endomembranes and for the crosstalk and lipid trafficking between mitochondria and ER.

Interestingly, we identified four enzymes (*lpxD*, COG1044; *lpxA,* COG1043; *lpxB,* COG0763 and *kdtA,* COG1519) involved in the biosynthesis of lipid A. As part of lipopolysaccharide (LPS), lipid A is an essential component of the outer membrane of gram-negative bacteria. It was only recently found to be present in certain eukaryotes, including some green algae and the vascular plant *Arabidopsis thaliana*, and its role in eukaryotes is largely unclear (Armstrong et al. 2006). Recently it has been suggested that in *A. thaliana* the lipid A is likely synthesized in mitochondria and subsequently transported to chloroplast (Duncan et al. 2011; Li et al. 2011a). In support of this finding, our results indicate that the lipid A biosynthesis pathway in eukaryotes was likely acquired from the mitochondrial ancestor. All four proteins are present in *A. thaliana* and *Phytophthora infestans* among the eukaryotic genomes in our analysis, while *lpxA* is present in *T. adhaerans* and was likely acquired separately from its own endosymbiont (Driscoll et al. 2013). It has been suggested that proteins with similar phylogenetic distribution are likely to functionally interact in the same biological process (Pellegrini et al. 1999). In light of this, we used phylogenetic profile analysis and identified one additional gene family (Group_1713) with

an unknown function that had the same distribution pattern. Although most of the members in this family were annotated as hypothetical proteins, COG assignment of this family hit the predicted nucleoside-diphosphate-sugar epimerase (COG3660) (evalue <= 1.3e-50), which is involved in the glycolipid metabolism (Li et al. 2011b). It is therefore reasonable to believe that this hypothetical gene is likely also involved in the lipid A biosynthesis. And it would be of particular interest to investigate the location of this protein *in vivo* and its potential role in the mitochondria and chloroplast function. Notably, 3 of the 4 eukaryotic members of this family showed a strong signal for mitochondrial localization in their protein sequences (TargetP specificity >= 0.9).

Other than the lipid biosynthesis, our results also shed light on the mitochondrial contribution to other eukaryotic metabolisms. For instance, we identified several genes (*purD*, COG0151; *purM*, COG0150; *mutT,* COG1051; *pyrD,* COG0167) involved in the *de novo* nucleotide biosynthesis as mitochondria-derived. Both *purD* and *purM* belong to the family of glycinamide ribonucleotide transformylase (GART) and catalyze different steps in the *de novo* purine biosynthesis. Mitochondria also contribute to the cytosolic purine biosynthesis by providing formate as the one-carbon unit. Consistently, the entire formate biosynthesis pathway is identified as mitochondria-derived in our results. On the other hand, *pyrD* is a mitochondria localized protein critical for the pyrimidine biosynthesis (Desler et al. 2010). *purD* and *purM* have been previously identified as of mitochondrial origin by a phylogenomic analysis using the *Wolbachia wMel* genome but were missing in the results of Gabaldon et al. 2007 (Wu et al. 2004; Gabaldon and Huynen 2007). Also, we identified a number of genes (*glmS*, COG0449; *wecB*, COG1940; *neuB*, COG2089; *murA*, COG0766) involved in the UDP-sugar biosynthesis. The UDP-sugar provides essential modifications to various target proteins such as nuclear pore proteins and cytoskeleton components (Hanover 2001). Thus it is tempting to speculate that these

mitochondria-derived genes might participate in controlling the activity of these eukaryotic-specific complexes.

Reconstructing the metabolism of pre-mitochondria

To reconstruct pre-mitochondria, first we need to place mitochondria firmly within a robust α-proteobacterial species tree. Of particular importance is to identify the closest contemporary relatives of mitochondria among α-proteobacteria. Using an integrated phylogenomic approach, we were able to refine the position of mitochondria and for the first time placed mitochondria unequivocally within the *Rickettsiales* order. Mitochondria form a sister clade to the *Rickettsiaceae* and *Anaplasmataceae* families, both subtended by the *Holosporaceae* family. We used this tree topology as the phylogenetic framework to reconstruct pre-mitochondria in this study.

We clustered mitochondrial and α-proteobacterial genes into orthologous gene families and treated each gene family as a character. We mapped the presence/absence of each gene family onto the leaves of the tree and used BayesTraits, a Bayesian character mapping software, to reconstruct the ancestral gene complement of pre-mitochondria. The Bayesian method has been shown to be superior to both the parsimony method and maximum likelihood method in accounting for uncertainties in both model parameters and phylogeny (Pagel et al. 2004; Vanderpoorten and Goffinet 2006). The mitochondria-derived nuclear genes, mitochondria-encoded genes and 148,485 genes of the 49 α-proteobacterial representatives were first classified into COGs. Sequences that cannot be assigned to a COG were then clustered into families using MCL (Enright et al. 2002), to create “expanded COGs”. Totally 4873 original COGs plus 3210 expanded COGs were created and mapped to the mitochondrial and α-proteobacterial species.

Using this approach, pre-mitochondria were predicted to possess 887 COGs. Based on the approximate linear relationships among the number of gene families, the number of genes and genome sizes (Supplementary Figure 2), we estimated the size of pre-mitochondrial genome to be 1.5 - 1.6 Mb, with 1100 - 1300 genes. This is typical of an obligate intracellular bacterium. Figure 5 shows the reconstructed metabolism of pre-mitochondria. Compared to highly specialized proto-mitochondria, pre-mitochondria were capable of much more diversified metabolism. In addition to the major pathways involved in translation (13.6%), cell wall, LPS and membrane biogenesis (8.3%), energy production (7.2%), and replication, recombination and repair (7.1%) (Figure 2), it was predicted to possess multiple key metabolic pathways including glycolysis, TCA cycle, pentose phosphate pathway, and fatty acid biosynthesis pathway, which indicates that pre-mitochondria were capable of generating ATP and at least several essential intermediates on its own. Also, pre-mitochondria possessed a large number of genes involved in synthesizing various cofactors, such as riboflavin, folate, biotin and ubiquinone. On the other hand, similar to most *Rickettsiales*, pre-mitochondria possessed a limited number of genes involved in amino acid biosynthesis. It was incapable of synthesizing any amino acid *de novo*, and was only able to synthesize certain amino acids (Glutamine, Leucine, Valine and Isoleucine) from metabolic intermediates. Therefore pre-mitochondria had to obtain most of its essential amino acids from the host. Accordingly, at least 5 known amino acid transporters were predicted in pre-mitochondria.

Pre-mitochondria were predicted to lack most of the genes involved in the *de novo* nucleotide biosynthesis pathway, except for a few genes such as *purD* and *pyrD* which were also present in proto-mitochondria. Among *Rickettsiales*, the *de novo* nucleotide biosynthesis pathway is present in the family *Anaplasmataceae* but absent in all other lineages. Hence one interpretation is that

the *de novo* nucleotide biosynthesis pathway was acquired in *Anaplasmataceae*, as indicated in our reconstruction. However, because the gain of the entire *de novo* nucleotide biosynthesis pathway, including 12 purine biosynthesis genes and 6 pyrimidine biosynthesis genes, is extremely unlikely in these intracellular bacteria, we think that the nucleotide biosynthesis pathway was most likely present in pre-mitochondria, and was subsequently lost multiple times in both *Rickettsiales* (except for *Anaplasmataceae*) and mitochondria.

Pre-mitochondrion was an energy parasite

Pre-mitochondria were predicted to have the plastid/parasite type of ATP/ADP translocase (posterior probability 0.93), the hallmark protein of many obligate intracellular bacteria that is used to import ATP from the host. The ATP/ADP translocase commonly functions as an ATP/ADP antiporter that exchanges bacterial ADP for the host cell ATP as a source of energy (Schmitz-Esser et al. 2004). In addition, it has been shown that some intracellular bacteria, including *Chlamydia* and *Rickettsia*, encode additional isoforms of this protein for the uptake of nucleotides to compensate for their inability to synthesize nucleotides *de novo* (Tjaden et al. 1999; Audia and Winkler 2006). Consistently, this gene family is absent in the *Anaplasmataceae* family of *Rickettsiales,* members of which all possess complete *de novo* nucleotide biosynthesis pathway (Wu et al. 2004; Brayton et al. 2005; Mavromatis et al. 2006). Previous studies have suggested that there were ancient lateral gene transfers of this gene between the ancestors of *Chlamydiales*, *Rickettsiales* and plastids (Amiri et al. 2003; Greub and Raoult 2003; Schmitz-Esser et al. 2004). Our phylogenetic analysis shows that the gene tree is largely congruent with the species tree of the *Rickettsiales* order (Supplementary Figure 3), suggesting that this gene has been vertically inherited in *Rickettsiales* and thus was most likely present in their last common ancestor and by inference, pre-mitochondria. If the nucleotide biosynthesis pathway was also

present in pre-mitochondria as we predicted, then this gene most likely functioned as an ATP/ADP exchanger instead of a nucleotide transporter in pre-mitochondria.

Remarkably, the plastid/parasite ATP/ADP translocase is evolutionarily unrelated to and functionally distinct from the ATP/ADP translocase in modern mitochondria, which exhibits an opposite polarity by exporting ATP into the host cytosol (Andersson and Kurland 1998; Wolf et al. 1999). Therefore, our reconstruction posits pre-mitochondria as an "energy scavenger" and suggests an energy parasitism between the endosymbiont and its host at the origin of mitochondria, as first proposed by Andersson et al. (Amiri et al. 2003; Andersson et al. 2003). This is in sharp contrast with the current role of mitochondria as the cell's energy producer and contradicts the traditional serial endosymbiotic theory that the symbiosis was driven by the symbiont supplying the host ATP (John and Whatley 1975; Whatley et al. 1979). The replacement of plastid/parasite ATP/ADP translocase by mitochondrial ATP/ADP translocase occurred subsequently, resulting in a reverse flow of ATP between the mitochondria and its host. This remarkable transformation in energy metabolism might mark the transition of mitochondria from a parasitic endosymbiont to a mutualistic organelle (Kurland and Andersson 2000).

A recent systematic survey of symbiosis has shown that bacterial mutualisms can originated either directly from environmental free-living bacteria or from intracellular parasites (Sachs et al. 2011). A key difference between these two evolutionary routes is that to initiate symbiosis, free-living bacteria need to offer immediate benefits to the host while parasitic bacteria do not (Ewald 1987). Our results suggest that mitochondria most likely originated from an obligate intracellular parasite and not from a free-living bacterium. Importantly it implies that when the endosymbiosis started, mitochondrial ancestor provided no benefits to the host. Accordingly, the benefits proposed by various hypotheses (e.g., oxygen scavenger hypothesis) should be irrelevant in

explaining the establishment of the initial symbiosis, even though they might be crucial in driving the transition of mitochondria from a parasite to a mutualistic organelle at a later stage.

Pre-mitochondrion possessed flagella

A recent study has suggested that the free-living mitochondrial ancestor possessed a flagellum (Sassera et al. 2011). This prediction was based on the presence of 26 flagellar genes in one *Rickettsiales* species, *Candidatus Midichloria mitochondrii*. We recently sequenced five novel and phylogenetically divergent members of *Candidatus Midichloriaceae* and *Holosporaceae* families in *Rickettsiales* (*Endosymbiont of Acanthamoebae UWC8, Candidatus Caedibacter acanthamoebae, Candidatus Paracaedibacter acanthamoebae, Candidatus Paracaedibacter symbiosus, NHP bacterium*) (Figure 4). Interestingly, most of these 26 flagellar genes were also found in four out of the five endosymbionts. The only exception is the *Candidatus Caedibacter acanthamoebae*, which possesses only 5 flagellar genes. The flagellar genes are also present in the recently sequenced *Holosporaceae* endosymbiont *Candidatus Odyssella thessalonicensis*. Phylogenetic analysis of the flagella genes indicated that they evolved vertically in *Rickettsiales* species. Consistently, they form syntenic gene clusters (Figure 6). It is therefore not surprising that pre-mitochondria were predicted to possess 25 COGs involved in the flagellum biosynthesis. These 25 COGs encode the core components of flagellum, including basal body, motor, hook, rod, filament and export apparatus (Liu and Ochman 2007). Therefore, these novel *Rickettsiales* lineages provide very strong evidence supporting the presence of a flagellum in pre-mitochondria. Electron microscopy of the *NHP bacterium* has shown flagella at the basal end of its cell (Bradley-Dunlop et al. 2004). It is interesting to note that one recent study also observed flagella in two endosymbionts of *Paramecium* belonging to the *Lyticum* genus of the *Midichloriaceae* family (Boscaro et al. 2013). However, electron microscopic examination of the

other four amoeba endosymbionts revealed no evidence of a flagellum (data not shown), thus it remains unclear how these flagellar genes actually function in these amoeba endosymbionts.

Pre-mitochondrion was capable of respiration at low oxygen condition

We predicted the presence of three COGs of cbb3-type cytochrome oxidase (ccoP, COG2010; ccoO, COG2993, ccoN, COG3278) and two COGs encoding its accessory proteins (ccoG, COG0348, ccoI, COG2217; posterior probability 0.56) in pre-mitochondria. cbb3-type cytochrome oxidases belong to the C-family cytochrome oxidase mainly functioning under the micro-oxic condition. All five components of cbb3 oxidases were identified in *Candidatus Midichloria mitochondrii* but were absent in other previously sequenced *Rickettsiales* species (Sassera et al. 2011). Of the five endosymbiont genomes we sequenced, cbb3 oxidases were found only in *Endosymbiont of Acanthamoeba UWC8*, the sister clade of *Candidatus Midichloria mitochondrii*. It is possible that cbb3 oxidases have been lost in mitochondria and other *Rickettsiales* lineages. However, we cannot exclude the possibility that cbb3 oxidases were gained in the *Candidatus Midichloriaceae* lineage. Interestingly, however, we also predicted the presence of two COGs of cytochrome bd-type quinol oxidase (COG1271, COG1294, posterior probability 0.89) in pre-mitochondria. Similar with the flagella genes, the bd-type oxidases are widely distributed in both *Holosporaceae* and *Candidatus Midichloriaceae*, including *Candidatus Odyssella thessalonicensis*, and four out of the five endosymbionts we sequenced, with the *Candidatus Caedibacter acanthamoebae* being the only exception. Thus this strongly suggests that the bd-type oxidases were present in pre-mitochondria and have been lost in mitochondria and other *Rickettsiales* lineages. Like cbb3-type cytochrome oxidases, the bd-type oxidases are functional under limited oxygen level. Our study therefore provides additional and stronger evidence that pre-mitochondria were capable of oxidative phosphorylation under low oxygen condition. Taken together, our reconstruction supports the hypothesis that the

mitochondrial ancestor was likely motile with an active role in interacting with its host, and it was capable of generating ATP under low oxygen condition under which the origin of mitochondria was initiated (Sassera et al. 2011).

Oxygen scavenger hypothesis or Hydrogen hypothesis

Reconstructing pre-mitochondria shines light on what might have driven constitute the initial symbiosis between the mitochondrial ancestor and its host. One key piece of evidence that could distinguish the oxygen scavenger hypothesis and the hydrogen hypothesis is whether the mitochondrial ancestor possessed a hydrogen-producing machinery. Two known hydrogen-producing pathways exist in bacteria. One is known as the nitrogenase-dependent hydrogen production, which is a side-reaction along with the nitrogen fixation, and is only present in certain nitrogen-fixing *Cyanobacteria*. In this pathway, hydrogen is produced by Fe-S-cluster-containing hydrogenase (NrfC, COG0437). The other pathway, which is more widely distributed, involves the anaerobic processing of pyruvate. Here the pyruvate reduces ferredoxin by pyruvate:ferredoxin oxidoreductase (PFO, COG0674). The ferredoxin is then oxidized by a ferredoxin hydrogenase (COG4624), reducing proton to H_2. The latter pathway is also found in hydrogenosomes of some aminochondriate eukaryotes. However, none of the components in either of the two pathways were found in our reconstructed pre-mitochondria. The PFO-related pathway is restricted only to certain α-proteobacteria lineages, such as *Rhodospirillum rubrum* and *Rhodopseudomonas palustris*, and was likely absent before the divergence of all *Rickettsiales* and mitochondria. Thus based on the current data, the hydrogen-producing machinery, the key component of the hydrogen hypothesis, seems unlikely to be present in the mitochondria ancestor.

Secondly, the hydrogen hypothesis requires the host (in this case, an archaea) being strictly autotrophic at the initial symbiotic event. This was argued for two main reasons: 1) all known hydrogen-dependent methanogens are autotrophy, including those utilizing acetate and any reduced one carbon compound as carbon source. 2) If both host and symbiont grew heterotrophically, competition is more likely than syntrophy (Martin and Muller, 1998). Therefore, the eukaryotic heterotrophy pathways, including both the carbohydrate importers and metabolism (glycolysis, pentose phosphate pathway, and carbohydrate interconversion) had to be later acquired from the endosymbiont to the host (Martin and Muller 1998). However, the mitochondrial origin of the eukaryotic heterotrophy was not supported by our results. Previous studies have shown that the eukaryotic glycolysis pathway was not originated from α-proteobacteria (Canback et al. 2002; Gabaldon and Huynen 2003; Gabaldon and Huynen 2007) but several components of the pentose phosphate pathway were (Gabaldon and Huynen 2007). Our results confirm the non-α-proteobacteria origin of the eukaryotic glycolysis pathway. However, our results disagree with Gabaldon et al. 2007 study and indicate that the pentose phosphate pathway is not of α-proteobacterial origin either. Eukaryotic sequences of these families were mostly found as sister clades with γ-proteobacteria in their phylogenetic study, which we argue should not be used as evidence for a mitochondrial origin. In addition, none of the eukaryotic glucose transport apparatus were of endosymbiotic origin in our results.

Taken together, our ancestral state reconstruction shows a lack of evidence for the hydrogen hypothesis. Instead, the reconstruction of complete aerobic pathways in pre-mitochondria, coupled with numerous antioxidant components, including one glutathione S-transferase (COG0625), one thioredoxin reductase (COG0492), two peroxiredoxin (COG0678, COG1225), two glutaredoxin-related proteins (COG0278, COG0695) and three thioredoxin-like proteins (COG0694, COG2143, COG3118), is consistent with the role of mitochondrial ancestor as an

oxygen detoxifier as suggested by the oxygen scavenger hypothesis. Sequencing additional α-proteobacterial species closely related to the mitochondria (i.e within *Holosporaceae* and *Candidatus Midichloriaceae*) might provide additional evidence to distinguish between these two alternative hypotheses. compatible with oxygen

Conclusion

Using a broad range of α-proteobacterial and eukaryotic genomes, our phylogenomic analysis significantly improves the accuracy and confidence of the mitochondrial ancestral reconstruction. In this study, we reconstructed the mitochondrial ancestors at two key points, pre-mitochondria and proto-mitochondria. In contrast to previous reconstructions suggesting that proto-mitochondria possessed a versatile metabolism, our results showed proto-mitochondria were already well adapted and functionally specialized as a primitive organelle. On the other hand, reconstruction of pre-mitochondria suggested that it was most likely an obligate intracellular energy parasite capable of oxidative phosphorylation under micro-oxic condition. Therefore, the massive metabolic turnover likely occurred from pre-mitochondria to proto-mitochondria, earlier than previously appreciated (Gabaldon and Huynen 2007). With the reconstruction of pre-mitochondria, we found no evidence supporting the hydrogen hypothesis. However, our results are consistent with the "oxygen scavanging" as the driving force for the origin of mitochondria.

Material and Methods

Selection of eukaryotic nuclear genomes for phylogenomic analysis

The phylogenetic distribution of all sequenced eukaryotic genomes was retrieved from GenomeOnline database (GOLD, http://genomesonline.org/). 30 eukaryotic genomes representing a broad range of phylogenetic diversity were selected and used for identifying the mitochondria-derived nuclear genes (*Allomyce macrogynus, Arabidopsis thaliana, Batrachochytrium dedrobatidis, Caenorhabditis elegans, Chlamydomonas reinhardtii, Cryptococcus neoformans, Dictyostelium discoideum, Drosophila melanogaster, Encephalitozoon intestinalis, Entamoeba histolytica, Enterocytozoon bieneusi, Giardia lamblia, Homo sapiens, Leishmania major, Micromonas pusilla, Monosiga brevicollis, Naegleria gruberi, Nectria haematococca, Nematostella vectensis, Nosema ceranae, Phytophthora infestans, Plasmodium falciparum, Saccharomyce cerevisiae, Schistosoma mansoni, Spizellomyces punctatus, Strongylocentrotus purpuratus, Tetrahymena thermophila, Thalassiosira pseudonana, Trichoplax adhaerens, Trypanosoma brucei*).

Identification of mitochondria-derived nuclear genes

For every single gene of 30 eukaryotic nuclear genomes, an initial BLASTP search was performed against all complete bacterial, archaeal, and mitochondrial genomes. A eukaryotic gene was retained for further phylogenetic analysis if its top five hits contained an α-proteobacterial or mitochondrial sequence (e-value cutoff 1e-4). All eukaryotic genes passing the initial BLAST search were clustered into families using the Markov Cluster Algorithm (Enright et al. 2002). Families that were present in at least two eukaryotic species were selected for phylogenetic analysis. For each of retained protein family, its homologs from all complete bacterial genomes were retrieved by BLASTP search (evalue cutoff 1e-15). Protein sequences were aligned using MAFFT (Katoh et al. 2002) and trimmed using ZORRO (Wu et al. 2012). Phylogenetic trees were constructed using FastTree 2 (Price et al. 2010). When possible, each individual tree was rooted using three different rooting methods, rooting with Archaea or

Deinococcus as the outgroup or midpoint rooting. Each of the rooted trees was scanned for a bipartition where eukaryotic genes were clustered with their α-proteobacterial or mitochondrial homologs. A partition was retained as one gene family if it contained at least two eukaryotes and two α-proteobacterial species. Paralogs, if existed in a family, were separated and each was treated as a new family of mitochondria-derived nuclear genes.

Functional annotation of mitochondria-derived nuclear genes

Mitochondria-derived nuclear genes were classified into Clusters of Orthologous Groups (COGs) by hidden Markov model search using HMMer3 (Eddy 1998). To reconstruct metabolic pathways, genes were mapped onto Kyoto Encyclopedia of Genes and Genomes (KEGG) database using KEGG Automatic Annotation Server (KAAS) (Moriya et al. 2007) with "bi-directional best hit (BBH)" as the assignment method.

α-proteobacteria and mitochondria species tree

We adopted a phylogenomic approach to reconstruct the α-proteobacteria and mitochondria phylogeny. We selected a set of 49 α-proteobacterial representatives using a tree-based greedy algorithm to maximize their phylogenetic diversity (Steel 2005). A set of six eukaryotic lineages (*Cryptococcus neoformans, Arabidopsis thaliana, Nematostella vectensis, Spizellomyces punctatus, Monosiga brevicollis, Phytophthora infestans*) were selected as mitochondrial representatives. We used a set of 29 mitochondria-derived nuclear genes identified previously as phylogenetic markers, which has been shown to have lower evolutionary rate and less compositional bias compared with the mitochondria-encoded genes. Protein sequences of the marker genes from selected genomes were identified, aligned, trimmed and concatenated using AMPHORA2 (Wu and Scott 2012). Bayesian consensus trees were reconstructed using

PhyloBayes (Lartillot and Philippe 2004) with the -CAT -GTR options, as recommended in the manual. Two independent MCMC chains were run and the chains were considered converged when the maxdiff dropped below 0.3, as suggested in the manual. The trees were sampled every 10 cycles and the beginning one fifth of the trees from each chain were discarded as burn-in.

Mitochondria ancestral state reconstruction

Using the species tree of 6 mitochondria and 49 α-proteobacteria as the phylogenetic framework, pre-mitochondria were reconstructed with a Bayesian character mapping inference algorithm implemented in BayesTraits V2 (Pagel et al. 2004). The mitochondrial genes were compiled by combining the mitochondria-derived nuclear genes with the mitochondria-encoded genes. Mitochondrial and α-proteobacterial genes were first assigned to COGs by hidden Markov model search using HMMer3 (Eddy 1998). Genes that cannot be assigned a COG were then clustered into families using the Markov Cluster Algorithm (Enright et al. 2002), creating "expanded COGs". The presence/absence of each COG in each species was treated as a binary trait and used for the ancestral state reconstruction. Gamma distribution was adopted as the prior distribution with its parameter estimated from an initial maximum likelihood analysis. The "hyperprior" option was used to reduce the uncertainty in choosing priors in the MCMC. A total number of 1,050,000 iterations were performed, with the first 50,000 cycles discarded as burn-in. The average value of each binary state in the remaining 1,000,000 cycles was then taken as the probability of the presence of each COG in the reconstructed ancestral state.

References

Amiri H, Karlberg O, Andersson SG. 2003. Deep origin of plastid/parasite atp/adp translocases. *J Mol Evol* 56:137-150.

Andersson SG, Karlberg O, Canback B, Kurland CG. 2003. On the origin of mitochondria: A genomics perspective. *Philos Trans R Soc Lond B Biol Sci* 358:165-177; discussion 177-169.

Andersson SG, Kurland CG. 1998. Reductive evolution of resident genomes. *Trends Microbiol* 6:263-268.

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG. 1998. The genome sequence of rickettsia prowazekii and the origin of mitochondria. *Nature* 396:133-140.

Ardail D, Popa I, Bodennec J, Louisot P, Schmitt D, Portoukalian J. 2003. The mitochondria-associated endoplasmic-reticulum subcompartment (mam fraction) of rat liver contains highly active sphingolipid-specific glycosyltransferases. *Biochem J* 371:1013-1019.

Armstrong MT, Theg SM, Braun N, Wainwright N, Pardy RL, Armstrong PB. 2006. Histochemical evidence for lipid a (endotoxin) in eukaryote chloroplasts. *FASEB J* 20:2145-2146.

Audia JP, Winkler HH. 2006. Study of the five rickettsia prowazekii proteins annotated as atp/adp translocases (tlc): Only tlc1 transports atp/adp, while tlc4 and tlc5 transport other ribonucleotides. *J Bacteriol* 188:6261-6268.

Boscaro V, Schrallhammer M, Benken KA, Krenek S, Szokoli F, Berendonk TU, Schweikert M, Verni F, Sabaneyeva EV, Petroni G. 2013. Rediscovering the genus lyticum, multiflagellated symbionts of the order rickettsiales. *Sci Rep* 3:3305.

Bradley-Dunlop DJ, Pantoja C, Lightner DV. 2004. Development of monoclonal antibodies for detection of necrotizing hepatopancreatitis in penaeid shrimp. *Dis Aquat Organ* 60:233-240.

Brayton KA, Kappmeyer LS, Herndon DR, Dark MJ, Tibbals DL, Palmer GH, McGuire TC, Knowles DP. 2005. Complete genome sequencing of anaplasma marginale reveals that

the surface is skewed to two superfamilies of outer membrane proteins. *Proceedings of the National Academy of Sciences of the United States of America* 102:844-849.

Canback B, Andersson SG, Kurland CG. 2002. The global phylogeny of glycolytic enzymes. *Proc Natl Acad Sci U S A* 99:6097-6102.

Desler C, Lykke A, Rasmussen LJ. 2010. The effect of mitochondrial dysfunction on cytosolic nucleotide metabolism. *J Nucleic Acids* 2010.

Driscoll T, Gillespie JJ, Nordberg EK, Azad AF, Sobral BW. 2013. Bacterial DNA sifted from the trichoplax adhaerens (animalia: Placozoa) genome project reveals a putative rickettsial endosymbiont. *Genome Biol Evol* 5:621-645.

Duarte A, Poderoso C, Cooke M, Soria G, Maciel FC, Gottifredi V, Podesta EJ. 2012. Mitochondrial fusion is essential for steroid biosynthesis. *Plos One* 7.

Duncan O, Taylor NL, Carrie C, Eubel H, Kubiszewski-Jakubiak S, Zhang B, Narsai R, Millar AH, Whelan J. 2011. Multiple lines of evidence localize signaling, morphology, and lipid biosynthesis machinery to the mitochondrial outer membrane of arabidopsis. *Plant Physiol* 157:1093-1113.

Eddy SR. 1998. Profile hidden markov models. *Bioinformatics* 14:755-763.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. *Nature* 440:623-630.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575-1584.

Esser C, Ahmadinejad N, Wiegand C, et al. 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol* 21:1643-1660.

Ewald PW. 1987. Transmission modes and evolution of the parasitism-mutualism continuum. *Ann N Y Acad Sci* 503:295-306.

Fitzpatrick DA, Creevey CJ, McInerney JO. 2006. Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the rickettsiales. *Mol Biol Evol* 23:74-85.

Gabaldon T, Huynen MA. 2003. Reconstruction of the proto-mitochondrial metabolism. *Science* 301:609.

Gabaldon T, Huynen MA. 2007. From endosymbiont to host-controlled organelle: The hijacking of mitochondrial protein synthesis and metabolism. *PLoS Comput Biol* 3:e219.

Georgiades K, Madoui MA, Le P, Robert C, Raoult D. 2011. Phylogenomic analysis of odyssella thessalonicensis fortifies the common origin of rickettsiales, pelagibacter ubique and reclimonas americana mitochondrion. *PLoS ONE* 6:e24857.

Gray MW, Burger G, Lang BF. 1999. Mitochondrial evolution. *Science* 283:1476-1481.

Greub G, Raoult D. 2003. History of the adp/atp-translocase-encoding gene, a parasitism gene transferred from a chlamydiales ancestor to plants 1 billion years ago. *Appl Environ Microbiol* 69:5530-5535.

Gupta RS. 1995. Evolution of the chaperonin families (hsp60, hsp10 and tcp-1) of proteins and the origin of eukaryotic cells. *Mol Microbiol* 15:1-11.

Hanover JA. 2001. Glycan-dependent signaling: O-linked n-acetylglucosamine. *FASEB J* 15:1865-1876.

Harris JK, Kelley ST, Spiegelman GB, Pace NR. 2003. The genetic core of the universal ancestor. *Genome Res* 13:407-412.

Holzl G, Leipelt M, Ott C, Zahringer U, Lindner B, Warnecke D, Heinz E. 2005. Processive lipid galactosyl/glucosyltransferases from agrobacterium tumefaciens and mesorhizobium loti display multiple specificities. *Glycobiology* 15:874-886.

Ingavale SS, Chang YC, Lee H, McClelland CM, Leong ML, Kwon-Chung KJ. 2008. Importance of mitochondria in survival of cryptococcus neoformans under low oxygen conditions and tolerance to cobalt chloride. *Plos Pathogens* 4.

John P, Whatley FR. 1975. Paracoccus denitrificans and the evolutionary origin of the mitochondrion. *Nature* 254:495-498.

Kanehisa M, Goto S. 2000. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27-30.

Karlberg O, Canback B, Kurland CG, Andersson SG. 2000. The dual origin of the yeast mitochondrial proteome. *Yeast* 17:170-187.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. Mafft: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res* 30:3059-3066.

Koonin EV. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol* 11:209.

Kurland CG, Andersson SGE. 2000. Origin and evolution of the mitochondrial proteome. *Microbiology and Molecular Biology Reviews* 64:786-+.

Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Gray MW. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* 387:493-497.

Lang BF, Gray MW, Burger G. 1999. Mitochondrial genome evolution and the origin of eukaryotes. *Annu Rev Genet* 33:351-397.

Lartillot N, Philippe H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21:1095-1109.

Li C, Guan Z, Liu D, Raetz CR. 2011a. Pathway for lipid a biosynthesis in arabidopsis thaliana resembling that of escherichia coli. *Proc Natl Acad Sci U S A* 108:11387-11392.

Li CL, Wang YQ, Liu LC, Hu YC, Zhang FX, Mergen S, Wang GD, Schlappi MR, Chu CC. 2011b. A rice plastidial nucleotide sugar epimerase is involved in galactolipid biosynthesis and improves photosynthetic efficiency. *Plos Genetics* 7.

Liu R, Ochman H. 2007. Stepwise formation of the bacterial flagellar system. *Proc Natl Acad Sci U S A* 104:7116-7121.

Mai ZM, Ghosh S, Frisardi M, Rosenthal B, Rogers R, Samuelson J. 1999. Hsp60 is targeted to a cryptic mitochondrion-derived organelle ("crypton") in the microaerophilic protozoan parasite entamoeba histolytica. *Molecular and Cellular Biology* 19:2198-2205.

Martin W, Muller M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature* 392:37-41.

Mavromatis K, Doyle CK, Lykidis A, et al. 2006. The genome of the obligately intracellular bacterium ehrlichia canis reveals themes of complex membrane structure and immune evasion strategies. *J Bacteriol* 188:4015-4023.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. Kaas: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35:W182-185.

Pagel M, Meade A, Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. *Syst Biol* 53:673-684.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci U S A* 96:4285-4288.

Perez-Moreno G, Sealey-Cardona M, Rodrigues-Poveda C, Gelb MH, Ruiz-Perez LM, Castillo-Acosta V, Urbina JA, Gonzalez-Pacanowska D. 2012. Endogenous sterol biosynthesis is important for mitochondrial function and cell morphology in procyclic forms of trypanosoma brucei. *International Journal for Parasitology* 42:975-989.

Peyretaillade E, Broussolle V, Peyret P, Metenier G, Gouy M, Vivares CP. 1998. Microsporidia, amitochondrial protists, possess a 70-kda heat shock protein gene of mitochondrial evolutionary origin. *Mol Biol Evol* 15:683-689.

Price MN, Dehal PS, Arkin AP. 2010. Fasttree 2--approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490.

Prince FP, Buttle KF. 2004. Mitochondrial structure in steroid-producing cells: Three-dimensional reconstruction of human leydig cell mitochondria by electron microscopic tomography. *Anat Rec A Discov Mol Cell Evol Biol* 278:454-461.

Rodriguez-Ezpeleta N, Embley TM. 2012. The sar11 group of alpha-proteobacteria is not related to the origin of mitochondria. *PLoS ONE* 7:e30520.

Roger AJ, Svard SG, Tovar J, Clark CG, Smith MW, Gillin FD, Sogin ML. 1998. A mitochondrial-like chaperonin 60 gene in giardia lamblia: Evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. *Proc Natl Acad Sci U S A* 95:229-234.

Sachs JL, Skophammer RG, Regus JU. 2011. Evolutionary transitions in bacterial symbiosis. *Proc Natl Acad Sci U S A* 108 Suppl 2:10800-10807.

Sassera D, Lo N, Epis S, et al. 2011. Phylogenomic evidence for the presence of a flagellum and cbb(3) oxidase in the free-living mitochondrial ancestor. *Mol Biol Evol* 28:3285-3296.

Schmitz-Esser S, Linka N, Collingro A, Beier CL, Neuhaus HE, Wagner M, Horn M. 2004. Atp/adp translocases: A common feature of obligate intracellular amoebal symbionts related to chlamydiae and rickettsiae. *J Bacteriol* 186:683-691.

Schnarrenberger C, Martin W. 2002. Evolution of the enzymes of the citric acid cycle and the glyoxylate cycle of higher plants. A case study of endosymbiotic gene transfer. *Eur J Biochem* 269:868-883.

Steel M. 2005. Phylogenetic diversity and the greedy algorithm. *Syst Biol* 54:527-529.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The cog database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28:33-36.

Thrash JC, Boyd A, Huggett MJ, Grote J, Carini P, Yoder RJ, Robbertse B, Spatafora JW, Rappe MS, Giovannoni SJ. 2011. Phylogenomic evidence for a common ancestor of mitochondria and the sar11 clade. *Sci Rep* 1:13.

Tjaden J, Winkler HH, Schwoppe C, Van Der Laan M, Mohlmann T, Neuhaus HE. 1999. Two nucleotide transport proteins in chlamydia trachomatis, one for net nucleoside triphosphate uptake and the other for transport of energy. *J Bacteriol* 181:1196-1202.

Vanderpoorten A, Goffinet B. 2006. Mapping uncertainty and phylogenetic uncertainty in ancestral character state reconstruction: An example in the moss genus brachytheciastrum. *Syst Biol* 55:957-971.

Viale AM, Arakaki AK. 1994. The chaperone connection to the origins of the eukaryotic organelles. *FEBS Lett* 341:146-151.

Viklund J, Ettema TJ, Andersson SG. 2012. Independent genome reduction and phylogenetic reclassification of the oceanic sar11 clade. *Mol Biol Evol* 29:599-615.

Viklund J, Martijn J, Ettema TJ, Andersson SG. 2013. Comparative and phylogenomic evidence that the alphaproteobacterium himb59 is not a member of the oceanic sar11 clade. *PLoS ONE* 8:e78858.

Whatley JM, John P, Whatley FR. 1979. From extracellular to intracellular: The establishment of mitochondria and chloroplasts. *Proc R Soc Lond B Biol Sci* 204:165-187.

Williams KP, Sobral BW, Dickerman AW. 2007. A robust species tree for the alphaproteobacteria. *J Bacteriol* 189:4578-4586.

Wolf YI, Aravind L, Koonin EV. 1999. Rickettsiae and chlamydiae: Evidence of horizontal gene transfer and gene exchange. *Trends Genet* 15:173-175.

Wu M, Chatterji S, Eisen JA. 2012. Accounting for alignment uncertainty in phylogenomics. *PLoS One* 7:e30288.

Wu M, Scott AJ. 2012. Phylogenomic analysis of bacterial and archaeal sequences with amphora2. *Bioinformatics* 28:1033-1034.

Wu M, Sun LV, Vamathevan J, et al. 2004. Phylogenomics of the reproductive parasite wolbachia pipientis wmel: A streamlined genome overrun by mobile genetic elements. *PLoS Biol* 2:E69.

Figures

Figure 1. Different time points through the mitochondrial evolution. Time point A (pre-mitochondria) represents the last common ancestor of mitochondria and alphaproteobacteria. Time point B (proto-mitochondria) represents the last common ancestor of all contemporary mitochondria. Time point C represents the origin of mitochondria.

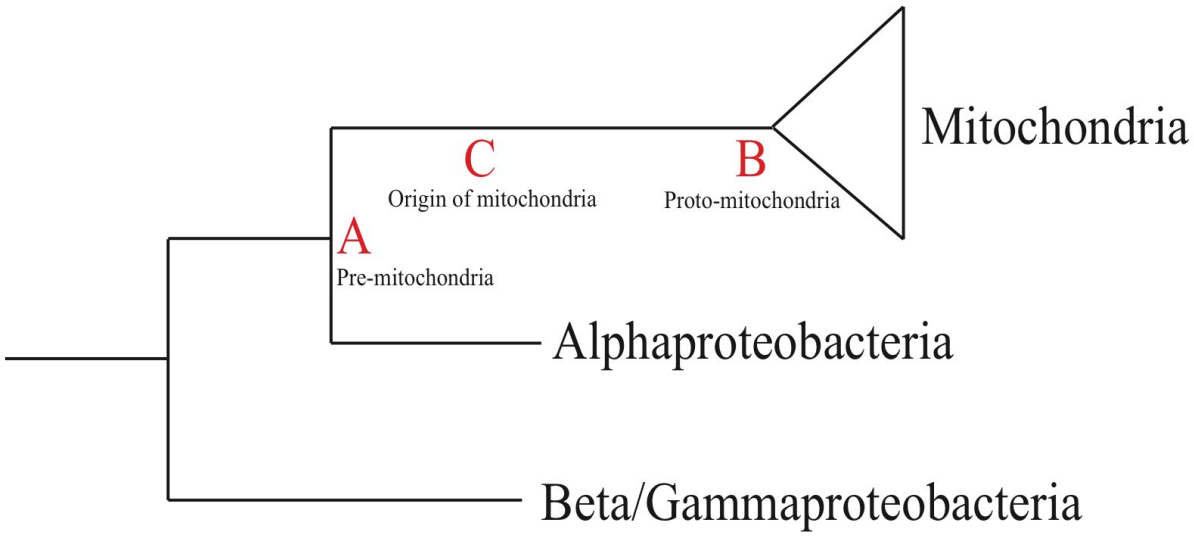


Figure 2. The distribution of COGs within each functional category in different mitochondria ancestral reconstructions. Within each class, from left to right are 1) the reconstructed proto-mitochondria in our study, 2) the reconstructed proto-mitochondria by Gabaldon et al. 2007, 3) the reconstructed pre-mitochondria in our study, 4) human mitochondrial proteome.

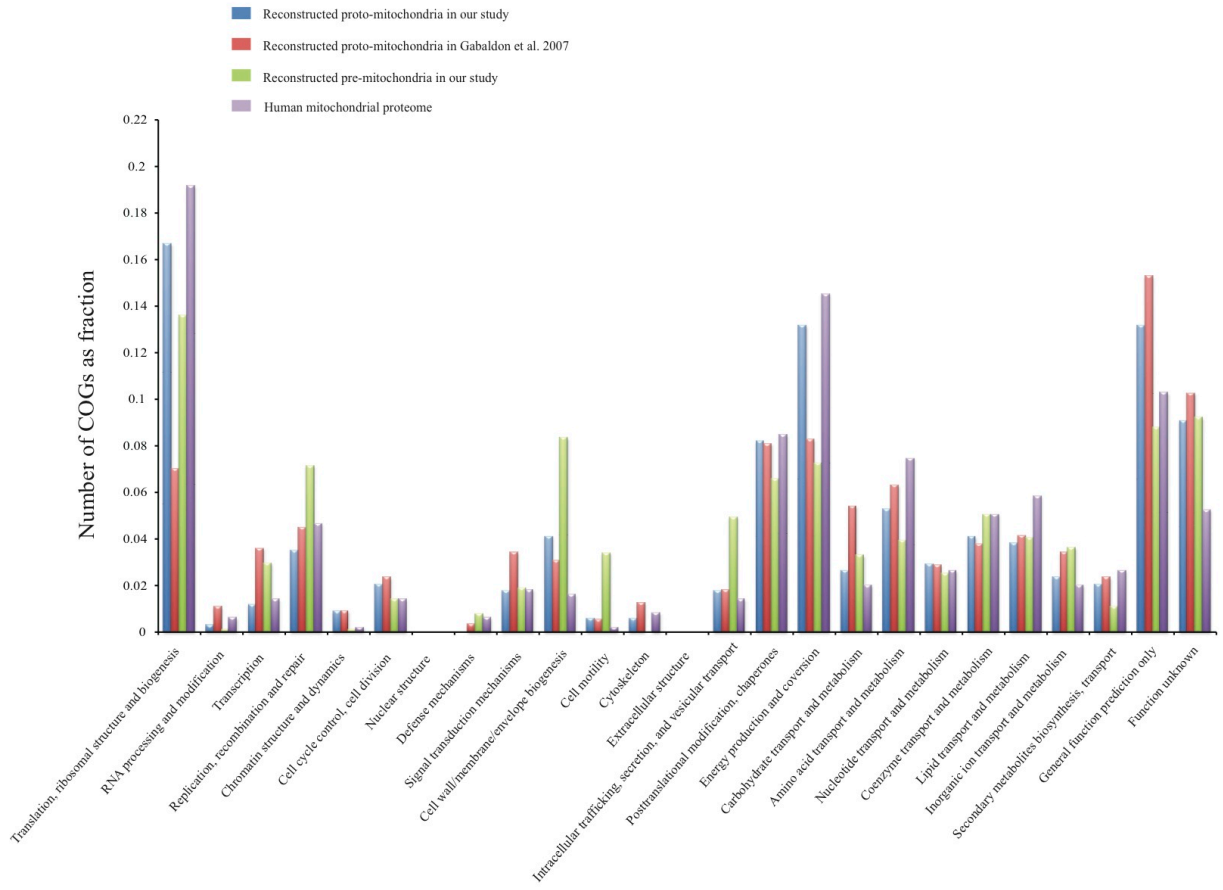


Figure 3. Reconstructed metabolism of proto-mitochondria. Black solid lines represent the genes identified only in our reconstruction. Dotted lines represent missing genes in an otherwise complete pathway in our reconstruction. Red solid lines represent the genes also present in Gabaldon et al. 2007.

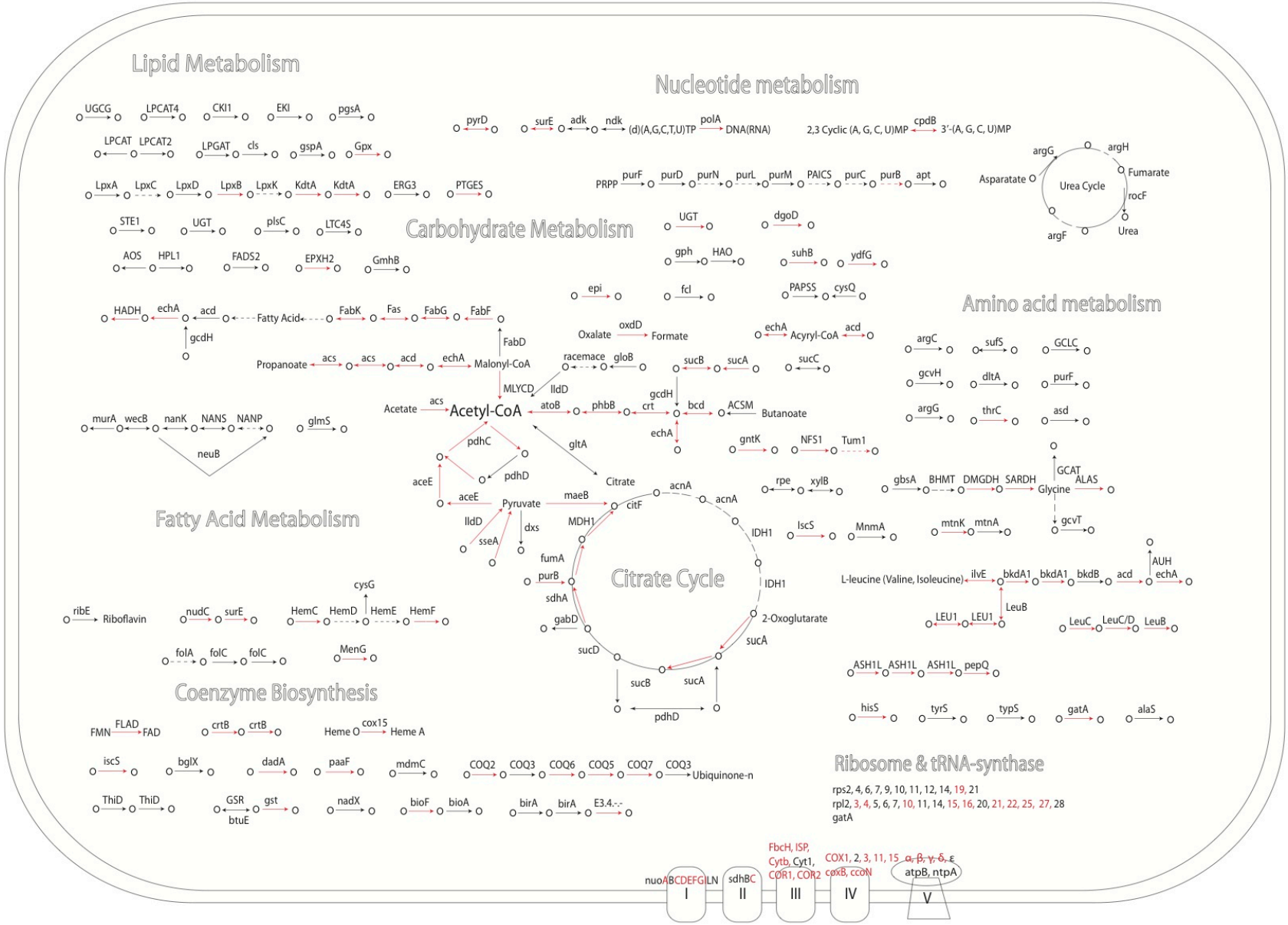

Lipid Metabolism
UGCG
LPCAT4
CKI1
EKI
pgsA
LPCAT
LPCAT2
LPGAT
cls
gspA
Gpx
LpxA
LpxC
LpxD
LpxB
LpxK
KdtA
KdtA
ERG3
PTGES
STE1
UGT
plsC
LTC4S
AOS
HPL1
FADS2
EPXH2
GmhB
Nucleotide metabolism
pyrD
surE
adk
ndk
(d)(A,G,C,T,U)TP
polA
DNA(RNA)
2,3 Cyclic (A, G, C, U)MP
cpdB
3'-(A, G, C, U)MP
PRPP
purF
purD
purN
purL
purM
PAICS
purC
purB
apt
Carbohydrate Metabolism
UGT
dgoD
gph
HAO
suhB
ydfG
epi
fcl
PAPSS
cysQ
Amino acid metabolism
argG
argH
Asparatate
Urea Cycle
Fumarate
rocF
argF
Urea
HADH
echA
acd
Fatty Acid
FabK
Fas
FabG
FabF
gcdH
FabD
Oxalate
oxdD
Formate
echA
Acyryl-CoA
acd
Propanoate
acs
acs
acd
echA
Malonyl-CoA
racemace
gloB
sucB
sucA
sucC
MLYCD
lldD
gcdH
Acetate
acs
Acetyl-CoA
atoB
phbB
crt
bcd
ACSM
Butanoate
echA
argC
sufS
GCLC
gcvH
dltA
purF
argG
thrC
asd
murA
wecB
nanK
NANS
NANP
glmS
neuB
pdhC
pdhD
aceE
aceE
gltA
gntK
NFS1
Tum1
Citrate
acnA
acnA
IDH1
IscS
MnmA
rpe
xylB
gbsA
BHMT
DMGDH
SARDH
Glycine
GCAT
ALAS
gcvT
mtnK
mtnA
Fatty Acid Metabolism
Pyruvate
maeB
citF
MDH1
fumA
purB
sdhA
gabD
sucD
sucB
sucA
pdhD
IDH1
2-Oxoglutarate
sucA
Citrate Cycle
lldD
sseA
dxs
L-leucine (Valine, Isoleucine)
ilvE
bkdA1
bkdA1
bkdB
acd
AUH
echA
LEU1
LEU1
LeuB
LeuC
LeuC/D
LeuB
ribE
Riboflavin
nudC
surE
HemC
HemD
HemE
HemF
cysG
folA
folC
folC
MenG
Coenzyme Biosynthesis
ASH1L
ASH1L
ASH1L
pepQ
hisS
tyrS
typS
gatA
alaS
FMN
FLAD
FAD
crtB
crtB
Heme O
cox15
Heme A
iscS
bglX
dadA
paaF
mdmC
COQ2
COQ3
COQ6
COQ5
COQ7
COQ3
Ubiquinone-n
ThiD
ThiD
GSR
btuE
gst
nadX
bioF
bioA
birA
birA
E3.4.-.-
Ribosome & tRNA-synthase
rps2, 4, 6, 7, 9, 10, 11, 12, 14, 19, 21
rpl2, 3, 4, 5, 6, 7, 10, 11, 14, 15, 16, 20, 21, 22, 25, 27, 28
gatA
nuoABCDEFGLN
I
sdhBC
II
FbcH, ISP,
Cytb, Cyt1,
COR1, COR2
III
COX1, 2, 3, 11, 15
coxB, ccoN
IV
atpB, ntpA
V

Figure 4. A rooted Bayesian consensus tree of 49 α-proteobacteria and 6 mitochondria made with 29 mitochondria-derived nuclear genes. The tree was rooted using β- and γ-proteobacteria as the outgroup. Asterisks indicate five endosymbiont genomes in *Rickettsiales* sequenced in our previous study. The posterior probability support values of the internal nodes are greater than 0.9 unless as indicated in the tree.


0.1
Rhodospirillum centenum SW
Micavibrio aeruginosavorus ARL-13
Rhodospirillum rubrum ATCC 11170
Magnetospirillum magneticum AMB-1
Granulibacter bethesdensis CGDNIH1
Gluconobacter oxydans 621H
Gluconacetobacter diazotrophicus PAI 5
Acidiphilium cryptum JF-5
Rhodospirillales
0.72
Zymomonas mobilis subsp. mobilis ZM4
Sphingomonas wittichii RW1
Sphingomonadales
Brevundimonas subvibrioides ATCC 15264
Asticcacaulis excentricus CB 48
Phenylobacterium zucineum HLK1
Caulobacterales
Maricaulis maris MCS10
Hirschia baltica ATCC 49814
Hyphomonas neptunium ATCC 15444
Parvularcula bermudensis HTCC2503
Rhodobacterales
Rhodomicrobium vannielii ATCC 17100
Hyphomicrobium denitrificans ATCC 51888
0.82
Xanthobacter autotrophicus Py2
Rhodopseudomonas palustris DX-1
Methylobacterium chloromethanicum CM4
Methylocella silvestris BL2
Rhizobium leguminosarum bv. trifolii WSM1325
Bartonella quintana str. Toulouse
Chelativorans sp. BNC1
Jannaschia sp. CCS1
Ruegeria sp. TM1040
0.68
Ketogulonicigenium vulgare Y25
Paracoccus denitrificans PD1222
Parvibaculum lavamentivorans DS-1
Rhizobiales
Candidatus Pelagibacter ubique SAR11 HTCC1002
Alphaproteobacterium HIMB114
Candidatus Pelagibacter IMCC9063
SAR11
Candidatus Puniceispirillum marinum IMCC1322
Mitochondria
Orientia tsutsugamushi str. Boryong
1
Rickettsia typhi str. Wilmington
Neorickettsia sennetsu str. Miyayama
Wolbachia endosymbiont of Culex quinquefasciatus Pel
Ehrlichia canis str. Jake
Anaplasma phagocytophilum HZ
0.95
Candidatus Midichloria mitochondrii IricVA
Endosymbiont of Acanthamoeba sp. UWC8 *
Ca. Midichloriaceae
Candidatus Paracaedibacter symbiosus *
Candidatus Odyssella thessalonicensis L13 *
Candidatus Paracaedibacter acanthamoebae *
Rickettsiales
NHP bacterium *
Candidatus Caedibacter acanthamoebae *
Holosporaceae
Magnetococcus marinus MC-1
Magnetococcales
Nitrosomonas sp. Is79A3
Ralstonia solanacearum GMI1000
Dechloromonas aromatica RCB
Chromobacterium violaceum ATCC 12472
Francisella tularensis subsp. tularensis FSC198
Legionella pneumophila str. Lens
Escherichia coli str. K-12 substr. MG1655
Pseudomonas aeruginosa PA7

Figure 5. Reconstructed metabolism of pre-mitochondria. Black solid lines represent the genes identified in our reconstruction while dotted lines represent missing genes in an otherwise complete pathway. Red lines represent the genes present in proto-mitochondria in Figure 3.

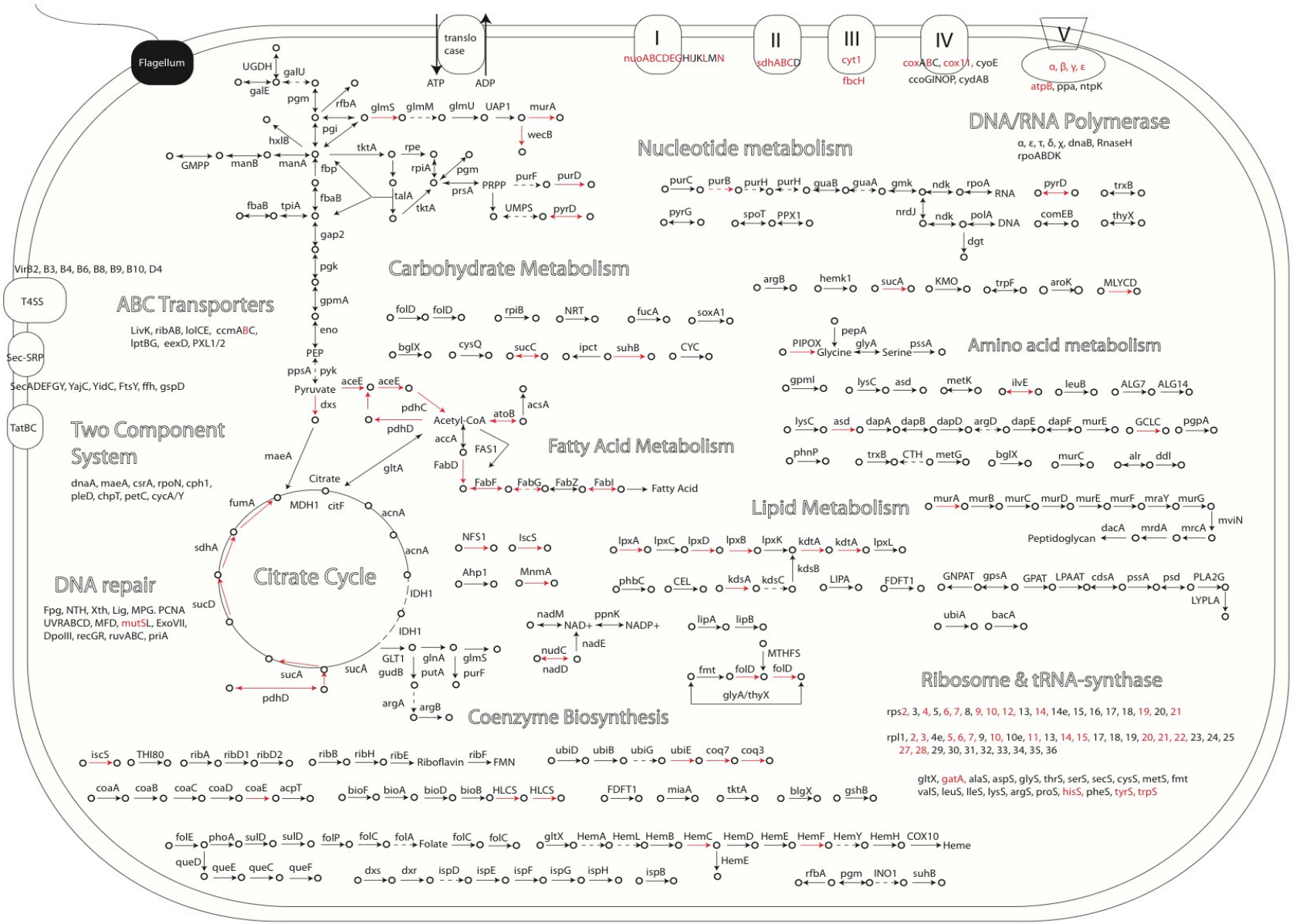
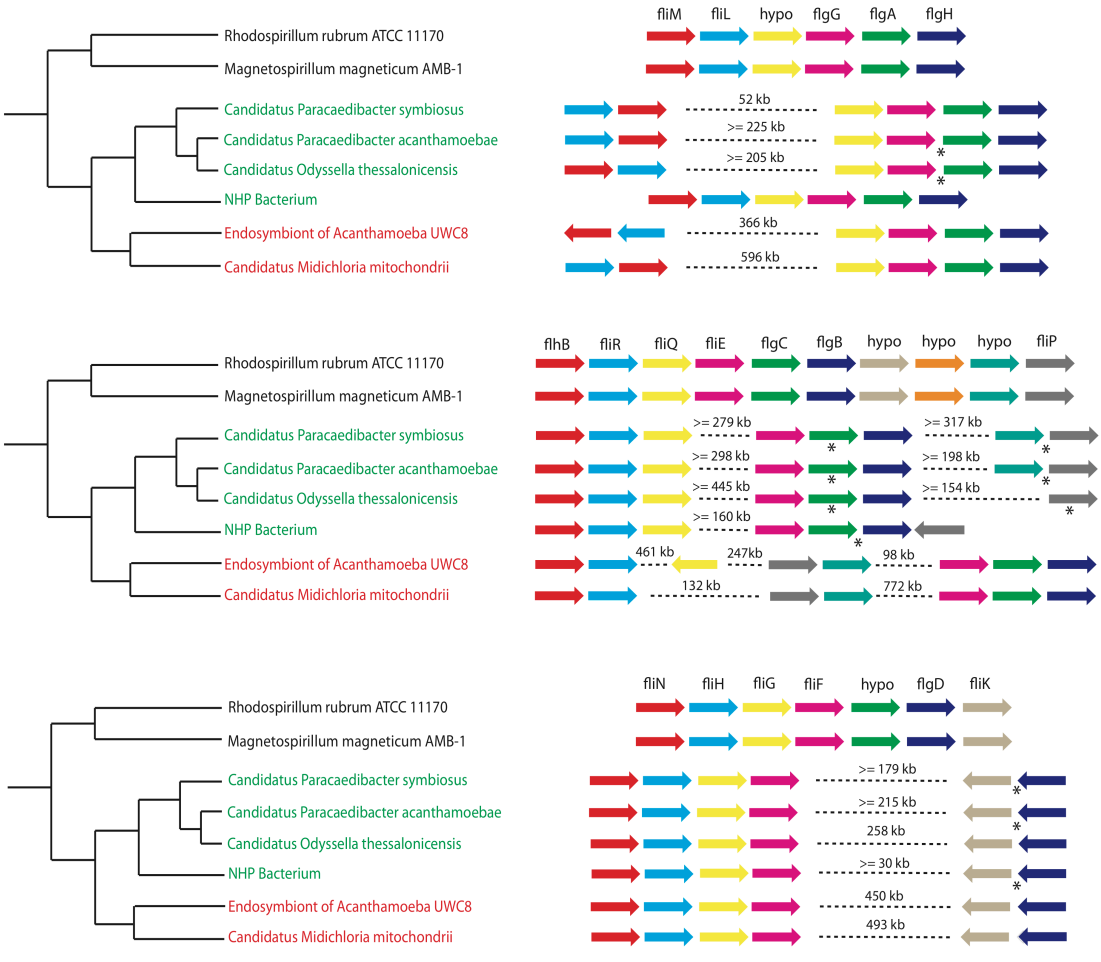

Flagellum
translo case
ATP
ADP
I
nuoABCDEGHIJKLMN
II
sdhABCD
III
cyt1
fbcH
IV
coxABC, cox1/, cyoE
ccoGINOP, cydAB
V
α, β, γ, ε
atpB, ppa, ntpK
DNA/RNA Polymerase
α, ε, τ, δ, χ, dnaB, RnaseH
rpoABDK
Nucleotide metabolism
Carbohydrate Metabolism
ABC Transporters
LivK, ribAB, lolCE, ccmABC, lptBG, eexD, PXL1/2
VirB2, B3, B4, B6, B8, B9, B10, D4
T4SS
Sec-SRP
SecADEFGY, YajC, YidC, FtsY, ffh, gspD
TatBC
Two Component System
dnaA, maeA, csrA, rpoN, cph1, pleD, chpT, petC, cycA/Y
Fatty Acid Metabolism
Amino acid metabolism
Lipid Metabolism
Peptidoglycan
Citrate Cycle
DNA repair
Fpg, NTH, Xth, Lig, MPG. PCNA
UVRABCD, MFD, mutSL, ExoVII,
DpoIII, recGR, ruvABC, priA
Coenzyme Biosynthesis
Ribosome & tRNA-synthase
rps2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 14e, 15, 16, 17, 18, 19, 20, 21
rpl1, 2, 3, 4e, 5, 6, 7, 9, 10, 10e, 11, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25
27, 28, 29, 30, 31, 32, 33, 34, 35, 36
gltX, gatA, alaS, aspS, glyS, thrS, serS, secS, cysS, metS, fmt
valS, leuS, ileS, lysS, argS, proS, hisS, pheS, tyrS, trpS
Fatty Acid
Riboflavin
FMN
Heme
Folate
Pyruvate
Citrate
Acetyl-CoA
Glycine
Serine
RNA
DNA
PRPP
UMPS
NAD+
NADP+
MTHFS
glyA/thyX

Figure 6. The flagella gene order in *Holosporaceae* (green), *Candidatus Midichloriaceae* (red) and free-living α-proteobacteria representatives (black). Each arrow represents a gene in the cluster. Genome rearrangements are shown as dotted lines between two genes, with the distance between them shown above the lines. Because of the incomplete nature of some genome assemblies, the exact distance between two genes could not be determined. In this case, a minimum distance was estimated as the sum of distances of each gene to the end of the contig it was located in. For the same reason, the orientation of some genes could not be determined (indicated by asterisks below the genes).



Tables

Table 1. Comparison between our reconstruction and Gabaldon et al. 2007.

		This study	Gabaldon et al. 2007
	Families	394	842
	COGs	300	501
Number of genomes	Total	1613	144
	α-proteobacteria/*Rickettsiales*	171/67	11/2
	Eukaryotes	30	16
Nuclear gene families in mitochondria proteome	Human	66.7%	34.1%
	Yeast	69.5%	46.8%
	Arabidopsis	63.3%	42.6%
Gene families with mitochondrial-targeted signal		56.1%	30.7%

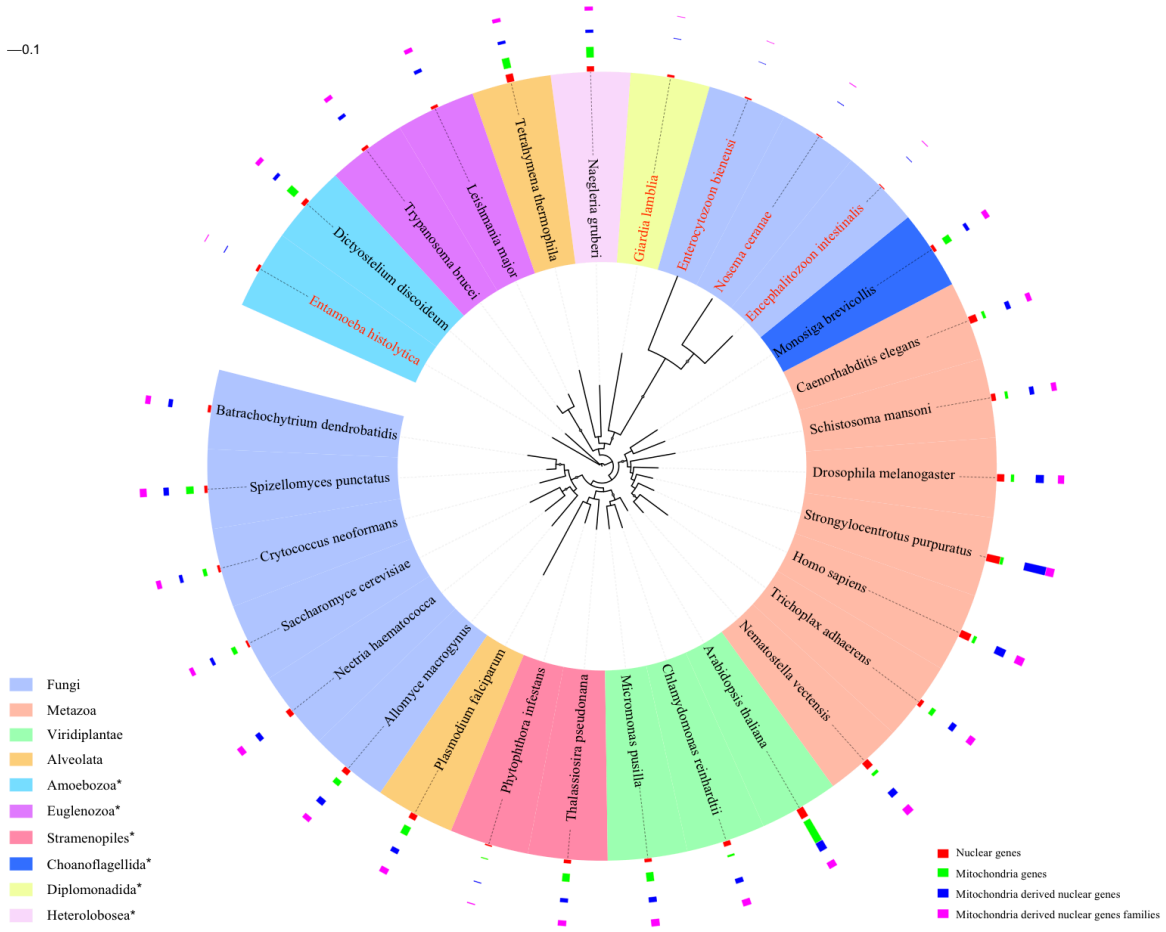
Table 2. List of mitochondria-derived nuclear genes involved in eukaryotic lipid metabolism.

Gene family	COG	Description	Lipid metabolized	Cellular localization	Identified in Gabaldon et al. 2007
Group_236	COG2867	cyclase/dehydrase	Oligoketide	Unknown	N
Group_267	COG1215	ceramide glucosyltransferase	Sphingolipid	ER (MAM)	Y
Group_268	COG1562	squalene/phytoene synthase	Cholesterol	ER	Y
Group_946	COG1154	1-deoxy-D-xylulose-5-phosphate synthase	Terpernoid	ER	Y
Group_1713	COG3660	hypothetical protein	Lipid A	Mitochondria	Y
Group_1971	COG5597	*sqdD* glycosyl transferase	Sulfolipid	Unknown	N
Group_2416	COG1044	*lpxD* UDP-3-O- 3-hydroxymyristoyl glucosamine N-acyltransferase	Lipid A	Mitochondria	N
Group_2710	COG4689	acetoacetate decarboxylase	Ketone body	Unknown	N
Group_3620	COG1043	*lpxA* UDP-N-acetylglucosamine	Lipid A	Mitochondria	N

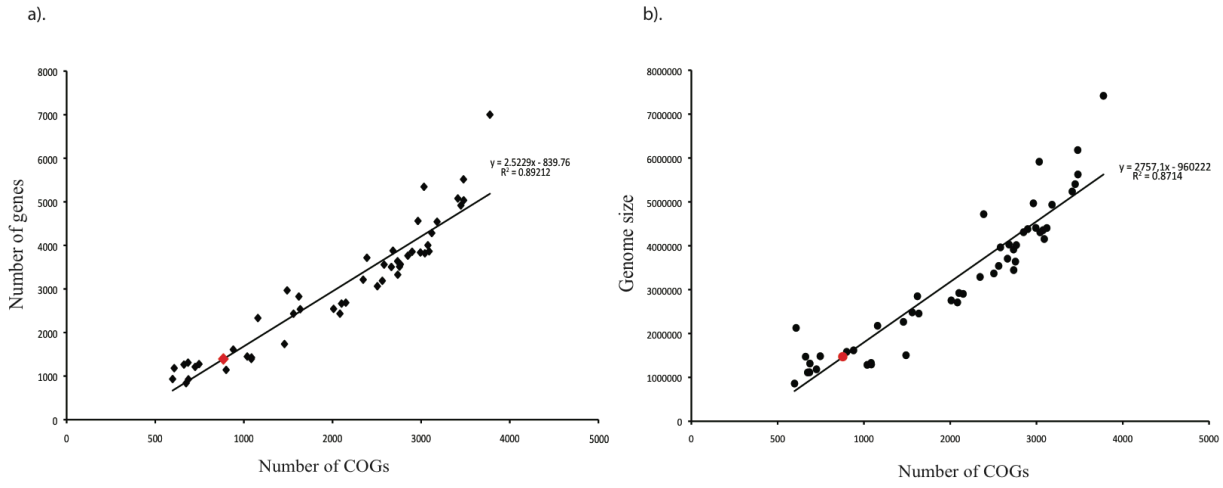
		acyltransferase			
Group_3864	COG3000	Sterol C5 desaturase	Sterol	ER	N
Group_4604	COG1519	*kdtA* 3-deoxy-D-manno-octulosonic-acid transferase	Lipid A	Mitochondria	Y
Group_4794	COG0763	*lpxB* lipid-A-disaccharide synthase	Lipid A	Mitochondria	N

Supplementary Figures

Supplementary Figure 1. Overview of mitochondria-derived nuclear genes. The eukaryotic species tree was reconstructed using a concatenation of 29 ribosomal proteins conserved among all three domains (Harris, et al. 2003). Within each species, from innermost to outermost are the numbers of 1) nuclear genes, 2) mitochondria genes, 3) mitochondria-derived nuclear genes, 4) mitochondria-derived nuclear gene families. The lengths of the bars were scaled for display purposes. Each color in the tree represents a different eukaryotic phylum. Six novel phyla that had not been sampled by previous studies are indicated by asterisks. Lineages highlighted in red represent amitochondriate eukaryotes. Branches with dots represent those with bootstrap support >= 80 (100 replicates).



Supplementary Figure 2. a) Correlation between the number of COGs and the number of genes, and b) between the number of COGs and the genome size of 49 α-proteobacterial representatives. In both graphs, the red dot represents the pre-mitochondria.



Supplementary Figure 3. A maximum-likelihood tree inferred from amino acid sequences of the ATP/ADP translocase in *Chlamydiales* (blue), *Rickettsiales* (orange), *Bacteroidetes* (yellow) and plastids (green). The tree was rooted by hypothetical proteins in *Microsporidia* (*Encephalitozoon intestinalis, Encephalitozoon cuniculi, Enterocytozoon bieneusi* and *Nosema ceranae*). Branches of several lineages are shortened for display purpose. Bootstrap values (out of 100 replicates) are above 80 unless as indicated in the tree.


Chlamydia muridarum tlc 1
Chlamydia trachomatis tlc1
Chlamydia pneumoniae tlc1
Simkania negevensis tlc1
Waddlia chondrophila tlc1
Parachlamydia acanthamoebae tlc1
Candidatus Protochlamydia amoebophila tlc1
Simkania negevensis tlc 2
Cardinium endosymbiont of Encarsia pergandiella tlc 2
Parachlamydia acanthamoebae tlc 2
Chlamydiaceae tlc 1
Parachla/Simkaniaceae/ Waddliaceae tlc1
Parachla/Simkaniaceae/ Waddliaceae tlc2
Thalassiosira pseudonana tlc 1
Chlamydomonas reinhardtii 2
Chlamydomonas reinhardtii tlc 3
Arabidopsis thaliana tlc1
Arabidopsis thaliana tlc2
Micromonas pusilla tlc1
Chlamydomonas reinhardtii tlc 1
Thalassiosira pseudonana tlc 2
Thalassiosira pseudonana tlc 3
Plastids
Candidatus Odyssella thessalonicensis tlc 1
Candidatus Paracaedibacter acanthamoebae tlc 1
Candidatus Paracaedibacter symbiosus tlc 1
Lawsonia intracellularis tlc 1
Holosporaceae
Endosymbiont of Acanthamoeba UWC8 tlc 1
Endosymbiont of Acanthamoeba UWC8 tlc 2
Candidatus Midichloria mitochondrii tlc 1
Midichloriaceae
Trichoplax adhaerens tlc 1
Rickettsia prowazekii tlc 1
Rickettsia conorii tlc 1
Orientia tsutsugamushi tlc 1
Rickettsiaceae tlc 1
Rickettsia prowazekii tlc 2
Rickettsia conorii tlc 2
Orientia tsutsugamushi tlc 2
Rickettsiaceae tlc 2
Candidatus Liberibacter solanacearum tlc 1
Candidatus Liberibacter asiaticus tlc 1
Rhizobiales
Candidatus Midichloria mitochondrii tlc 2
Rickettsia prowazekii tlc 5
Rickettsia conorii tlc 5
Orientia tsutsugamushi tlc 5
Rickettsiaceae tlc 5
Rickettsia prowazekii tlc 4
Rickettsia conorii tlc 4
Orientia tsutsugamushi tlc 4
Rickettsiaceae tlc 4
Rickettsia endosymbiont of Ixodes scapularis tlc 1
Rickettsia endosymbiont of Ixodes scapularis tlc 2
Rickettsia conorii tlc 3
Rickettsia prowazekii tlc 3
Cardinium endosymbiont of Encarsia pergandiella tlc 4
Orientia tsutsugamushi tlc 3
Rickettsiaceae tlc 3
Waddlia chondrophila tlc 5
NHP bacterium tlc 1
Candidatus Caedibacter acanthamoebae tlc 1
Holosporaceae
Waddlia chondrophila tlc 3
Candidatus Protochlamydia amoebophila tlc 3
Parachlamydia acanthamoebae tlc 3
Parachla/Simkaniaceae/ Waddliaceae tlc3
Candidatus Protochlamydia amoebophila tlc 2
Chlamydia muridarum tlc 2
Chlamydia trachomatis tlc 2
Chlamydia pneumoniae tlc 2
Waddlia chondrophila tlc 2
Chlamydiaceae tlc 2
Candidatus Protochlamydia amoebophila tlc 4
Waddlia chondrophila tlc 4
Parachlamydia acanthamoebae tlc 4
Simkania negevensis tlc 3
Parachla/Simkaniaceae/ Waddliaceae tlc4
Candidatus Caedibacter acanthamoebae tlc 2
Candidatus Amoebophilus asiaticus tlc 2
Cardinium endosymbiont of Encarsia pergandiella tlc 3
Cardinium endosymbiont of Encarsia pergandiella tlc 1
Bacteroidetes
Enterocytozoon bieneusi tlc 2
Encephalitozoon intestinalis tlc 1
Encephalitozoon cuniculi tlc 1
Nosema ceranae tlc 1
Enterocytozoon bieneusi tlc 1
Nosema ceranae tlc 2
Nosema ceranae tlc 4
Encephalitozoon cuniculi tlc 2
Encephalitozoon intestinalis tlc 2
Encephalitozoon intestinalis tlc 3
Encephalitozoon cuniculi tlc 3
Nosema ceranae tlc 3
Encephalitozoon cuniculi tlc 4
Encephalitozoon intestinalis tlc 4

Chapter 4. Genomic insights into an obligate epibiotic bacterial predator: *Micavibrio aeruginosavorus* ARL-13[1]

[1]Formatted as a co-authored manuscript and published as:

Wang Z, Kadouri DE, Wu M. BMC Genomics 2011, 12:453. doi:10.1186/1471-2164-12-453

Referenced supplementary material is available online at:

http://www.biomedcentral.com/1471-2164/12/453

Abstract

Background

Although bacterial predators play important roles in the dynamics of natural microbial communities, little is known about the molecular mechanism of bacterial predation and the evolution of diverse predatory lifestyles.

Results

We determined the complete genome sequence of *Micavibrio aeruginosavorus* ARL-13, an obligate bacterial predator that feeds by "leeching" externally to its prey. Despite being an obligate predator depending on prey for replication, *M. aeruginosavorus* encodes almost all major metabolic pathways. However, our genome analysis suggests that there are multiple amino acids that it can neither make nor import directly from the environment, thus providing a simple explanation for its strict dependence on prey. Remarkably, despite apparent genome reduction, there is a massive expansion of genomic islands of foreign origin. At least nine genomic islands encode many genes that are likely important for *Micavibrio*-prey interaction such as hemolysin-related proteins. RNA-Seq analysis shows substantial transcriptome differences between the attack phase, when *M. aeruginosavorus* seeks its prey, and the attachment phase, when it feeds and multiplies. Housekeeping genes as well as genes involved in protein secretion were all dramatically up-regulated in the attachment phase. In contrast, genes involved in chemotaxis and flagellum biosynthesis were highly expressed in the attack phase but were shut down in the attachment phase. Our transcriptomic analysis identified additional genes likely important in *Micavibrio* predation, including porins, pilins and many hypothetical genes.

Conclusions

The findings from our phylogenomic and transcriptomic analyses shed new light on the biology and evolution of the epibiotic predatory lifestyle of *M. aeruginosavorus.* The analysis reported here and the availability of the complete genome sequence should catalyze future studies of this organism.

Keywords

Bacterial predation, Predator-prey interaction, Integrative and conjugative elements (ICEs), Hemolysin-related protein, Quorum sensing, RNA-Seq

Background

Predatory bacteria are a diverse group of bacteria that attack and feed on other bacteria. They live in various habitats and likely play an important role in microbial ecosystems (Casida 1980; Germida and Casida 1983; Chen, et al. 2011). Predation probably has originated multiple times in Bacteria, as examples of predators have been found in dispersed major lineages including *Proteobacteria*, *Chloroflexi*, *Cytophagaceae*, and Gram-positive bacteria (Jurkevitch 2007). Bacterial predators prey using a number of strategies. For example, *Myxobacteria* are facultative predators. They attack as a "wolf pack" and feed on, among other substrates, various live and dead bacteria. On the other hand, *Bdellovibrio* and like organisms (BALOs) are obligate predatory bacteria — they can only survive by preying on other bacteria (Jurkevitch and Davidov 2007). Unlike *Myxobacteria*, which use excreted hydrolytic enzymes to degrade prey cells, obligate predation requires close and irreversible contact between the predator and the prey. *Bdellovibrio* invade the periplasmic space of their prey, where they replicate at the expense of the prey's cellular content and eventually lyse the cell. *Micavibrio*, on the other hand, feed by "leeching" externally to the surface of the prey cell and therefore has an epibiotic lifestyle (Lambina, et al. 1983; Davidov, et al. 2006; Kadouri, et al. 2007; Dashiff, et al. 2010).

First isolated in 1983 from wastewater, *Micavibrio aeruginosavorus* is Gram-negative, relatively small in size (0.5 to 1.5 μm long), rod shaped, curved and has a single polar flagellum (Lambina, et al. 1983). Like BALOs, *Micavibrio spp.* are characterized by an obligatory parasitic life cycle. *Micavibrio's* life cycle is believed to consist of an attack phase, in which motile *Micavibrio* seek their prey, and an attachment phase, in which *Micavibrio* attach irreversibly to the cell surfaces of prey bacteria. At this point the attached *Micavibrio* feed on their prey and divide by binary fission, leading to the death of the infected prey cells (Lambina, et al. 1982; Lambina, et al.

1983; Afinogenova, et al. 1987; Davidov, et al. 2006). *Micavibrio* usually exhibit a high degree of prey specificity. For example, *M. aeruginosavorus* was initially reported to prey only on *Pseudomonas aeruginosa*, *Burkholderia cepacia* and *Klebsiella pneumoniae* (Lambina, et al. 1983; Kadouri, et al. 2007). However a breach in prey specificity was recently demonstrated and *M. aeruginosavorus* was found to be able to prey on many other bacterial species including *Escherichia coli* (Dashiff, et al. 2010).

Myxobacteria and *Bdellovibrio*, both belonging to the δ-proteobacteria, have been extensively studied (Berleman and Kirby 2009; Sockett 2009). Members from both groups (*M. xanthus* DK1622 and *B. bacteriovorus* HD100) have recently been sequenced (Rendulic, et al. 2004; Goldman, et al. 2006). In comparison, *Micavibrio*, members of the α-proteobacteria, have received much less attention, at least partly due to the difficulty to obtain axenic culture and partly due to the lack of good genetic tools to study them. In order to gain greater insights into its predatory lifestyle and to further understand the evolution of bacterial predation in general, we sequenced one of the better studied strains, *Micavibrio aeruginosavorus* ARL-13 (Davidov, et al. 2006; Kadouri, et al. 2007; Dashiff, et al. 2010) and characterized its transcriptome during the attachment and attack stages of its growth cycle.

Results and Discussion

Genome summary

The complete genome of *Micavibrio aeruginosavorus* ARL-13 consists of 2,481,983 base pairs on a single circular molecule with a G+C content of 54.7%. Major features of the genome are summarized in Table 1 and Figure 1. The genome exhibits two clear GC skew transitions that likely correspond to the DNA replication origin and terminus (Figure 1). 90.3% of the genome is

predicted to code for 2434 open reading frames (ORFs), 40 tRNA genes and one rRNA operon. Only 50.5% of the predicted ORFs can be assigned to a putative function. No extragenomic DNA molecules (plasmid or phage) were identified from the genome sequence assembly. CRISPRs (**C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats) function as the immune system of bacteria and archaea that defends against exogenous DNA such as phages and plasmids (Horvath and Barrangou 2010). Accordingly, no CRISPRs elements were identified from the genome.

Repetitive DNAs facilitate genome arrangement and increase the genome plasticity through homologous recombination. Strikingly, only 0.10% of the *M. aeruginosavorus* genome is repetitive (at least 50 bp with at least 97% identity; in comparison, 2.7% of *E. coli* genome contains repeats). The only large repeat (>100bp) that can be identified from the genome is a 1200 bp fragment encoding the elongation factor Tu gene, whose duplication is known to be widespread among proteobacteria (Lathe and Bork 2001). The genome is completely devoid of mobile genetic elements including transposons, retrotransposons and insertion sequences. The paucity of repetitive DNA has been attributed to extensive genome streamlining (Andersson, et al. 1998). Observations of genomes with such an infrequent occurrence of repeats have been limited to obligate intracellular bacteria (e.g., *Buchnera*, *Rickettsia* and *Chlamydiales*) and the free-living bacteria *Prochlorococcus* and *Pelagibacter* that have gone through extensive genome reduction (Andersson, et al. 1998; Stephens, et al. 1998; Shigenobu, et al. 2000; Rocap, et al. 2003; Giovannoni, et al. 2005). *Micavibrio*'s genome is moderate in size. At 2.4 Mbp, it is almost twice as large as most obligate intracellular α-proteobacteria, but is still substantially smaller than most free-living α-proteobacteria, and about 35% smaller than *B. bacteriovorus* HD100 (3.7 Mbp) (Rendulic, et al. 2004). *M. aeruginosavorus'* genome does not have the

extreme GC% bias typical of intracellular bacteria and is almost completely devoid of pseudogenes.

Phylogeny and taxonomy

Micavibrio spp. have many morphological and physiological features resembling those of the *Bdellovibrio spp.* As a result, historically, *Micavibrio spp.* have been affiliated with *Bdellovibrio* and classified as δ-proteobacteria (Garrity, et al. 2004). However, recent studies based on the 16s rRNA and several protein-coding genes have placed *Micavibrio* as a deep branch lineage within the α-proteobacteria (Davidov, et al. 2006), which is strongly supported by our genome-level phylogenetic analysis using 31 housekeeping genes (Figure 2). Its closest relative with a sequenced genome is "*Candidatus Puniceispirillum marinum"*, a member of the ubiquitous marine bacterioplankton SAR116 group (Oh, et al. 2010). Together, they form a sister clade to the *Rhodospirillales* order that is otherwise distinct from all the major α-proteobacterial groups that are currently recognized. Based on our own and previous phylogenetic analyses, we recommend that the taxonomy of *Micavibrio* to be revised.

General metabolic features

Although an obligate predator depending on prey for cell replication, *M. aeruginosavorus* has a free-living attack phase during which it swims around and seeks out the prey. Analysis of the genome shows that it has many features of a free-living bacterium (Additional file 1). For example, it has an elaborate suite of genes involved in cell wall and lipopolysaccharide (LPS) biosynthesis; it is predicted to cover all major metabolic pathways, including glycolysis, the tricarboxylic acid (TCA) cycle, the electron transport and respiration systems and ATP synthase, indicating that it is fully capable of generating ATP on its own by converting carbohydrate, fats and proteins into carbon dioxide and water. It also possesses a complete pentose phosphate

pathway and a full set of genes for nucleotide metabolism, allowing it to synthesize nucleotides from scratch. Not surprisingly, it does not encode any known nucleotide transporters. It has a slightly reduced set of 43 genes devoted to biosynthesis of cofactor, prosthetic groups and carriers. Obligate intracellular bacteria such as *Buchnera* depend on their hosts for most of their nutrients, and as a result of the reduced selection pressure, they have lost a lot of biosynthetic genes (Shigenobu, et al. 2000). The gene loss in *M. aeruginosavorus* is modest in comparison, suggesting that there is considerable selective pressure acting on the remaining genes. This is consistent with the finding that there are rarely any pseudogenes or signs of active gene degradation in the genome.

Amino acid biosynthesis and transport

Since *M. aeruginosavorus* is an obligate predator and has not been cultured axenically, it is of particular interest to use the genome sequence to understand its nutritional needs. Analysis of the genome sequence revealed that *M. aeruginosavorus* encodes genes to synthesize 13 amino acids needed for protein synthesis. However, it is missing almost the entire biosynthesis pathways for the other 7 amino acids: Alanine, Arginine, Histidine, Isoleucine, Methionine, Tryptophan and Valine, suggesting that it can not synthesize these amino acids either de novo or from metabolic intermediates, and has to obtain them directly from external sources. Strikingly, the genome is completely devoid of any known transporters for amino acids, peptides and amines, although it contains 82 ORFs predicted to transport ions, carbohydrates, organic alcohols and acids and other unknown substrates.

Our genome analysis suggested that *M. aeruginosavorus* is deficient in amino acid biosynthesis and uptake from the environment, which at least partially explains why *M. aeruginosavorus* could not be cultured in nutrient rich media (Lambina, et al. 1983; Davidov, et al. 2006) (Daniel

Kadouri, unpublished data). It would be extremely difficult for *Micavibrio* to revert to a lifestyle independent of prey, as it would entail the acquisition of many eliminated genes including those involved in amino acid metabolism. This could explain the failure to isolate prey-independent variants of *Micavibrio* using rich media as described for *Bdellovibrio* (Seidler and Starr 1969; Ishiguro 1974) (Daniel Kadouri, unpublished data). In contrast, although *B. bacteriovorus* is capable of synthesizing only 11 amino acids (Rendulic, et al. 2004), it has a large repertoire of 113 transporters for transporting amino acids, peptides or amines. Therefore, *Bdellovibrio* is capable of importing amino acids that it cannot make on its own from the environment. Accordingly, spontaneous mutants of *Bdellovibrio* that grow in rich media have been isolated at a frequency of 10^{-6} to 10^{-7} and higher (Seidler and Starr 1969; Dashiff and Kadouri 2009).

Among all bacterial and archaeal species sequenced to date, only a few species such as *Buchnera spp.* and *Nanoarchaeum equitans* encode no known amino acid transporters in their genomes. *Buchnera* are bacterial endosymbionts engaged in a classical example of metabolic symbiosis with their host aphids: *Buchnera* supply aphids with essential amino acids and in return, aphids provide complementary non-essential amino acids to the bacteria. The shuttling of the amino acids between the host and the endosymbiont is most likely carried out by transporters encoded by the host genome but not the bacterial genome itself (Shigenobu, et al. 2000; Wilson, et al. 2010). *Nanoarchaeum equitans* represents a more interesting analogy to *Micavibrio spp.* It is an obligate epibiotic parasite that lives on another archaeon *Ignicoccus*. It attaches to the surface of the host cell and presumably acquires its nutrients from the host cell because its tiny genome of 0.5 Mbp does not encode genes for biosynthesis of amino acids, nucleotides or cofactors, nor does it encode transporters for these substrates that allow direct import from the environment (Waters, et al. 2003). Consequently, *Nanoarchaeum* must stay in direct contact with the host organism to survive.

Recently, it has been shown that bacteria can exchange cellular constituents (small molecules, proteins and DNAs) through intercellular nanotubes that connect neighboring cells, even between evolutionarily distant species (Dubey and Ben-Yehuda 2011). It remains unclear how epibiotic parasites and predators extract nutrients from the host or prey, however. For *Nanoarchaeum equitans*, electron microscopy showed a close attachment of the parasite to the surface of the host, although no fixed structure was observed (Huber, et al. 2002). In the case of the bacterial predators *Vampirococcus* and *Ensifer adhearens*, they adhere to the exterior of the prey and appear to attack via a specialized cytoplasmic bridge that is clearly visible as electron-dense materials under the electron microscope (Casida 1982; Jurkevitch and Davidov 2007). The outer membrane of the predator is breached where the dense material appears. Presumably, nutrients can be imported into predators through this junction. It is possible that *Micavibrio* use a similar mechanism to acquire substrates from their prey, as close attachment of *Micavibrio spp.* to prey cells has been shown for strains ARL-13, ARL-14 and EPB previously (Lambina, et al. 1982; Lambina, et al. 1983; Davidov, et al. 2006; Dashiff, et al. 2010).

Hemolysin-related proteins

Micavibrio grow at the expense of the prey eventually leading to its death. Therefore, it is interesting that *M. aeruginosavorus* encodes six hemolysin-related proteins that belong to the RTX (repeats in the toxin) toxin family, as they all bear the calcium-binding, tandem-repeated GGXGXD signature motif in their sequences (Table 2). RTX toxins are produced by a broad range of bacteria and represent a diverse group of hemolysins, cytolysins, proteases and bactericides. They bind to the host cell membrane and play important roles in bacteria-host interactions (Lally, et al. 1999). Functions of many RTX toxins have been well studied, among which the alpha-hemolysin from *E. coli* has been best characterized. After secretion, alpha-

hemolysin inserts itself into the host cell membrane, forms a transmembrane pore and lyses the cell (Bhakdi, et al. 1986). It has been suggested that bacteria may use hemolysin to obtain nutrients from the host cells (e.g., irons released from lysed red blood cells) (Litwin and Calderwood 1993).

The hemolysin-related proteins encoded in the *M. aeruginosavorus* genome vary greatly in length and structural features (Table 2). Further examination of their sequences suggests that they might play important roles in prey recognition and adhesion as well. In addition to the glycine-rich tandem-repeats, two proteins also contain motifs known to mediate cell adhesion and recognition. For instance, GMV2456 contains a bacterial lectin-like domain. Numerous bacterial species produce surface lectins, which are calcium-dependent carbohydrate binding modules typically associated with pili. It is well known that bacterial lectins mediate cell-cell recognition and play key roles in infection by promoting bacterial adherence to the host cells (Sharon and Lis 1989). An early study demonstrated that carbohydrate receptors are involved in *Micavibrio*-prey interaction (Chemeris and Afinogennova 1986), although a recent study suggested this needs to be further investigated (Dashiff, et al. 2011). Cell adhesion can be boosted further with two Von Willebrand factor (VWF) type A domains identified in GMV0107. VWF domain mediates cell-cell adhesion via metal ion-dependent adhesion sites (Ruggeri and Ware 1993). It was originally discovered in extracellular eukaryotic proteins but recently was found to be widespread in bacteria as well.

Notably, hemolysin-related protein is one of few protein families that have been expanded in the *Micavibrio* genome. Phylogenetic analysis indicated that the expansion is not a result of recent gene duplications. In light of the strong genome streamlining in *Micavibrio*, we argue that hemolysin-related proteins play an important role in predation in order for the family to expand

and to be maintained in the genome. This is supported by our transcriptomic analysis showing five of the six hemolysin-related genes were actively expressed in either the attack, the attachment, or both stages (Table 2). It is possible that once *M. aeruginosavorus* attaches to a prey cell, it releases hemolysins into the cell junction, which can then insert themselves into the cell membrane of the prey cell, form pores and open up channels for substrates trafficking. The finding that *Bdellovibrio* insert their own outer membrane pore proteins into the prey cell membrane supports this hypothesis (Tudor and Karp 1994; Beck, et al. 2005).

Secretion system and degradative hydrolytic enzymes

The genome of *M. aeruginosavorus* contains a complete type I and a functional type II secretion systems for protein secretion. However, there is no evidence for the presence of type III or IV secretion system. Type I secretion system transports various substances like RTX-toxins, proteases, lipases, and S-layer proteins to the extracellular space, many of which are important in bacteria pathogenesis. The six hemolysin-related genes in *M. aeruginosavorus* genome all possess type I secretion signals and therefore are predicted to be extracellularly translocated by the type I secretion pathway. In *E. coli* and other bacteria, the genes encoding alpha-hemolysin (*hlyA*) and type I secretion system components (*hlyB* and *hlyD*) are transcribed as one operon (Frey 2006). Interestingly, GMV0107, the largest hemolysin-protein in *M. aeruginosavorus* genome with 2892 amino acids, is located immediately upstream of a cluster of genes encoding type I secretion system components *TolC* (GMV0108), *hlyB* (GMV0110) and *hlyD* (GMV0111). It has been suggested that this arrangement allows the timely export of toxins without damage to the membrane of the bacteria producing them (Frey 2006).

Type II is responsible for the extracellular secretion of toxins and hydrolytic enzymes, many of which contribute to pathogenesis in both plants and animals. Proteins secreted through the type II

system depend on the Sec or twin-arginine translocation (TAT) system for initial transport into the periplasm. The genome encodes a complete TAT secretion system (*TatABCD*), and a complete Sec secretion system (*SecABDEFGY, YajC, FtsY, SRP*). The type II secretion apparatus is composed of at least 12 different gene products that are thought to form a multiprotein complex. Some components of the type II secretion system, including *GspCGHK*, are absent in the genome annotation. It is possible that they can be substituted by type IV pilus proteins encoded in the genome, as they are homologous and functionally equivalent (Sandkvist 2001). Based on the presence of the complete TAT and Sec transport systems, we think the type II secretion system is likely to be functional.

M. aeruginosavorus encodes an impressive arsenal of hydrolytic enzymes. A large fraction of the genome (4.3%) was predicted to encode 49 proteases and peptidases, 12 lipases, 2 DNAses, 4 RNAases and 37 other hydrolases (Additional file 2). Although hydrolytic enzymes are required for the routine maintenance of cellular structures, we expect a sizeable portion of *Micavibrio*'s hydrolytic enzymes to be devoted to digest the prey cell macromolecules. For example, it has been demonstrated that a lytic proteinase of around 39 kDa (+/- 1.5 KDa) isolated from *Micavibrio admirandus* is able to lyse *E. coli* cells (Severin, et al. 1987). *M. aeruginosavorus* encodes one proteinase in this molecular weight range — GMV0053 is predicted to encode a 40 kDa peptidase M23 family protein. Although their roles in *Micavibrio* predation remain to be elucidated, with the gene sequences now it is possible to have the hydrolases heterologously expressed and experimentally characterized, as they may be valuable for the development of enzyme-based anti-microbial agents.

Flagellum and pili

Micavibrio spp. are motile and possess a single, sheathless, polar flagellum. Motility gives *Micavibrio* the advantage of being able to actively search for prey. In addition, *M. aeruginosavorus* is capable of biofilm predation (Kadouri, et al. 2007; Dashiff, et al. 2010). Flagellum might provide the necessary force for the predator to penetrate and attack biofilms, as demonstrated in *Bdellovibrio* (Medina, et al. 2008). As expected, *M. aeruginosavorus* encodes a plethora of genes related to flagellum biosynthesis and chemotaxis (Additional file 3). The genome also possesses multiple dispersed *pil* genes encoding type IV pili, including three operons encoding eight proteins with prepilin-type cleavage/methylation signal at the N-terminus. Proteins with prepilin-like leader sequences are typically involved in type IV pili biogenesis or type II secretion system (Mattick 2002). Type IV pili in bacteria are in general involved in adherence and invasion of host cells (Mattick 2002) and is believed to play a role in *B. bacteriovorus* predation (Evans, et al. 2007; Mahmoud and Koval 2010). Although *Micavibrio* are epibiotic predators and do not invade prey cells, type IV pili can play an important role in predation by mediating cell adhesion. This is supported by our transcriptomic data showing that four pili-related genes were highly expressed in the attack or attachment phase (GMV0530, 0902, 0903,1530, see Additional file 4). Notably, gene GMV0530 encoding a *flp/Fap* pilin component family protein was one of the most actively transcribed genes in the attack phase.

Signal transduction and quorum-sensing

Unlike other obligate parasitic bacteria such as *Mycoplasma* that live exclusively inside the prey cell, *M. aeruginosavorus* is an epibiotic predator constantly exposed to the environment. Moreover, in the attack phase it has to actively search for its next prey. *M. aeruginosavorus* is poised to respond to diverse environmental cues through a suite of signal transduction pathways and processes. For example, the organism has at least 41 genes of two-component signal transduction systems, which is remarkable given its genome size. Intriguingly, the *M.*

aeruginosavorus genome encodes at least four genes involved in quorum-sensing: one autoinducer synthase (*LuxI*, GMV1999), two autoinducer binding proteins (*LuxR*, GMV0289 and 0290) and one regulator protein (*LuxO*, GMV1999). Quorum sensing is important for group predation, which requires a quorum of predators and coordinated release of hydrolytic enzymes to degrade the prey. "Wolf pack" predation has been observed in *Myxobacteria* and *Lysobacter* but not in *Micavibrio* or *Bdellovibrio*, at least under laboratory conditions. *Micavibrio* is known to attack the prey on an one-to-one basis (Lambina, et al. 1982; Lambina, et al. 1983; Davidov, et al. 2006), so it is not clear what the biological role of the quorum-sensing genes is. One possibility is that *Micavibrio* can use quorum-sensing to detect their own density and avoid having two or more predators attacking the same prey cell. Multi-predation on a single cell can spell disaster because one prey cell usually does not have enough resource to support the replication of multiple predators. It is also possible that *Micavibrio* can use quorum-sensing to detect the density of the prey population when predating on biofilm. Our RNA-Seq data show that *LuxO* was expressed at low level during the attack phase but not in the attachment phase, *LuxR* was expressed at low level in both phases while *LuxI* was not expressed in either phase (Additional file 4). It will be extremely interesting to elucidate the biological function of the quorum-sensing genes in *Micavibrio*, to investigate whether *Micavibrio* are capable of quorum-sensing, and if so, to deduce its role in the evolution of predation.

Lateral gene transfers

Since *M. aeruginosavorus* preys on other Gram-negative bacteria, it has the potential to take up prey's DNAs during the feeding process and incorporate them into its own genome. Using BLAST search, we did not find any examples of highly similar stretches of DNA (>100bp and 97% identity) shared between *M. aeruginosavorus* and *P. aeruginosa,* the strain that has been used in the laboratory to maintain *Micavibrio*. Similarly, there is no evidence of recent lateral

gene transfer from prey into *B. bacteriovorus* (Rendulic, et al. 2004). Foreign DNA usually has a nucleotide composition distinct from that of the native DNA and therefore can be detected using chi-square test of base homogeneity, although sequence bias can arise from other sources as well. Our tri-nucleotide chi-square analysis identified numerous regions deviating significantly from the rest of the genome (Figure 1). Among them are operons encoding the rRNA genes and ribosomal proteins, where sequence biases are most likely due to either secondary structure constraint (rRNAs) or biased codon usage (ribosomal proteins). However, we also identify nine genomic islands of possible foreign origins (Additional file 5). Their sizes range from 11.4 Kbp to 27.4 Kbp.

Features found on these islands suggest that they belong to a group of integrative and conjugative elements (ICEs). Four out of nine islands are flanked by tRNA genes on one side and seven out of nine contain the signature integrase related to lambda phages (Additional file 5). tRNA genes are known hotspots for ICE insertion (Burrus and Waldor 2004; Wozniak and Waldor 2010). Some also contain helicases, DNA primase, resolvase and reverse transcriptase, mobilization gene (e.g., *mobA/L*) and addiction modules important for ICE maintenance. ICEs normally replicate as part of the host chromosome. But under certain conditions, they can excise from the chromosome, circularize and then transfer to new hosts by conjugation. ICEs therefore combine features of phages and plasmids and can mediate lateral gene flow between distantly related bacterial species (Burrus and Waldor 2004; Wozniak and Waldor 2010). It is not immediately clear whether any of the *Micavibrio* ICEs are still functional, i.e., whether they can move within the genome or to other bacterial species. Our transcriptomic data show that at least five integrases were actively expressed during the attachment or attack phase, suggesting that the ICEs can be active.

ICEs allow bacteria to rapidly adapt to new environmental niches (Burrus and Waldor 2004) and often carry genes such as antibiotic resistance genes and virulence genes (e.g., adhesins, toxins, invasins on the pathogenicity island) (Schmidt and Hensel 2004; Gal-Mor and Finlay 2006) that confer selective advantages to the cell. *M. aeruginosavorus* strain ARL-13 was originally isolated from sewage water. Not surprisingly, heavy metal (copper, cobalt, zinc, cadmium) resistance genes are found within the *M. aeruginosavorus* genomic islands. Interestingly, three hemolysin genes are also located on the ICEs, in addition to a few genes encoding peptidoglycan binding proteins (Additional file 5).

Since ICEs can move between distantly related species by conjugation, it is natural to ask where did the ICEs in *Micavibrio* come from? ICEs have been found in many bacteria including *Micavibrio*'s prey, *P. aeruginosa*. It is possible, at least in theory, that ICEs are passed from the prey to *Micavibrio* during predation. After all, epibiotic predation and conjugation share an unmistakable common ground — both involve intimate cell-cell contact and interaction. Phylogenetic analysis of the integrase genes does not support prey being the ICE source. Instead, it indicates that *Micavibrio* ICEs are mostly closely related to those of other α-proteobacteria. Therefore, these ICEs either only move among α-proteobacteria, or they were present in the ancestor of *Micavibrio* and have been inherited through vertical descent.

Transcriptome analysis

To identify genes important in the predatory life cycle of *Micavibrio*, we analyzed the transcriptomes of *M. aeruginosavorus* in the attachment and attack phases using RNA sequencing (RNA-Seq). We obtained a total of 8,451,083 reads by Illumina sequencing. 96% of the attack and 60% of the attachment reads were mapped unambiguously to the *M. aeruginosavorus* genome. Of the unmapped reads, the vast majority (92%) were actually the

sequences of the prey *P. aeruginosa*. This shows that the prey cells coexisted with the predator cells in the attachment phase but were nearly absent in the attack phase, indicating our strategy of obtaining *Micavibrio* cells at both stages was working. Although we estimated that more than 90% of ribosomal RNAs had been removed during the mRNA preparation, they still constituted the bulk of our illumina reads, as seen previously (Oliver, et al. 2009).

Approximately 72.6% of the genome (coding and non-coding) is covered by at least one read, suggesting that more than 27.4% of the genome was not transcribed or was transcribed at low levels in either phase. In addition, 91.6% of reads match predicted ORFs, indicating that there was very little background noise due to potential DNA contamination in our mRNA preparation. RNA-Seq has provided reliable quantitative estimates of gene expression in yeast and bacteria (Nagalakshmi, et al. 2008; Oliver, et al. 2009; Yoder-Himes, et al. 2009). To allow for quantitative comparisons between samples, we calculated the gene expression index (GEI) as the mean coverage depth of the gene normalized by the total number of reads mapped to non-rRNA regions of the genome. Additional file 6 shows a tight correlation between GEI and the transcript level determined by real-time quantitative reverse transcription PCR (qRT-PCR, $R^2 = 0.85$), confirming that our RNA-Seq data provide reliable estimates of gene expression. In addition, as we show below, the expression levels of genes within a particular pathway are fairly consistent, indicating that there was little bias in our RNA-Seq library construction. For example, our RNA-Seq data show strong up-regulation of gene expression in all 54 ribosomal proteins encoded in the genome in the attachment phase.

The transcriptome differs substantially between the attack and attachment phases. Overall, 80.0% of genes were transcribed in the attachment phase, but only 33.4% of genes were transcribed in the attack phase. Genes that were up-regulated in the attack phase are flagellar genes, chemotaxis

genes and many hypothetical genes. Genes that were up-regulated in the attachment phase include housekeeping genes involved in DNA replication (e.g., chromosome replication initiation protein, DNA polymerase, DNA topoisomerase, helicase, gyrase), transcription (e.g., RNA polymerase, sigma 70, transcription terminator), translation (e.g., ribosomal proteins, translation initiation and elongation factors), energy production (e.g., TCA cycles, electron transport system, ATP synthase) and cell division (e.g., Fts proteins, cell shape determining factor *MreB*) (Additional file 4). The gene expression pattern is consistent with what we know about the life cycle of *Micavibrio*. During the attack phase, powered by a single polar flagellum attached at one end of the cell, *Micavibrio* seek out their prey. Once attached to the prey, *Micavibrio* lose their motility, start to feed on their prey, grow, and multiply by binary fission (Lambina, et al. 1982; Lambina, et al. 1983; Davidov, et al. 2006). Accordingly, genes involved in chemotaxis and flagella biosynthesis were highly expressed in the attack phase but were shut down in the attachment phase. Genes of the two-component signal transduction system were also up-regulated in the attack phase. On the other hand, genes involved in active cell growth and division were highly expressed in the attachment phase, providing the necessary energy and other resources for the cell to replicate. Our genome-wide expression data is consistent with the fact that *M. aeruginosavorus* is an obligate predator that depends on prey to multiply and lacks the ability to propagate in rich media.

Genes involved in protein secretion were also substantially up-regulated in the attachment phase. For example, our RNA-Seq data reveal a uniform increase of gene expression of the entire Sec secretion system (*SecABDEFGY, YajC, FtsY, SRP*), averaging a 17-fold increase when compared to the attack phase. Similarly, the entire twin-arginine translocation (TAT) system, the type I secretion system, and most of the type II secretion system were also significantly up-regulated. This is in agreement with the idea that while attached to the prey cells, *Micavibrio* actively inject

hydrolytic enzymes and toxins into prey cells for prey degradation and nutrient uptake. The expression levels of hydrolytic enzymes were nearly unchanged (attachment/attack = 1.29). It is possible that hydrolytic enzymes are produced and accumulate in the attack phase, which can then be readily discharged in the next round of attachment phase.

Interestingly, three cold-shock protein genes (GMV0274, 1414, 2249) were highly expressed in the attachment phase but were not transcribed in the attack phase. Cold-shock proteins of *E. coli* act as mRNA chaperons to promote single-strandedness of mRNA molecules at low temperature to facilitate their translation (Jiang, et al. 1997). A recent study in *Bacillus subtilis* demonstrated that cold-shock proteins are also essential for cellular growth and efficient protein synthesis at optimal growth temperature (Graumann, et al. 1997). Since the attachment cells were never exposed to cold shock before they were mixed with RNAlater, we believe the up-regulation of cold-shock protein genes in *M. aeruginosavorus* serves to maximize the translation efficiency (Sommerville 1999). This is consistent with our observation that genes involved in the translation process were all up-regulated in the attachment phase. Intriguingly, although the heat-shock protein sigma 32 was highly expressed in both phases, its expression was further boosted in the attack phase by 12-fold. Heat shock has been shown to induce axenic growth of *B. bacteriovorus* in rich media, possibly by generating or simulating signals normally derived from prey (Gordon, et al. 1993). Sigma 32 is one of the few functionally characterized genes that were up-regulated in *Micavibrio* during the attack phase, suggesting that it might play an important role in the attack phase by promoting the transcriptions of other genes.

The most highly expressed gene (other than the rRNA genes) in the attachment phase is a porin-encoding gene GMV0043. Porins form aqueous channels on the outer membrane of Gram-negative bacterial cells, and control the diffusion of small metabolites like sugars, ions and

amino acids across the outer membrane. GMV0043 was expressed at low level in the attack phase but was dramatically up-regulated in the attachment phase by more than 400-fold. The timing and intensity of the gene expression strongly argue that it plays a critical role in the attachment phase by facilitating the uptake of small metabolites derived from degrading prey cells. Similarly, Lambert et al. have showed that the maltose porin gene in *Bdellovibrio* is highly upregulated during predation, when sugars derived from the prey degradation are available for uptake (Lambert, et al. 2009). Of the five other porin-encoding genes identified in the *Micavibrio* genome, four were actively transcribed in the attachment phase, albeit at subdued levels (GMV0953, 1742, 1033, 0975, see Additional file 4).

Strikingly, most of the highly expressed genes in the attack phase are hypothetical genes. This is in sharp contrast to the gene expression pattern of the attachment phase, where most of the highly expressed genes are well-known housekeeping genes. The fact that the hypothetical genes are highly expressed and the RNA-Seq reads match nicely to the gene models suggest that they are real genes. While uncharacterized, they most likely code for actual proteins that play cryptic but important functions in the unique lifestyle of *Micavibrio*.

Conclusions

The phylogenomic and transcriptomic analyses of *M. aeruginosavorus* revealed many features consistent with what we know about its epibiotic predatory lifestyle. Analysis of the genome has also provided new perspectives on the biology of this species and the evolution of bacterial predation in general. Because of the lack of good genetic tools for *Micavibrio*, their predation has remained molecularly enigmatic. The analysis reported here and the availability of the

complete genome sequence should open up new opportunities and catalyze future studies of this organism.

Competing interests

The authors have declared that no conflicts of interest exist.

Materials and methods

Bacteria culture and genomic DNA preparation

M. aeruginosavorus strain ARL-13 was used in this study (Lambina, et al. 1983; Kadouri, et al. 2007). *M. aeruginosavorus* was maintained as plaques in double-layered diluted nutrient broth (DNB) agar, a 1:10 dilution of nutrient broth amended with 3 mmol l^{-1} $MgCl_2 \cdot 6H_2O$ and 2 mmol l^{-1} $CaCl_2 \cdot 2H_2O$ [pH 7·2] and agar (0·6% agar in the top layer). To initiate a lysate, cocultures were obtained by adding a plug of agar containing *M. aeruginosavorus* plaque to washed overnight grown *P. aeruginosa* PA14 prey cells (1×10^9 CFU ml^{-1}) in DNB and incubated at 30°C on a rotary shaker set at 200 rev min^{-1} until the coculture became clear (stock lysate). To harvest the predators, cocultures were prepared in which 20 ml of washed *P. aeruginosa* PA14 cells were incubated with 20 ml of stock lysate in 200 ml of DNB and incubated for 48 hrs. Thereafter, the cocultures were passed 10 times through a 0.45-μm Millex pore-size filter (Millipore) to remove residual prey and cell debris. The filtered lysate was spun down for 30 min at 15,000xg. The supernatant was removed and the pelleted cells were taken for chromosomal DNA extraction using Puregene-Genomic DNA purification kit (Gentra systems) (Dashiff, et al. 2010).

Genome sequencing and annotation

The genome was sequenced by 3Kbp paired-end 454 pyrosequencing, in the University of Virginia Department of Biology Genome Core Facility, and was assembled using GS De Novo Assembler (Newbler). The initial Newbler assembly contained 21 contigs in one scaffold. The Phred/Phrap/Consed software package was used for quality assessment in genome assembly. PCR and Sanger sequencing was used to close the gaps between contigs to get the complete genome sequence, which was then annotated by the IGS annotation engine . The complete sequence has been assigned GenBank accession number CP002382. DNA repeats of at least 50 bp with at least 97% sequence identity were identified using the program Vmatch (Kurtz 2004).

Genome tree construction

Protein sequences of 31 housekeeping genes *(dnaG, frr, infC, nusA, pgk, pyrG, rplA, rplB, rplC, rplD, rplE, rplF, rplK, rplL, rplM, rplN, rplP, rplS, rplT, rpmA, rpoB, rpsB, rpsC, rpsE, rpsI, rpsJ*, *rpsK, rpsM, rpsS, smpB, tsf)* from genomes of interest were identified, aligned, trimmed and concatenated using the software AMPHORA (Wu and Eisen 2008). The concatenated protein sequence alignment was then used to build a maximum likelihood tree using Phyml (Guindon and Gascuel 2003).

RNA isolation, library construction, and transcriptome sequencing

To isolate RNA from attachment phase *M. aeruginosavorus* cells*,* cocultures were prepared as before using *P. aeruginosa* PA14 as the prey. The cocultures were incubated for 8 hrs to allow attachment of the predator to its prey. Thereafter, the cocultures were collected in a 50 ml tube and a fraction containing mainly prey-attached *M. aeruginosavorus* cells was isolated by low speed centrifugation at 4,000xg for 5 min at room temperature. The pellet was then resuspended in 0.5 ml of RNAlater stabilization solution (Applied Biosystems). For isolating RNA from attack phase *M. aeruginosavorus* cells, the cocultures were incubated for 48 hrs allowing the

killing of the prey cells and growth and enrichment of the predator. The clear culture was collected and passed 5 times through a 0.45-µm Millex pore-size filter to remove any residual prey and *M. aeruginosavorus* cells which are still firmly attached to the prey. The filtered lysate was spun down at 4°C for 30 min at 15,000xg and the pellet containing attack phase *M. aeruginosavorus* was resuspended in RNAlater stabilization solution until RNA extraction.

Total RNA for both attachment and attack samples were isolated from bacteria pellet using RiboPure-Bacteria Kit (Ambion) according to the manufacturer's instructions, with genomic DNA removed using DNase I. RNA was quantified using Quant-iT™ RNA Assay Kit (Invitrogen). 23S and 16S rRNA were removed for mRNA enrichment using MICROBExpress Kit (Ambion). RNA quality analysis using Bioanalyzer (Agilent) indicated that about 90% rRNA was removed. cDNA libraries for Illumina sequencing were then constructed using NEBNext mRNA Sample Prep Master Mix Set 1 (New England Biolabs) following the manufacturer's protocol. Libraries were tagged, amplified by 15 cycles of PCR and sequenced with one lane of Illumina GA IIx 43 cycle single-end sequencing.

RNA-Seq reads mapping and visualization

FASTX-Toolkit was used to split the pooled reads into separate attachment and attack phase categories, and to eliminate the tag barcodes from the reads. We mapped reads from both attachment and attack sample to the *M. aeruginosavorus* genome using Maq , allowing up to 2 mismatches to occur. The gene expression index (GEI) was calculated as the mean coverage depth of the gene, normalized by the total number of reads mapped to non-rRNA regions of the genome. The medium coverage of intergenic regions calculated this way was 0.7. Therefore, based on the RNA-Seq coverage, genes were classified into 4 categories using a schema similar to the one described in (Oliver, et al. 2009): 1) not expressed (coverage < 0.7), 2) low expression

(0.7 <= coverage < 10), 3) medium expression (10 <= coverage < 25), 4) high expression (coverage >= 25). The gene expression levels were plotted and visualized in Artemis (Rutherford, et al. 2000).

Quantitative real-time PCR

Total RNA for attachment phase sample was reverse transcribed to cDNA using *SuperScript*® II Reverse Transcriptase (Invitrogen). The primer premier 5 software was used to design and select optimum primers for an amplification product of about 350bp. The quantitative RT-PCR was performed with Fast SYBR-Green master mix (Applied Biosystems) in 7500/7500 Fast Real-Time PCR system. Three replicates were conducted for each gene and the average Ct value was obtained (the cycle number when the fluorescence is detected above the background level). The relative abundance for each gene was calculated based on the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen 2001).

List of abbreviations

ORF, open reading frame; BALOs, *Bdellovibrio* and like organisms (BALOs); GEI, gene expression index; **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats, CRISPRs; LPS, lipopolysaccharide; TCA, the tricarboxylic acid; RTX, repeats in the toxin; VWF, Von Willebrand factor; TAT, twin-arginine translocation; ICEs, integrative and conjugative elements; RNA-Seq, RNA sequencing; qRT-PCR, real-time quantitative reverse transcription PCR; T4SSs, type IV secretion systems; CFU, colony forming unit.

Authors' contributions

ZW carried out laboratory work of genome and transcriptome sequencing and bioinformatic analysis and contributed to manuscript writing. DK prepared the cell cultures for genome and

transcriptome sequencing and helped to draft the manuscript. MW conceived and designed the experiments, analyzed the data and wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Dr. Michelle Gwinn-Giglio for her help on annotating the genome sequence. Part of this study was supported by a grant to MW by the Thomas F. Jeffress and Kate Miller Jeffress Memorial Trust.

References

FASTX-Toolkit: FASTQ/A short-reads pre-processing tools [Internet]. Available from: http://hannonlab.cshl.edu/fastx_toolkit/index.html

IGS Annotation Engine [Internet]. Available from: http://ae.igs.umaryland.edu/cgi/index.cgi

Maq: Mapping and Assembly with Qualities [Internet]. Available from: http://maq.sourceforge.net/

Afinogenova AV, Markelova N, Lambina VA. 1987. Analysis of the interpopulational interactions in a 2-component bacterial system of *Micavibrio admirandus* – Escherichia coli. Nauchnye Doki Vyss Shkoly Biol Nauki 6:101–104.

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133-140.

Beck S, Schwudke D, Appel B, Linscheid M, Strauch E. 2005. Characterization of outer membrane protein fractions of *Bdellovibrionales*. FEMS Microbiol Lett 243:211-217.

Berleman JE, Kirby JR. 2009. Deciphering the hunting strategy of a bacterial wolfpack. FEMS Microbiol Rev 33:942-957.

Bhakdi S, Mackman N, Nicaud JM, Holland IB. 1986. Escherichia coli hemolysin may damage target cell membranes by generating transmembrane pores. Infect Immun 52:63-69.

Burrus V, Waldor MK. 2004. Shaping bacterial genomes with integrative and conjugative elements. Res Microbiol 155:376-386.

Casida LE. 1980. Bacterial Predators of *Micrococcus luteus* in Soil. Appl Environ Microbiol 39:1035-1041.

Casida LE. 1982. *Ensifer adhaerens* gen. nov., sp. nov.: a bacerial predator of bacteria in soil. Int J Syst Bacteriol 32:339-345.

Chemeris NA, Afinogennova AV. 1986. Role of carbohydrate receptors in the interaction of *Micavibrio admirandus* and host-bacterium. Zentbl. Mikrobiol. 141:557-560.

Chen H, Athar R, Zheng G, Williams HN. 2011. Prey bacteria shape the community structure of their predators. Isme J.

Dashiff A, Junka RA, Libera M, Kadouri DE. 2010. Predation of human pathogens by the predatory bacteria *Micavibrio aeruginosavorus* and *Bdellovibrio bacteriovorus*. J Appl Microbiol 110:431-444.

Dashiff A, Kadouri DE. 2009. A New Method for Isolating Host-Independent Variants of *Bdellovibrio bacteriovorus* Using E. coli Auxotrophs. Open Microbiol J 3:87-91.

Dashiff A, Keeling TG, Kadouri DE. 2011. Inhibition of Predation by *Bdellovibrio bacteriovorus* and *Micavibrio aeruginosavorus* via Host Cell Metabolic Activity in the Presence of Carbohydrates. Appl Environ Microbiol 77:2224-2231.

Davidov Y, Huchon D, Koval SF, Jurkevitch E. 2006. A new alpha-proteobacterial clade of *Bdellovibrio*-like predators: implications for the mitochondrial endosymbiotic theory. Environ Microbiol 8:2179-2188.

Dubey GP, Ben-Yehuda S. 2011. Intercellular nanotubes mediate bacterial communication. Cell 144:590-600.

Evans KJ, Lambert C, Sockett RE. 2007. Predation by *Bdellovibrio bacteriovorus* HD100 requires type IV pili. J Bacteriol 189:4850-4859.

Frey J. 2006. Genetics and phylogeny of RTX cytolysins. In: Alouf JE, Popoff MR, editors. The comprehensive sourcebook of bacterial protein toxins. Amsterdam ; Boston: Elsevier. p. 570-577.

Gal-Mor O, Finlay BB. 2006. Pathogenicity islands: a molecular toolbox for bacterial virulence. Cell Microbiol 8:1707-1719.

Garrity GM, Bell JA, Lilburn TG. 2004. Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology, Second Edition.: Springer-Verlag, New York.

Germida JJ, Casida LE. 1983. *Ensifer adhaerens* Predatory Activity Against Other Bacteria in Soil, as Monitored by Indirect Phage Analysis. Appl Environ Microbiol 45:1380-1388.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, et al. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science 309:1242-1245.

Goldman BS, Nierman WC, Kaiser D, Slater SC, Durkin AS, Eisen JA, Ronning CM, Barbazuk WB, Blanchard M, Field C, et al. 2006. Evolution of sensory complexity recorded in a myxobacterial genome. Proc Natl Acad Sci U S A 103:15200-15205.

Gordon RF, Stein MA, Diedrich DL. 1993. Heat shock-induced axenic growth of *Bdellovibrio bacteriovorus*. J Bacteriol 175:2157-2161.

Graumann P, Wendrich TM, Weber MH, Schroder K, Marahiel MA. 1997. A family of cold shock proteins in *Bacillus subtilis* is essential for cellular growth and for efficient protein synthesis at optimal and low temperatures. Mol Microbiol 25:741-756.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696-704.

Horvath P, Barrangou R. 2010. CRISPR/Cas, the immune system of bacteria and archaea. Science 327:167-170.

Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 417:63-67.

Ishiguro EE. 1974. Minimum nutritional requirements for growth of host-independent derivatives of *Bdellovibrio bacteriovorus* strain 109 Davis. Can J Microbiol 20:263-264.

Jiang W, Hou Y, Inouye M. 1997. CspA, the major cold-shock protein of *Escherichia coli*, is an RNA chaperone. J Biol Chem 272:196-202.

Jurkevitch E. 2007. Predatory prokaryotes : biology, ecology, and evolution. Berlin ; New York: Springer.

Jurkevitch E, Davidov Y. 2007. Phylogenetic Diversity and Evolution of Predatory Prokaryotes. Microbiology Monographs 4:11-56.

Kadouri D, Venzon NC, O'Toole GA. 2007. Vulnerability of pathogenic biofilms to *Micavibrio aeruginosavorus*. Appl Environ Microbiol 73:605-614.

Kurtz S. 2004. The Vmatch large scale sequence analysis software. In.

Lally ET, Hill RB, Kieba IR, Korostoff J. 1999. The interaction between RTX toxins and target cells. Trends Microbiol 7:356-361.

Lambert C, Hobley L, Chang CY, Fenton A, Capeness M, Sockett L. 2009. A predatory patchwork: membrane and surface structures of *Bdellovibrio bacteriovorus*. Adv Microb Physiol 54:313-361.

Lambina VA, Afinogenova AV, Romai Penabad S, Konovalova SM, Pushkareva AP. 1982. *Micavibrio admirandus* gen. et sp. nov. Mikrobiologiia 51:114–117.

Lambina VA, Afinogenova AV, Romay Penobad Z, Konovalova SM, Andreev LV. 1983. New species of exoparasitic bacteria of the genus *Micavibrio* infecting gram-positive bacteria. Mikrobiologiia 52:777–780.

Lathe WC, 3rd, Bork P. 2001. Evolution of tuf genes: ancient duplication, differential loss and gene conversion. FEBS Lett 502:113-116.

Litwin CM, Calderwood SB. 1993. Role of iron in regulation of virulence genes. Clin Microbiol Rev 6:137-149.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25:402-408.

Mahmoud KK, Koval SF. 2010. Characterization of type IV pili in the life cycle of the predator bacterium *Bdellovibrio*. Microbiology 156:1040-1051.

Mattick JS. 2002. Type IV pili and twitching motility. Annu Rev Microbiol 56:289-314.

Medina AA, Shanks RM, Kadouri DE. 2008. Development of a novel system for isolating genes involved in predator-prey interactions using host independent derivatives of *Bdellovibrio bacteriovorus* 109J. BMC Microbiol 8:33.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344-1349.

Oh HM, Kwon KK, Kang I, Kang SG, Lee JH, Kim SJ, Cho JC. 2010. Complete genome sequence of "*Candidatus Puniceispirillum marinum*" IMCC1322, a representative of the SAR116 clade in the Alphaproteobacteria. J Bacteriol 192:3240-3241.

Oliver HF, Orsi RH, Ponnala L, Keich U, Wang W, Sun Q, Cartinhour SW, Filiatrault MJ, Wiedmann M, Boor KJ. 2009. Deep RNA sequencing of L. monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. BMC Genomics 10:641.

Rendulic S, Jagtap P, Rosinus A, Eppinger M, Baar C, Lanz C, Keller H, Lambert C, Evans KJ, Goesmann A, et al. 2004. A predator unmasked: life cycle of *Bdellovibrio bacteriovorus* from a genomic perspective. Science 303:689-692.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424:1042-1047.

Ruggeri ZM, Ware J. 1993. von Willebrand factor. Faseb J 7:308-316.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944-945.

Sandkvist M. 2001. Biology of type II secretion. Mol Microbiol 40:271-283.

Schmidt H, Hensel M. 2004. Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev 17:14-56.

Seidler RJ, Starr MP. 1969. Isolation and characterization of host-independent Bdellovibrios. J Bacteriol 100:769-785.

Severin AI, Markelova NY, Afinogenova AV, Kulaev IS. 1987. Isolation and some physicochemical properties of lytic proteinase of the parasitic bacterium *Micavibrio admirandus*. Biokhimiya 52:1594-1599.

Sharon N, Lis H. 1989. Lectins as cell recognition molecules. Science 246:227-234.

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature 407:81-86.

Sockett RE. 2009. Predatory lifestyle of *Bdellovibrio bacteriovorus*. Annu Rev Microbiol 63:523-539.

Sommerville J. 1999. Activities of cold-shock domain proteins in translation control. Bioessays 21:319-325.

Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, et al. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282:754-759.

Tudor JJ, Karp MA. 1994. Translocation of an outer membrane protein into prey cytoplasmic membranes by bdellovibrios. J Bacteriol 176:948-952.

Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M, Beeson KY, Bibbs L, Bolanos R, Keller M, et al. 2003. The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci U S A 100:12984-12988.

Wilson AC, Ashton PD, Calevro F, Charles H, Colella S, Febvay G, Jander G, Kushlan PF, Macdonald SJ, Schwartz JF, et al. 2010. Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*. Insect Mol Biol 19 Suppl 2:249-258.

Wozniak RA, Waldor MK. 2010. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol 8:552-563.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9:R151.

Yoder-Himes DR, Chain PS, Zhu Y, Wurtzel O, Rubin EM, Tiedje JM, Sorek R. 2009. Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. Proc Natl Acad Sci U S A 106:3976-3981.

Figures

Figure 1. Main features of the *M. aeruginosavorus* chromosome

From the outside inward the circles show: (1) and (2) predicted protein-coding regions on the plus and minus strands (colors were assigned according to the color code of functional classes; (3) tRNA genes (purple) and rRNA genes (blue); (4) gene expression level as measured by the natural logarithm of Gene Expression Index (GEI), attack phase (green) and attachment phase (red); (5) GC skew plot; (6) GC%; (7) tri-nucleotide chi-square score; (8) genomic islands.

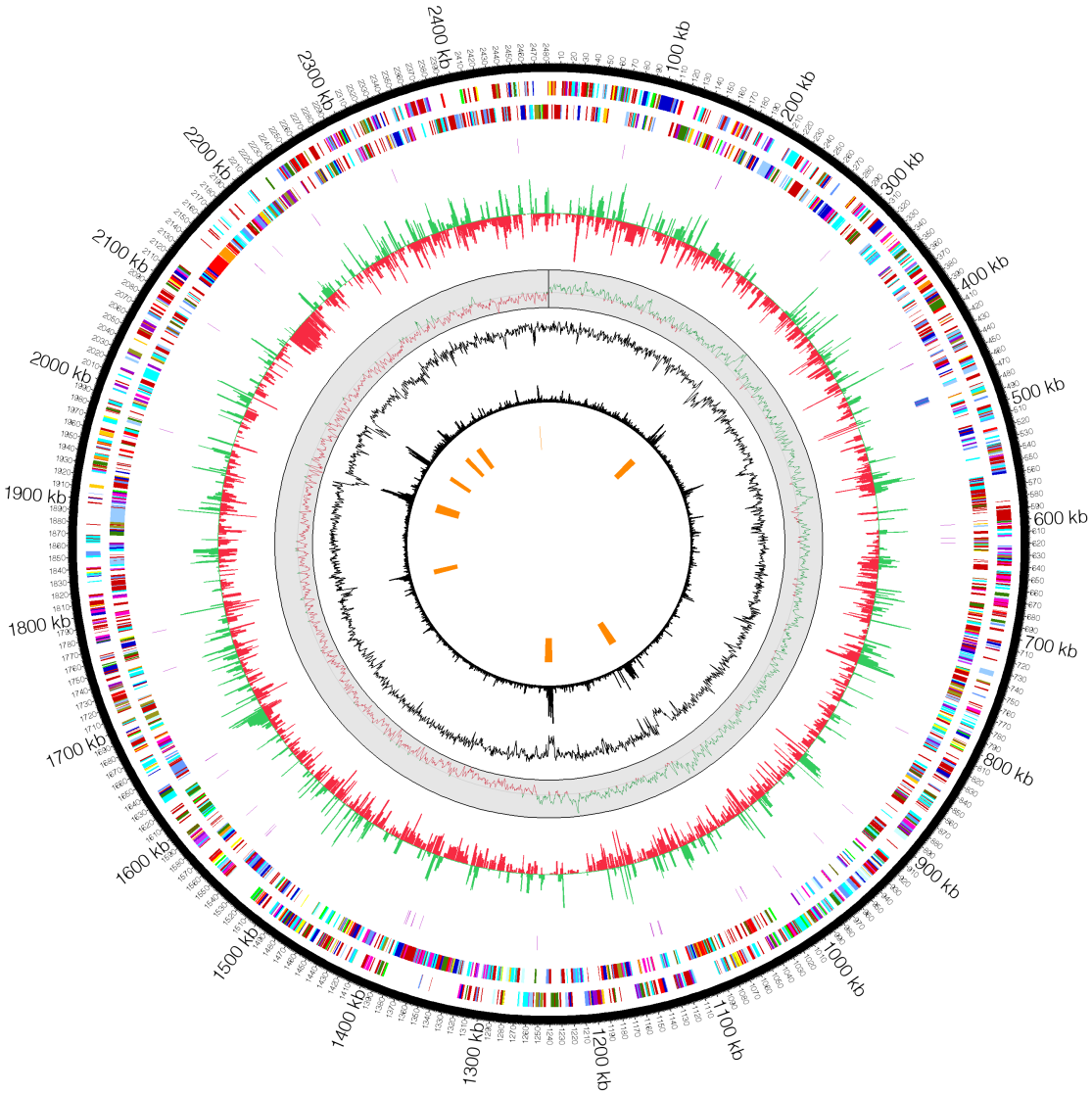
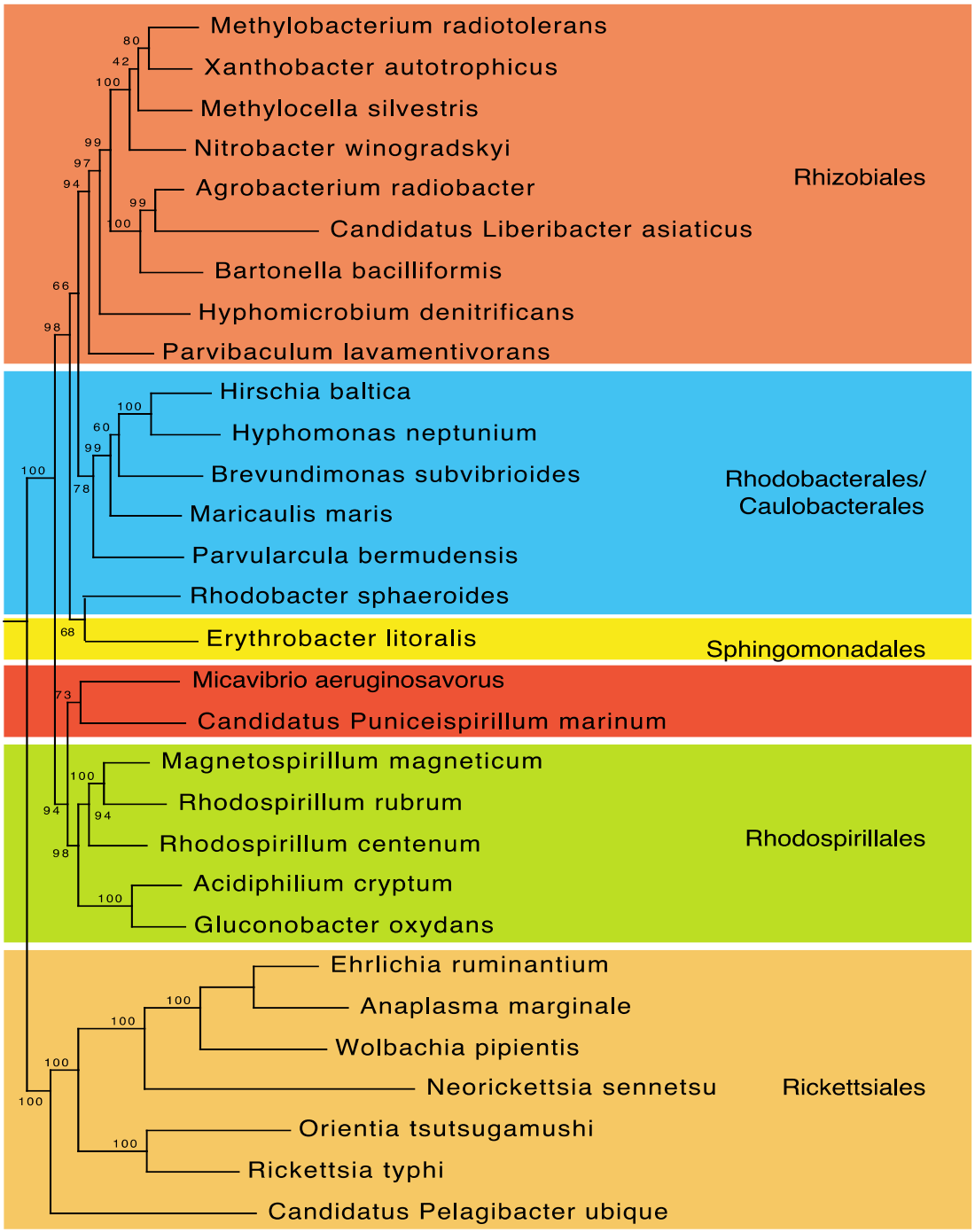


Figure 2. A maximum likelihood genome tree of α-proteobacterial representatives

A maximum likelihood tree was built from concatenated protein sequences of 31 universal housekeeping genes and rooted by γ- and β-proteobacteria. Bootstrap support values (out of 100 runs) for branches of interest are shown beside them.



Tables

Table 1. Main features of the genome of *M. aeruginosavorus* ARL-13

Feature	**Value**
Genome Size, bp	2,481,983
GC%	54.7
Predicted open reading frames (ORFs)	2434
ORFs with assigned function	1228 (50.5%)
Conserved hypothetical ORF	193 (7.9%)
Unknown function ORF	124 (5.1%)
Hypothetical ORF	746 (30.6%)
Average ORF length, bp	919
Percent of genome that is coding	90.3
Ribosomal RNA operon	3
Transfer RNA	40
CRISPR element	0
Plasmid	0

Table 2. Hemolysin-related proteins encoded by *M. aeruginosavorus*

Gene	Length (aa)	No. of Hemolysin-type calcium binding repeat	Other Motifs	Type I secretion system signal	GEI[a] in attachment/attack phase	Located within a genomic Island
GMV0092	559	6		+	30.0/124.4	
GMV0093	495	0		+	1.8/27.2	
GMV0107	2892	5	Von Willebrand factor	+	16.9/0.3	
GMV0287	1876	11		+	4.1/1.5	+
GMV1777	1296	17		+	4.8/2.2	+
GMV2456	1238	18	Lectin	+	0.4/0.1	+

[a]**GEI:** gene expression index

Additional files

Additional file 1. Comparison of major metabolic pathways between *Micavibrio aeruginosavorus*, *Bdellovibrio bacteriovorus* and *Escherichia coli*.

Main metabolic features	***M. aeruginosavorus***	***B. bacteriovorus***	***E. coli***
Amino acid biosynthesis	37 (1.52%)	76 (2.09%)	115 (2.13%)
Purines, pyrimidines, nucleosides, and nucleotides	30 (1.23%)	76 (2.09%)	82 (1.52%)
Fatty acid and phospholipid metabolism	32 (1.31%)	82 (2.26%)	70 (1.30%)
Biosynthesis of cofactors, prosthetic groups, and carriers	40 (1.64%)	101 (2.78%)	104 (1.93%)
Central intermediary metabolism	5 (0.21%)	145 (4.00%)	73 (1.35%)
Energy metabolism	99 (4.07%)	369 (10.17%)	397 (7.36%)
Transport and binding proteins	89 (3.66%)	344 (9.48%)	321 (5.95%)
DNA metabolism	86 (3.53%)	151 (4.16%)	107 (1.98%)
Transcription	35 (1.44%)	73 (2.01%)	45

			(0.83%)
Protein synthesis	117 (4.81%)	156 (4.30%)	121 (2.24%)
Protein fate	140 (5.75%)	230 (6.34%)	117 (2.17%)
Regulatory functions & Signal transduction	107 (4.40%)	191 (5.26%)	175 (3.24%)
Cell envelope	164 (6.74%)	391 (10.77%)	180 (3.34%)
Cellular processes	83 (3.41%)	263 (7.25%)	190 (3.52%)
Mobile and extrachromosomal element functions	16 (0.66%)	8 (0.22%)	50 (0.93%)

Additional file 2. Hydrolytic Enzymes encoded by *M. aeruginosavorus.*

Categories	Predicted location	Number	Gene
Proteases/Peptidases		Σ49	
	Extracellular	2	GMV0435, GMV1493
	Periplasmic	2	GMV2106, GMV0400
	OuterMembrane	1	GMV1190
	CytoplasmicMembrane	15	GMV1189, GMV1323, GMV1332, GMV1734, GMV1942, GMV2330, GMV2447, GMV2468, GMV2469, GMV0251, GMV0252, GMV0253, GMV0254, GMV0520, GMV0801
	Cytoplasmic	20	GMV1210, GMV1381, GMV1382, GMV1482, GMV1485, GMV1495, GMV0187, GMV0191, GMV2327, GMV2336, GMV2342, GMV2343, GMV2344, GMV0239, GMV0036, GMV0498, GMV0500, GMV0719, GMV0728, GMV0095

	Unknown	9	GMV1237, GMV1694, GMV1849, GMV2269, GMV0402, GMV0053, GMV0542, GMV0733, GMV0929
Lipases		Σ12	
	Extracellular	1	GMV1133
	CytoplasmicMembrane	1	GMV1286
	Cytoplasmic	9	GMV1056, GMV1100, GMV1106, GMV1602, GMV1603, GMV2060, GMV2379, GMV0860, GMV0890
	Unknown	1	GMV0881
Other Hydrolases		Σ37	
	Periplasmic	1	GMV0451
	OuterMembrane	1	GMV0827
	CytoplasmicMembrane	6	GMV1020, GMV2460, GMV0411, GMV0425, GMV0448, GMV0737
	Cytoplasmic	19	GMV1024, GMV1309, GMV1318,

			GMV1325, GMV1327, GMV1329, GMV1335, GMV1788, GMV1891, GMV1913, GMV2181, GMV2396, GMV0316, GMV0447, GMV0476, GMV0491, GMV0604, GMV0079, GMV0934
	Unknown	10	GMV1188, GMV1565, GMV1566, GMV2004, GMV0318, GMV0450, GMV0633, GMV0780, GMV0078, GMV0927
DNase		Σ2	
	Cytoplasmic	1	GMV0206
	Unknown	1	GMV0804
RNase		Σ4	
	Cytoplasmic	3	GMV1208, GMV1952, GMV0356

Unknown	1	GMV0812

Additional file 3. Flagellum biosynthesis and chemotaxis genes of *M. aeruginosavorus*.

Gene	Description
Regulators	
NA	
Export apparatus	
GMV0749	flagellar biosynthesis protein FlhA
GMV1696	flagellar biosynthetic protein FlhB
GMV1697	flagellar biosynthetic protein FliR
GMV1698	flagellar biosynthetic protein FliQ
GMV1707	flagellar biosynthetic protein FliP
GMV0751	flagellar protein export ATPase FliI
MS-, P- and L-rings	
GMV1717	flagellar basal body P-ring formation protein FlgA
GMV0767	flagellar M-ring protein FliF
GMV1723	flagellar P-ring protein (Basal body P-ring protein)
GMV1718	flagellar L-ring family protein
Hook and basal body	
GMV1716	flagellar basal-body rod protein FlgG
GMV1724	flagellar FlgJ-like protein
GMV2433	flagellar hook-associated protein FlgK
GMV1700	flagellar hook-basal body complex protein FliE family protein
GMV1701	flagellar basal-body rod protein FlgC
GMV1702	flagellar basal-body rod protein FlgB
GMV1727	flagellar hook capping family protein

Rotor	
GMV0766	flagellar motor switch protein FliG
GMV0764	flagellar motor switch protein fliN
GMV1713	flagellar motor switch protein FliM
Motor	
GMV0763	motA/TolQ/ExbB proton channel family protein
Filament	
GMV0775	flagellin
Chemotaxis	
GMV2029	chemotaxis regulator transmitting signal to flagellar motor component
GMV2033	chemotaxis regulator transmitting signal to flagellar motor component
GMV1044	methyl-accepting chemotaxis protein
GMV0470	cheR methyltransferase, all-alpha domain protein
GMV0711	cheR methyltransferase, SAM binding domain protein
GMV2027	cheR methyltransferase, SAM binding domain protein
GMV2312	methyl-accepting chemotaxis (MCP) signaling domain protein
Unknown function within flagellar structure	
GMV0991	flagellar basal body-associated protein FliL family protein
GMV1714	flagellar basal body-associated protein FliL family protein

Additional file 4*. Gene expression index (GEI) derived from RNA-Seq.

A excel file listing the gene expression index for all ORFs of *M. aeruginosavorus* in the attachment and the attack phases.

*Additional file 4 is available at:

http://www.biomedcentral.com/1471-2164/12/453

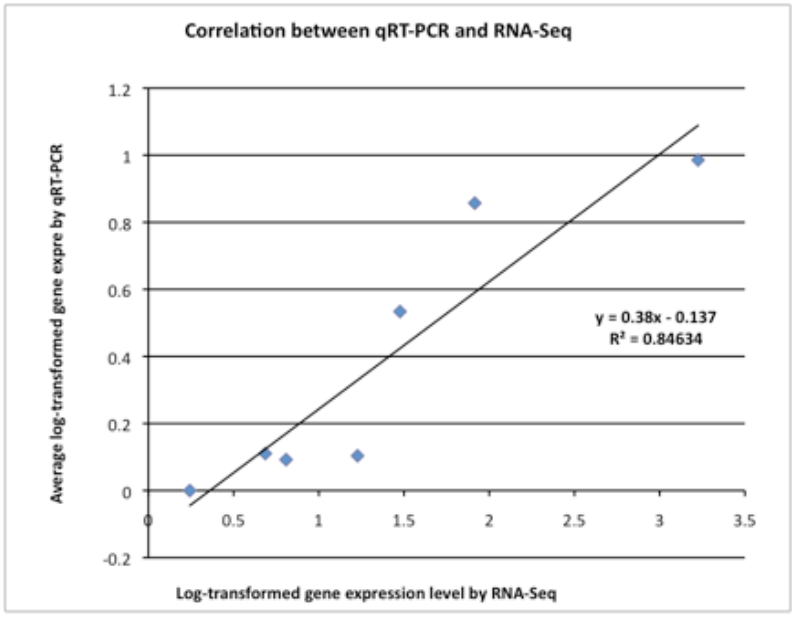
Additional file 5. Genomic islands in *M. aeruginosavorus* ARL-13.

Location	Length (Kbp)	Genes of interest
GMV0276-0292	21.6	integrase, primase, helicase, hemolysin-related proteins, luxR, tRNA-Thr
GMV1004-1026	24.8	polysaccharide biosynthesis, asparagine synthase
GMV1243-1260	24.2	integrase, copper resistance gene, mobilization gene, type I restriction enzyme, tRNA-Ser
GMV1766-1779	15.9	integrase, hemolysin-related proteins, addiction module, peptidoglycan binding protein
GMV1981-2002	27.4	resolvase, helicase, autoinducer luxI, peptidoglycan binding protein, cadmium resistance gene
GMV2073-2082	13.3	integrase, primase, reverse transcriptase

GMV2168-2176	14.9	integrase, tRNA-Gly
GMV2213-2226	16.2	integrase, cadmium resistance gene, mobilization gene, tRNA-Tyr
GMV2451-2460	11.4	integrase, hemolysin-related protein, addiction module

Additional file 6. Correlation between qRT-PCR and RNA-Seq.

An image file in PNG format showing the correlation between qRT-PCR and RNA-Seq data for selected genes in the *Micavibrio* attachment sample. Genes were selected to represent a broad range of gene expression levels. They were: GMV0043 (porin), GMV0092, GM0093, GMV0107 (hemolysin-related proteins), GMV1700 (flagellar hook-basal body complex FliE family), GMV2023 (bacterial regulatory tetR family protein) and GMV2138 (ribosomal protein S7).



Appendix 1. Comparative genomic insights into amoeba endosymbionts belonging to the families of "*Holosporaceae*" and "*Candidatus Midichloriaceae*" within *Rickettsiales*

Abstract

Rickettsiales are a group of obligate intracellular bacteria with greatly diversified lifestyles and host ranges. Previous sequenced *Rickettsiales* genomes mostly fall on two families *Rickettsiaceae* and *Anaplasmataceae*, with two other families *Holosporaceae* and *Candidatus Midichloriaceae* largely overlooked. We sequenced the genomes of four amoeba endosymbionts, three of which belonging to *Holosporaceae* and one to *Candidatus Midichloriaceae*. All the four endosymbionts have streamlined genomes and are completely devoid of *de novo* amino acid and nucleotide biosynthesis pathways, suggesting that they have to strictly depend on their host resources to multiple. Strikingly, we identified multiple gene expansions in most of these endosymbionts, such as *UmuCD, ProPQ, FadD*, *CheY* and *LysR*-type transcriptional regulator (LTTR), all of which are related to the spectacular intracellular lifestyle of these amoeba endosymbionts and their interaction with hosts. Genome analysis of these endosymbionts identified a large number of genes most closely related to homologs in various other amoeba-associated bacteria, supporting the hypothesis that amoeba serve as a "melting pot" facilitating the genetic exchange among distantly related intra-amoeba bacteria. Finally, phylogenomic analysis with these endosymbiont genomes indicates both genome reduction and expansion within *Rickettsiales.*

Introduction

Rickettsiales are a deep-branched order of α-proteobacteria consisting of obligate intracellular bacteria. It encompasses at least four distinct families, *Rickettsiaceae, Anaplasmataceae*, *Holosporaceae* and *Candidatus Midichloriaceae*. Most members of *Rickettsiaceae* and *Anaplasmataceae* are pathogens of a wide range of multicellular eukaryotic hosts, including arthropods, nematodes, and mammals (Andersson, et al. 1998; Wu, et al. 2004; Brayton, et al. 2005; Cho, et al. 2007). In comparison, a large number of bacteria within *Holosporaceae* and *Candidatus Midichloriaceae* are endosymbionts of unicellular protists, including *Paramecium* and *Acanthamoeba* (Horn, et al. 1999; Beier, et al. 2002; Montagna, et al. 2013).

Acanthamoeba is known to harbor a remarkably wide range of intracellular bacteria belonging to dispersed phylogenetic lineages throughout the bacterial tree of life, such as *Chlamydia, Legionella, Burkholderia, Francisella, Listeria,* and *Mycobacteria* (Schmitz-Esser, et al. 2008; Moliner, et al. 2010). Interestingly, many of these bacteria are also pathogens of multicellular eukaryotes such as livestocks and human being. It has been suggested that *Acanthamoeba* serves as an evolutionary "training ground" for the emergence of these specialized bacterial pathogens (Molmeret, et al. 2005).

Within *Rickettsiales*, the families of *Rickettsiaceae* and *Anaplasmataceae* have been extensively studied because of their biomedical and pathological importance. However, very little is known about members of *Holosporaceae* and *Candidatus Midichloriaceae*. As part of our effort to fill the gaps in the tree of life and pinpoint the origin of mitochondria, we sequenced the genomes of four amoeba endosymbionts, three of which belonging to the family *Holosporaceae* (*Candidatus Caedibacter acanthamobae* (Cca), *Candidatus Paracaedibacter acanthamoebae* (Cpa),

Candidatus Paracaedibacter symbiosus (Cps)) and one belonging to the family *Candidatus Midichloriaceae* (*Endosymbiont of Acanthamoeba UWC8* (Eau)). Comparative genomic analyses of these endosymbionts provide many novel insights into their unique biology and mechanisms of host-symbiont interaction.

Results and Discussion

Major features of the four genomes are summarized in Table 1. The complete assembly of Eau consists of one single circular molecule of 1,615,277 bp (Figure 1). In comparison, the complete assembly of Cca comprises five replicons, a circular chromosome of 1,722,347 bp and four circular plasmids (Figure 1). The statistics of the other two incomplete genomes are in general comparable with Cca, although the presence of plasmids in these two genomes cannot be determined due to the incomplete nature of their assemblies (Table 1).

All four endosymbionts have streamlined genomes and limited metabolic capacities. Nevertheless, they are predicted to cover several major metabolic pathways, including glycolysis, the pentose phosphate pathway, the TCA cycle, the electron transport system and ATP synthase, indicating that they are capable of generating ATP on its own. On the other hand, all genomes are completely devoid of *de novo* amino acid and nucleotide biosynthesis pathways. Therefore, these endosymbionts have to uptake most of these nutrients from hosts. Accordingly, at least 13 transporters for amino acids, peptides or amines were identified in these genomes. In nucleotide transport, ATP/ADP translocase, the hallmark protein of many obligate intracellular bacteria, is present in all four genomes, and is duplicated in Eau and Cca.

Despite extensive genome reduction, a number of gene families have been dramatically expanded, most of which are related with the specialized intracellular lifestyles of these endosymbionts and their interaction with hosts (Table 2). For example, *UmuCD* are specialized DNA polymerases involved in the SOS response of DNA damage caused by agents such as UV light (Woodgate and Sedgwick 1992; Murli and Walker 1993). The proliferation of these genes are likely a result of adaptation of these endosymbionts to living within amoebae, which are frequently exposed to UV light in its aquatic habitat. Several other expanded genes, such as *FadD*, *LysR*-type transcriptional regulator (LTTR) and murein lytic transglycosylase (LTs) are associated with virulence of multiple bacterial pathogens (Betzner and Keck 1989; Maddocks and Oyston 2008; Kang, et al. 2010) and therefore likely play important roles in the amoeba-symbiont interaction.

Remarkably, all four endosymbionts possessed a large number of genes involved in response to environmental cues such as the two-component systems and chemotaxis (Table 2). Furthermore, each of the Cpa and Cps genomes encode at least five proteins in quorum sensing, which has been so far absent in obligate intracellular bacteria. The enrichment of two-component system is likely reflective of relatively unpredictable environment of *Acanthamoeba* compared to multicellular eukaryotes, such as fluctuation in temperature, osmolarity, pH and exposure to oxidizing agents. Also, both two-component system and quorum sensing system are involved in the regulation of virulence secretion for several pathogens such as *Salmonella enterica*, *Staphylococcus aureus* and *Pseudomonas aeruginosa* (Beier and Gross 2006; Antunes, et al. 2010). Thus it is tempting to speculate that endosymbionts residing within the same host might use a similar strategy to mediate intra-amoeba communication and to spread their virulence for host cell exploitation.

Acanthamoeba is known to harbor multiple distantly related endosymbionts, thereby providing a fertile ground for DNA exchanges between endosymbionts (Moliner, et al. 2010; Bertelli and Greub 2012). Phylogenetic analyses by Ogata et al. suggested that the whole *tra* cluster of *Rickettsia bellii* was likely acquired from *Protochlamydiae amoebaphila*, both of which are capable of infecting *Acanthamoeba* (Ogata, et al. 2006). Based on this, they proposed that amoeba could serve as a genetic "melting pot" for endosymbionts living within the same host (Ogata, et al. 2006). Comparative genomic analysis indicates that all four endosymbionts sequenced in this study contain a substantial fraction of genes whose best hits are in other distantly related amoeba-associated bacteria (Figure 2). This suggests that these four amoeba endosymbionts have potentially undergone extensive lateral gene transfers (LGTs) with the other species. A large number of gene families subject to such lateral genetic exchanges are transposases and recombinases. Accordingly, there are abundant of mobile genetic elements in the three sequenced *Holosporaceae* species (Table 1). This is in sharp contrast to other *Rickettsiales* lineages, which are free of mobile genetic elements in general. Intriguingly, among these LGT candidates, we identified several gene families related to bacterial virulence, including 40 histidine kinases, 29 patatin-like genes and 15 toxin-antitoxin genes. Therefore, our results further support the "melting pot" evolution, and suggest that it plays an important role in shaping the virulence and host cell interaction of these endosymbionts.

Finally, genome sequences of the four endosymbionts significantly increased the taxon representation of two poorly sampled families *Holosporaceae* and *Candidatus Midichloriaceae*, thereby allowing us to better understand the diversification and genome evolution of *Rickettsiales*. Our results suggest both genome reduction and expansion in the *Rickettsiales* evolution, with genome reduction leading to *Rickettsiaceae* and *Anaplasmataceae* most of which are pathogens of multicellular host, and genome expansion leading to *Holosporaceae* and

Candidatus Midichloriaceae which contain endosymbionts of unicellular host. A large number of genes acquired by *Holosporaceae* or *Candidatus Midichloriaceae* encode transcriptional regulators, type IV pilus proteins, amino acid transporters and two-component system, which are all important for the intracellular metabolism and virulence of the amoeba endosymbionts. On the other hand, the gene loss in *Rickettsiaceae* or *Anaplasmataceae* are mostly related to biosynthesis pathways, such as flagella biosynthesis, cell envelope biosynthesis and *de novo* nucleotide biosynthesis, which is likely a result of adaptation of these specialized pathogens to the more stable intracellular niches of multicellular hosts. The genome expansion leading to *Holosporaceae* and *Candidatus Midichloriaceae* challenges the well-established notion that genome reductive evolution is a common feature of obligate intracellular bacteria. And the genome reduction leading to *Anaplasmataceae* and *Rickettsiaceae* is consistent with the role of amoeba as the "training ground" leading to the adaptation of bacterial pathogens.

References

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG. 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature 396:133-140.

Antunes LC, Ferreira RB, Buckner MM, Finlay BB. 2010. Quorum sensing in bacterial virulence. Microbiology 156:2271-2282.

Beier CL, Horn M, Michel R, Schweikert M, Gortz HD, Wagner M. 2002. The genus Caedibacter comprises endosymbionts of Paramecium spp. related to the Rickettsiales (Alphaproteobacteria) and to Francisella tularensis (Gammaproteobacteria). Appl Environ Microbiol 68:6043-6050.

Beier D, Gross R. 2006. Regulation of bacterial virulence by two-component systems. Curr Opin Microbiol 9:143-152.

Bertelli C, Greub G. 2012. Lateral gene exchanges shape the genomes of amoeba-resisting microorganisms. Front Cell Infect Microbiol 2:110.

Betzner AS, Keck W. 1989. Molecular cloning, overexpression and mapping of the slt gene encoding the soluble lytic transglycosylase of Escherichia coli. Mol Gen Genet 219:489-491.

Brayton KA, Kappmeyer LS, Herndon DR, Dark MJ, Tibbals DL, Palmer GH, McGuire TC, Knowles DP. 2005. Complete genome sequencing of Anaplasma marginale reveals that the surface is skewed to two superfamilies of outer membrane proteins. Proceedings of the National Academy of Sciences of the United States of America 102:844-849.

Cho NH, Kim HR, Lee JH, Kim SY, Kim J, Cha S, Kim SY, Darby AC, Fuxelius HH, Yin J, et al. 2007. The Orientia tsutsugamushi genome reveals massive proliferation of conjugative type IV secretion system and host-cell interaction genes. Proc Natl Acad Sci U S A 104:7981-7986.

Horn M, Fritsche TR, Gautom RK, Schleifer KH, Wagner M. 1999. Novel bacterial endosymbionts of Acanthamoeba spp. related to the Paramecium caudatum symbiont Caedibacter caryophilus. Environ Microbiol 1:357-367.

Kang Y, Zarzycki-Siek J, Walton CB, Norris MH, Hoang TT. 2010. Multiple FadD acyl-CoA synthetases contribute to differential fatty acid degradation and virulence in Pseudomonas aeruginosa. PLoS One 5:e13557.

Maddocks SE, Oyston PC. 2008. Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. Microbiology 154:3609-3623.

Moliner C, Fournier PE, Raoult D. 2010. Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. FEMS Microbiol Rev 34:281-294.

Molmeret M, Horn M, Wagner M, Santic M, Abu Kwaik Y. 2005. Amoebae as training grounds for intracellular bacterial pathogens. Applied and Environmental Microbiology 71:20-28.

Montagna M, Sassera D, Epis S, Bazzocchi C, Vannini C, Lo N, Sacchi L, Fukatsu T, Petroni G, Bandi C. 2013. "Candidatus Midichloriaceae" fam. nov. (Rickettsiales), an ecologically widespread clade of intracellular alphaproteobacteria. Appl Environ Microbiol 79:3241-3248.

Murli S, Walker GC. 1993. SOS mutagenesis. Curr Opin Genet Dev 3:719-725.

Ogata H, La Scola B, Audic S, Renesto P, Blanc G, Robert C, Fournier PE, Claverie JM, Raoult D. 2006. Genome sequence of Rickettsia bellii illuminates the role of amoebae in gene exchanges between intracellular pathogens. PLoS Genet 2:e76.

Schmitz-Esser S, Toenshoff ER, Haider S, Heinz E, Hoenninger VM, Wagner M, Horn M. 2008. Diversity of bacterial endosymbionts of environmental acanthamoeba isolates. Appl Environ Microbiol 74:5822-5831.

Woodgate R, Sedgwick SG. 1992. Mutagenesis induced by bacterial UmuDC proteins and their plasmid homologues. Mol Microbiol 6:2213-2218.

Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, Brownlie JC, McGraw EA, Martin W, Esser C, Ahmadinejad N, et al. 2004. Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: a streamlined genome overrun by mobile genetic elements. PLoS Biol 2:E69.

Figures

Figure 1. Main features of the complete genomes of Eau and Cca. On the top are the main chromosomes of Eau (left) and Cca (right). On the bottom are the four plasmids of Cca (pCca1-4). From the outside inward the circle shows (1) and (2) predicted ORFs on the plus and minus strands (colors were assigned according to the color code of functional classes); (3) tRNA (green) and rRNA (grey) genes (absent in the four plasmids); (4) tri-nucleotide chi-square score; (5) GC% (green represents above average and red represents below average); (6) GC skew plot (yellow represents plus and blue represents minus).

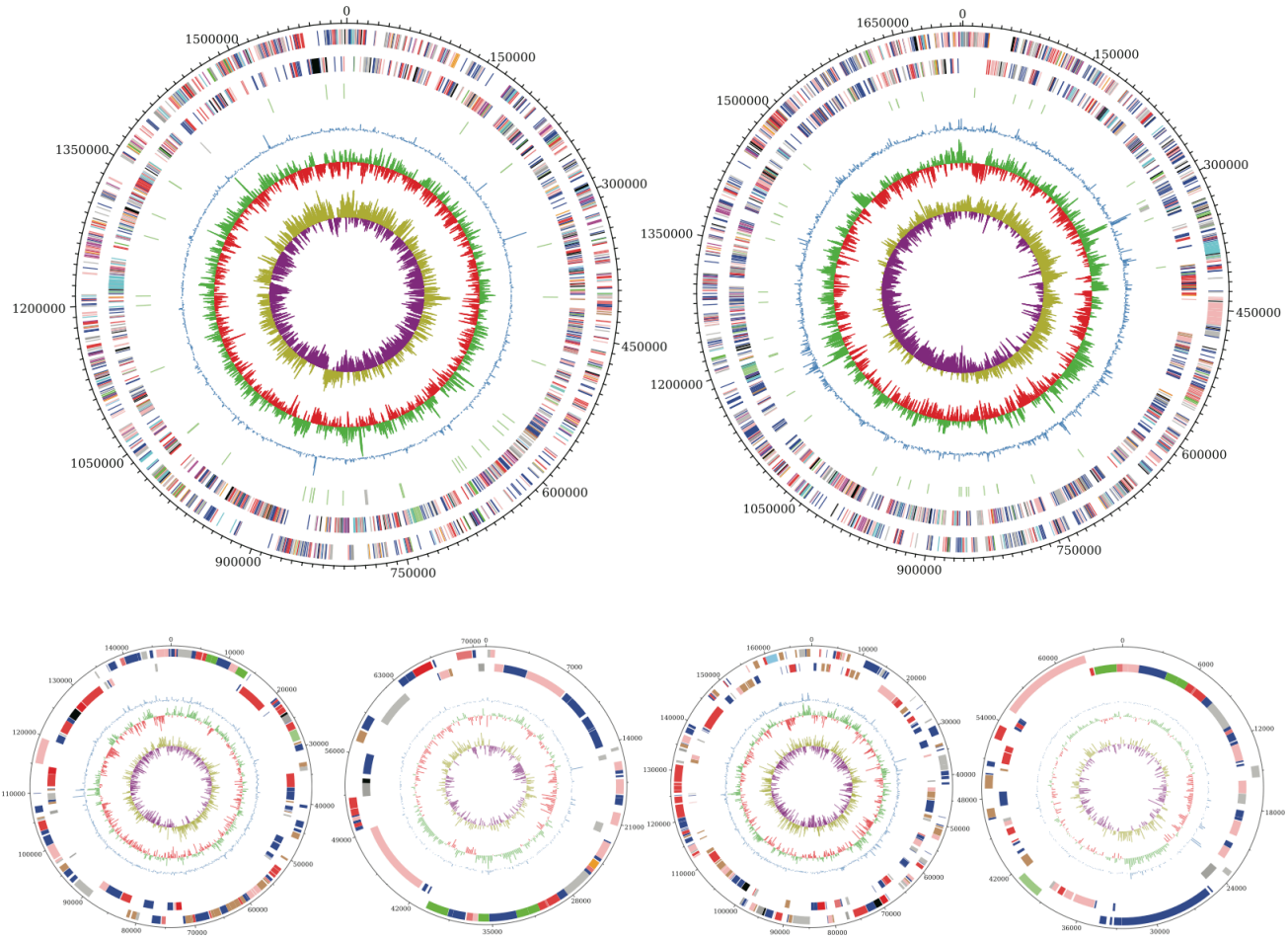


Figure 2. “Melting-pot” evolution of *Rickettsiales*. The histogram on the top right represents the taxonomic distribution of best hits (Y-axis) when genes in the four amoeba endosymbionts (highlighted in red) and the 11 *Rickettsiales* representatives (X-axis) were BLASTP searched against all complete bacterial genomes. The 10 top-ranked non α-proteobacterial lineages with the most hits are shown. The amoeba-associated bacteria are highlighted by asterisks. The histogram is enlarged and shown at bottom panel.

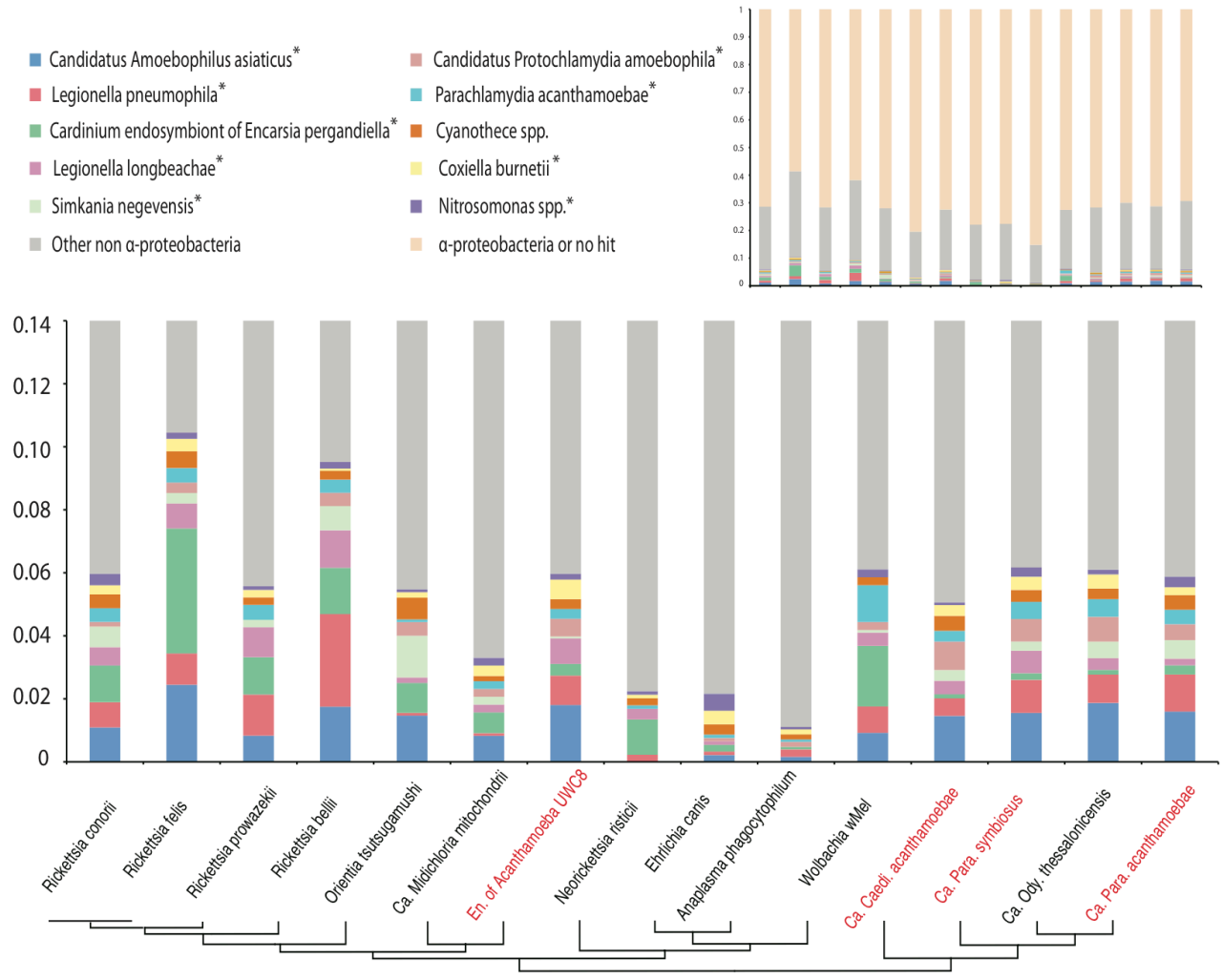
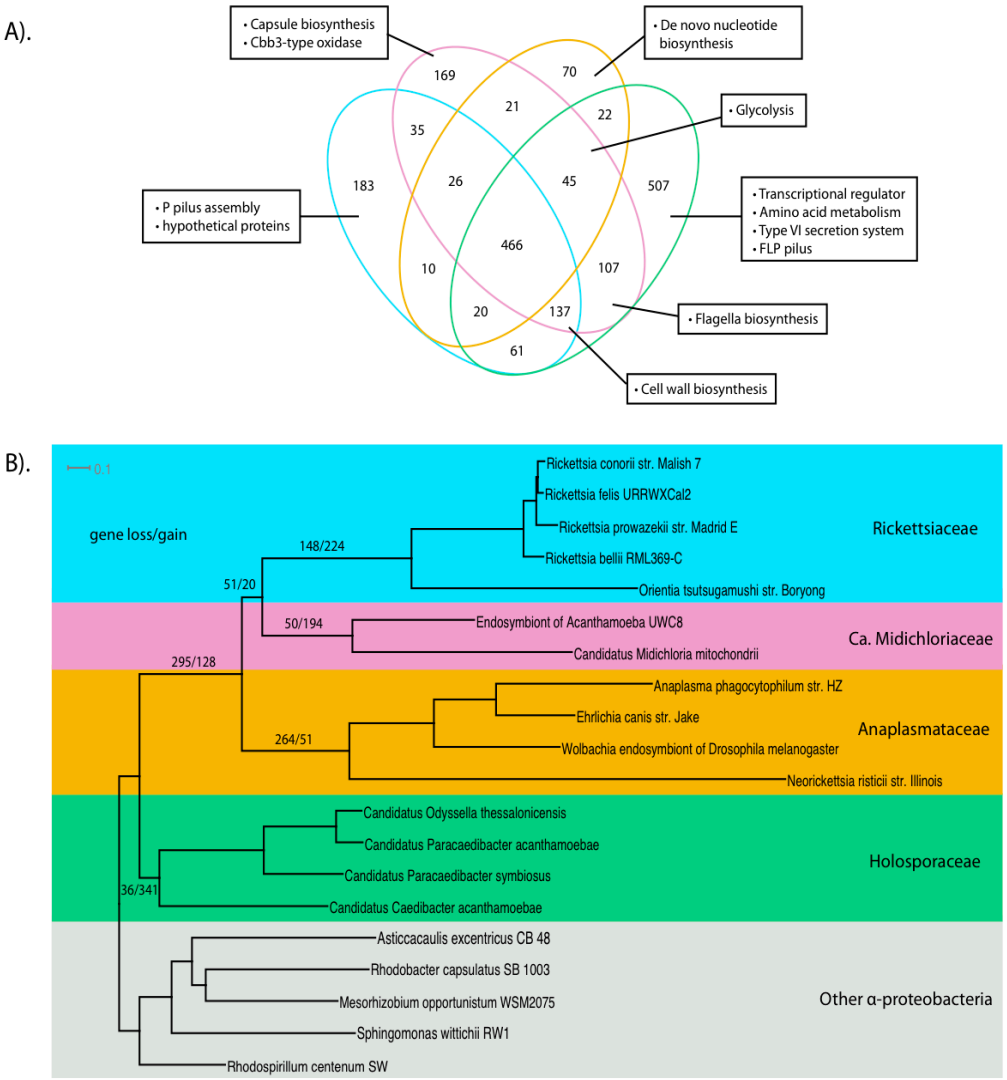


Figure 3. *Rickettsiales* genome evolution. Figure 3A is a Venn diagram illustrating the gene content comparison of the last common ancestor of *Holosporaceae* (green), *Candidatus Midichloriaceae* (pink), *Rickettsiaceae* (blue) and *Anaplasmataceae* (orange). Figure 3B represents a Bayesian genome tree of *Rickettsiales* based on concatenated protein sequences of 200 α-proteobacterial phylum-level markers. The numbers of gene gain and loss events are displayed beside each branch of interest. Posterior probability values are 1.0 for all the internal nodes.



Tables

Table 1. Main features of the four endosymbiont genomes.

	Cca	**Cpa**	**Cps**	**Eau**
Genome Size, bp	2,175,773	2,455,062	2,665,575	1,615,277
GC %	37.9%	41.0%	41.3%	34.7%
Plasmids	4	NA	NA	0
Predicted ORFs	2,332	2,382	2,383	1,608
ORFs with assigned functions	56.1%	53.6%	53.3%	65.2%
Average ORF length, bp	808	866	793	898
Percent of genome that is coding	86.4%	84.3%	71.1%	89.4%
Ribosomal RNA operon	1	2	2	1
Transfer RNA	42	40	41	36
Transposase	17	11	15	9
Mobile genetic element	98	114	63	1

Table 2. A list of gene families and functional pathways that were expanded in the four endosymbiont genomes. The number of genes in each family/pathway is shown for each of the four endosymbiont genomes (Cca, Cpa, Cps and Eau), as well as four other *Rickettsiales* genomes (Cot (*Candidatus Odyssella thessaloniscensis*), Cmm (*Candidatus Midichloria mitochondrii*), Rbe (*Rickettsia bellii RML369-C*) and Rpr (*Rickettsia prowazekii str. Madrid E*)).

	UmuC	***UmuD***	***ProP***	***ProQ***	***FadD***	**LTTR**	**Murein LTs**	***CheY***	**Flagella synthesis**	**T6SS***	**2CS***
Cca	8	6	1	12	8	6	7	0	5	7	13
Cpa	2	2	7	3	8	3	6	7	27	8	26
Cps	2	2	13	2	4	12	7	4	28	8	32
Eau	1	2	1	0	1	0	4	0	29	0	26
Cot	4	5	8	2	2	10	7	9	28	8	34
Cmm	0	0	1	0	0	0	4	0	26	0	14
Rbe	0	1	8	0	0	0	4	0	0	0	10
Rpr	0	0	7	0	0	0	3	0	0	0	9

* T6SS: Type VI secretion system * 2CS: Two-component system