SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
ENVIRONMENTAL MICROBIOLOGY
EVOLUTION
PHYLOGENETICS
MICROBIAL GENETICS

Received
16 February 2011

Accepted
10 May 2011

Published
14 June 2011

Correspondence and requests for materials should be addressed to S.J.G. (steve.giovannoni@oregonstate.edu)

* Current address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, 45 Center Drive, Bethesda, MD 20892

Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade

J. Cameron Thrash[1], Alex Boyd[2], Megan J. Huggett[3], Jana Grote[3], Paul Carini[1], Ryan J. Yoder[2], Barbara Robbertse[2]*, Joseph W. Spatafora[2], Michael S. Rappé[3] & Stephen J. Giovannoni[1]

Departments of [1]Microbiology, and [2]Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, [3]Hawaii Institute of Marine Biology, SOEST, University of Hawaii at Manoa, Kaneohe, HI 96744.

Mitochondria share a common ancestor with the *Alphaproteobacteria*, but determining their precise origins is challenging due to inherent difficulties in phylogenetically reconstructing ancient evolutionary events. Nonetheless, phylogenetic accuracy improves with more refined tools and expanded taxon sampling. We investigated mitochondrial origins with the benefit of new, deeply branching genome sequences from the ancient and prolific SAR11 clade of *Alphaproteobacteria* and publicly available alphaproteobacterial and mitochondrial genome sequences. Using the automated phylogenomic pipeline Hal, we systematically studied the effect of taxon sampling and missing data to accommodate small mitochondrial genomes. The evidence supports a common origin of mitochondria and SAR11 as a sister group to the *Rickettsiales.* The simplest explanation of these data is that mitochondria evolved from a planktonic marine alphaproteobacterial lineage that participated in multiple inter-specific cell colonization events, in some cases yielding parasitic relationships, but in at least one case producing a symbiosis that characterizes modern eukaryotic life.

Vital to the evolution of all known eukaryotic cells was the endosymbiotic event that resulted in the permanent acquisition of bacteria that through time were transformed into mitochondria. Even eukaryotes that were previously thought to lack mitochondria have now been shown to contain some remnants of that original endosymbiosis, either through the maintenance of hydrogenosomes, mitosomes, or nuclear genes of mitochondrial origin[1,2]. Although argument continues about the circumstances of the founding endosymbiotic event[2,3], current hypotheses, making use of mitochondrial genome and proteome data, agree that the mitochondrial ancestor was most closely related to *Alphaproteobacteria*[4–11]. The rooting of the ancestral mitochondria within the *Alphaproteobacteria* has not yet been resolved, but better phylogenetic methodology and new mitochondrial and bacterial genome sequences offer opportunities to refine understanding of this singular evolutionary event.

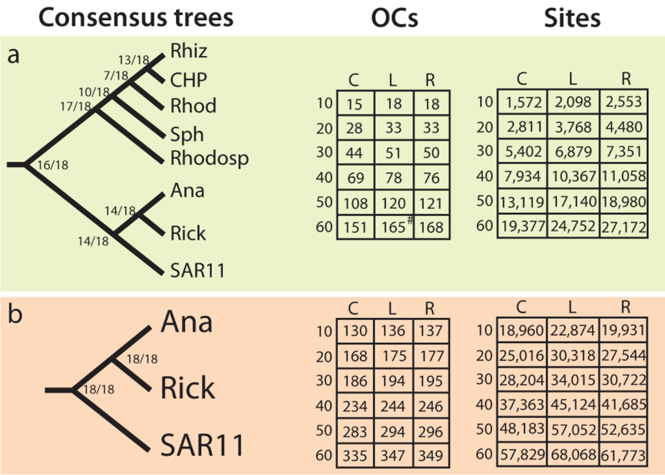
The sequencing of the first *Rickettsia* genome provided phylogenomic evidence for the previously established theory that the mitochondrial root was within the *Rickettsiales*[5,10,12,13] (and references therein). This theory was intellectually satisfying because all known *Rickettsia* have extremely reduced genomes and obligate intracellular lifestyles, although there was evidence to suggest that the genome reduction in mitochondria and *Rickettsia* probably happened independently[5,10,11]. While the rooting of mitochondria within the *Rickettsiales* has been debated based on other phylogenomic studies[4,7,14], it has gained support more recently with the addition of more genome sequences from taxa in the *Rickettsiales*, including the globally important SAR11 clade[6,15].

As free-living, predominantly marine bacteria, the phylogenetically diverse[16] organisms of the ubiquitous SAR11 clade are unique in the *Rickettsiales* order. Like other *Rickettsiales*, SAR11 have small genomes and, consequently, limited metabolic versatility[5,7,17–19]. However, SAR11 genomes bear distinctly different evolutionary signatures from other *Rickettsiales* genomes, indicating that selection for a streamlined genome, rather than genetic drift, has been the path of genome reduction in these organisms[9,17,19]. For example, the sequence of *Candidatus* Pelagibacter ubique HTCC1062 has few recent gene duplications, phage, transposons, or pseudogenes, has extremely small intergenic regions, and is optimized for a highly efficient lifestyle[17], whereas organisms like *Rickettsia* and those of the *Anaplasmataceae* contain large numbers of duplications, putatively inactive genes, larger intergenic regions, and/or mobile elements, indicative of the different selective pressures of obligate intracellular life[7,18,20,21]. Despite of the differences in evolutionary history and lifestyle, there is phylogenomic evidence for a shared common ancestor between SAR11 organisms and the *Rickettsiales*, and for rooting of the mitochondria in this group[6], but the latter is still a matter of debate[22]. Here, we systematically tested the phylogenomic relationship of the SAR11 clade to the mitochondria using an automated pipeline and with the benefit of new sequences from deeply branching SAR11 genomes[23].

Results

Experimental design and logic. We framed our phylogenomic testing around the following logic. First, we analyzed a subset (127) of available alphaproteobacterial genomes, including seven SAR11 genomes, with six outgroup species from the *Gamma*-, *Beta*-, and *Deltaproteobacteria* (Table S1, Fig. 1a). These outgroup taxa had a range of G+C content (40–67%, mean ~55%). Second, we tested the SAR11 genomes with the *Rickettsiales* separately for comparison to see if there was an observable effect of removing high G+C content taxa, using a smaller "median" outgroup with a mean G+C content of ~53% (Table S2, Fig. 1b). Third, with these established topologies, we then introduced mitochondrial sequences (Table S3) and evaluated the stability of the topology to variations in taxon sampling, gap-removal strategy, and allowed missing data using the same outgroup (Fig. 2). Fourth, we examined the consistency of these topologies to G+C bias of the outgroup taxa at low (mean ~40%) and high (mean ~67%) values (Figs. 4, 5).

Within the Hal pipeline, the cutoff percentage for the minimum number of taxa required for an orthologous cluster (OC) to be included (missing data) can be specified. The deleterious effect of missing data on phylogenetic analysis can be offset by the advantageous effect of additional taxon sampling if there are a large number of total sites, and can improve accuracy[24,25]. Thus, taxa with fewer protein coding genes, such as mitochondria, can be incorporated into a robust phylogenetic analysis that makes use of the signal generated by larger amounts of sequence from bacterial taxa with higher gene content. We analyzed each initial dataset over the range of 10–60% allowed missing data at 10% increments to assess the stability of tree topology across differing numbers of total included OCs, and to allow for the inclusion of taxa with varying amounts of coding sequence. Biased outgroup tests were then done on a subset of these missing data values (40 and 60%).



	C	L	R
10	15	18	18
20	28	33	33
30	44	51	50
40	69	78	76
50	108	120	121
60	151	165#	168

	C	L	R
10	1,572	2,098	2,553
20	2,811	3,768	4,480
30	5,402	6,879	7,351
40	7,934	10,367	11,058
50	13,119	17,140	18,980
60	19,377	24,752	27,172

	C	L	R
10	130	136	137
20	168	175	177
30	186	194	195
40	234	244	246
50	283	294	296
60	335	347	349

	C	L	R
10	18,960	22,874	19,931
20	25,016	30,318	27,544
30	28,204	34,015	30,722
40	37,363	45,124	41,685
50	48,183	57,052	52,635
60	57,829	68,068	61,773

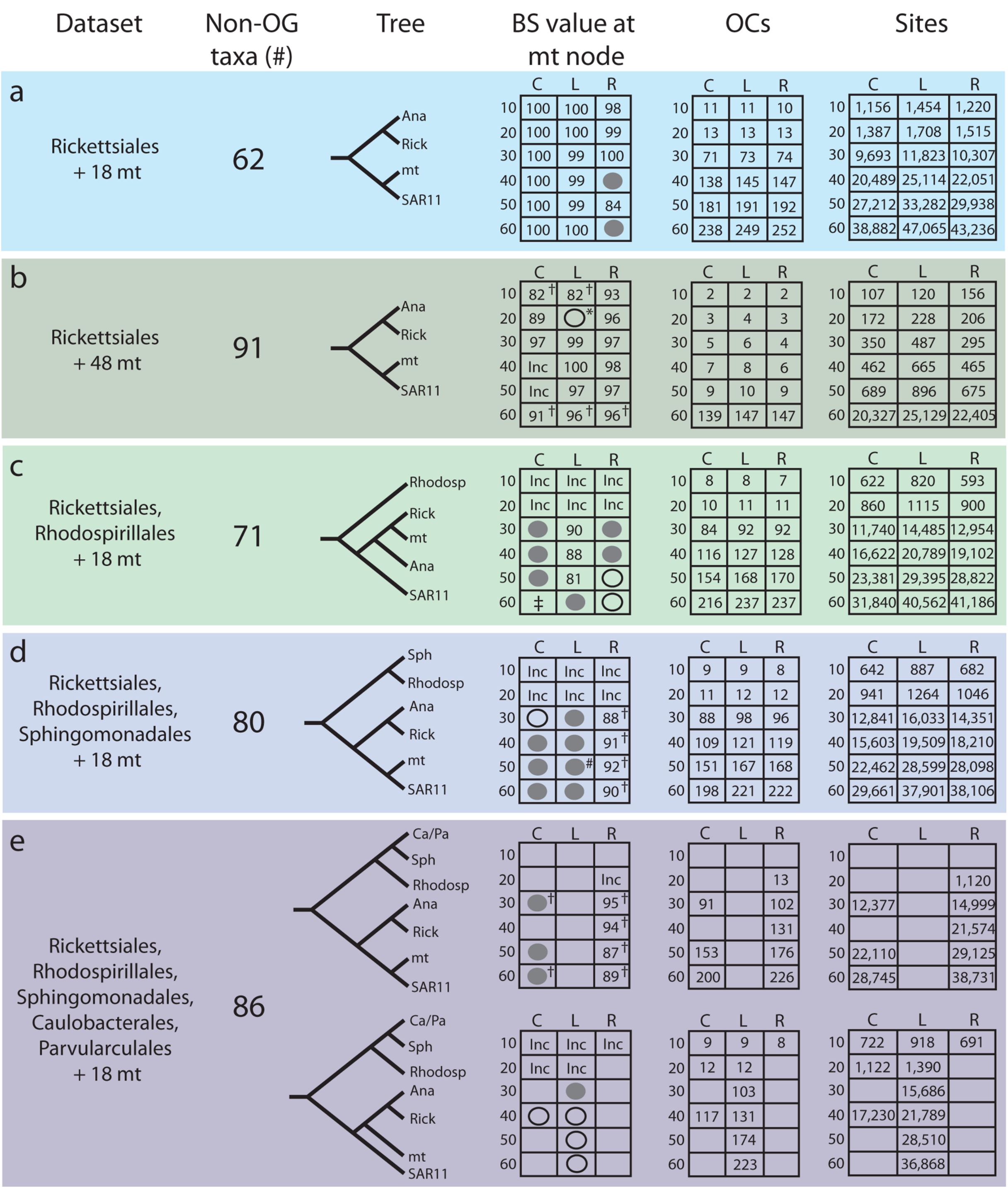
Figure 1 | Summary of phylogenomic analyses without mitochondrial sequences. The consensus trees for both the *Alphaproteobacteria* (a), and the *Rickettsiales* (b) across 10–60% allowed missing data are shown with concomitant information regarding number of OCs and amino acid sites for each individual tree. Rows are allowed missing data (10–60%), columns are gap-removal strategy (C- conservative, L- liberal, R- remgaps). Branching frequencies are noted at the nodes. # - this tree is featured in Fig. 6. Other abbreviations: Ana- *Anaplasmataceae*, Rick- *Rickettsiaceae*, Rhodosp- *Rhodospirillales*, Sph- *Sphingomonadales*, CHP- *Caulobacterales + Hyphomonadaceae + Parvularculales*, Rhod- *Rhodobacteraceae*.

The variations in taxon sampling, gap-removal strategy, and allowed missing data were put in place specifically to address potential systematic error known to effect phylogenomic analyses[26,27] (See Discussion). Changing these parameters causes different OCs, as well as different numbers of OCs, to be used for each analysis. This strategy can reveal error resulting from gene-specific heterotachy; likewise, altering the sampling of both ingroup and outgroup taxa with aberrant G+C content can reveal perturbations of branch order caused by compositional bias. In addition to these parameter variations, we also used the CAT model of rate heterogeneity, for which a Bayesian implementation has been shown to be helpful for overcoming both compositional signal error and long-branch attraction brought on by rate signal error[26,28,29,30].

Phylogenomic analysis of the *Alphaproteobacteria* and *Rickettsiales.* The majority of trees generated for the 127 *Alphaproteobacteria* genomes over 10–60% allowed missing data with different gap-removal stringencies supported predicted family- and order-level groupings, with the previously established exception that the *Caulobacterales*, *Hyphomonadaceae*, and *Parvularculales* always grouped together, thus making the *Rhodobacterales* polyphyletic[6] (Fig. 1a). Two of the 18 trees failed to reconstruct the *Rickettsiales* as a monophyletic group, and two separated the *Rhodospirillales*, although all four had poor bootstrap support at these nodes (Fig. S1, SI tree files). Sixteen of eighteen trees formed the characteristic bifurcation of the *Rickettsiales* separate from the remaining orders of the phylum, which has been demonstrated previously by different methods[6,31] (Fig. 1a). The branching order of clades between the *Rhodospirillales* and the *Rhizobiales* was unstable (Fig. S1); however, the consensus tree demonstrates that this was usually limited to the positioning of the *Rhodobacteraceae* and the *Sphingomonadales* (Fig. 1a). Regardless, the *Rickettsiales* consistently grouped away from the remainder of the tree and, in accordance with previous studies[6,32], the SAR11 clade was monophyletic with the *Rickettsiaceae* and the *Anaplasmataceae* in 16 of 18 trees; 2 of these 16 grouped SAR11 with the *Rickettsiaceae* instead of basal to the *Rickettsiaceae* and the *Anaplasmataceae*. The two trees where the *Rickettsiales* was paraphyletic had poorly supported nodes for the *Rickettsiaceae*/SAR11 group and the *Anaplasmataceae* (Fig. S1). A separate analysis adding the sequence of *Magnetococcus* sp. MC-1, for which a potential position within the *Alphaproteobacteria* has been suggested[8], placed this sequence among the outgroup taxa in all trees at 50 and 20% allowed missing data (data not shown).

We further focused our analysis on the *Rickettsiales* alone to verify the topology revealed by the complete alphaproteobacterial dataset with more included OCs and all 44 available genomes for the order, and to more firmly establish the relationship of the SAR11 clade to the *Rickettsiales* without the compositional signal of other, higher G+C content organisms. *Gamma*- and *Betaproteobacteria* sequences from the previous analysis served as outgroups (Table S2). The tree topology was consistent across 10–60% missing data, which included from 130 to 349 OCs (Fig. 1b). Importantly, at the family and genus level, this consensus tree agreed with the topology of the *Rickettsiales* order that was observed in the complete alphaproteobacterial phylogeny.

Inclusion of mitochondrial genome sequences. To test the phylogenomic placement of the mitochondria and SAR11 clades among the *Alphaproteobacteria* and monitor the stability of the resulting topologies to sampling perturbation, we adopted the following strategy: six different datasets with variation in sampling among mitochondrial or alphaproteobacterial taxa (Table S2) were tested across the range of 10–60% included missing data (Fig. 2). Established orders of *Alphaproteobacteria* were added in



a — Rickettsiales + 18 mt — 62

	BS C	BS L	BS R	OCs C	OCs L	OCs R	Sites C	Sites L	Sites R
10	100	100	98	11	11	10	1,156	1,454	1,220
20	100	100	99	13	13	13	1,387	1,708	1,515
30	100	99	100	71	73	74	9,693	11,823	10,307
40	100	99	●	138	145	147	20,489	25,114	22,051
50	100	99	84	181	191	192	27,212	33,282	29,938
60	100	100	●	238	249	252	38,882	47,065	43,236

b — Rickettsiales + 48 mt — 91

	BS C	BS L	BS R	OCs C	OCs L	OCs R	Sites C	Sites L	Sites R
10	82†	82†	93	2	2	2	107	120	156
20	89	○*	96	3	4	3	172	228	206
30	97	99	97	5	6	4	350	487	295
40	Inc	100	98	7	8	6	462	665	465
50	Inc	97	97	9	10	9	689	896	675
60	91†	96†	96†	139	147	147	20,327	25,129	22,405

c — Rickettsiales, Rhodospirillales + 18 mt — 71

	BS C	BS L	BS R	OCs C	OCs L	OCs R	Sites C	Sites L	Sites R
10	Inc	Inc	Inc	8	8	7	622	820	593
20	Inc	Inc	Inc	10	11	11	860	1115	900
30	●	90	●	84	92	92	11,740	14,485	12,954
40	●	88	●	116	127	128	16,622	20,789	19,102
50	●	81	○	154	168	170	23,381	29,395	28,822
60	‡	●	○	216	237	237	31,840	40,562	41,186

d — Rickettsiales, Rhodospirillales, Sphingomonadales + 18 mt — 80

	BS C	BS L	BS R	OCs C	OCs L	OCs R	Sites C	Sites L	Sites R
10	Inc	Inc	Inc	9	9	8	642	887	682
20	Inc	Inc	Inc	11	12	12	941	1264	1046
30	○	●	88†	88	98	96	12,841	16,033	14,351
40	●	●	91†	109	121	119	15,603	19,509	18,210
50	●	●#	92†	151	167	168	22,462	28,599	28,098
60	●	●	90†	198	221	222	29,661	37,901	38,106

e — Rickettsiales, Rhodospirillales, Sphingomonadales, Caulobacterales, Parvularculales + 18 mt — 86

	BS C	BS L	BS R	OCs C	OCs L	OCs R	Sites C	Sites L	Sites R
10									
20			Inc			13			1,120
30	●†		95†	91		102	12,377		14,999
40			94†			131			21,574
50	●		87†	153		176	22,110		29,125
60	●†		89†	200		226	28,745		38,731

	BS C	BS L	BS R	OCs C	OCs L	OCs R	Sites C	Sites L	Sites R
10	Inc	Inc	Inc	9	9	8	722	918	691
20	Inc	Inc		12	12		1,122	1,390	
30		●			103			15,686	
40	○	○		117	131		17,230	21,789	
50		○			174			28,510	
60		○			223			36,868	

Figure 2 | Summary of phylogenomic analyses including mitochondrial sequences. Data is divided by each dataset and corresponding tree topology (a–e). Internal tables for each dataset show the relevant information for each individual tree output. Rows are allowed missing data (10–60%), columns are gap-removal strategy (C- conservative, L- liberal, R- remgaps). Bootstrap support values ≥ 80% are reported, grey circles indicate a value between 79 and 51%, open circles ≤ 50%, Inc- incongruence in non-mitochondrial branch order compared with trees built without mitochondrial sequences and other studies. † - HIMB59 grouped with the mitochondria. * - two mitochondrial sequences grouped with the *Ehrlichia*. ‡ - mitochondria grouped with the SAR11 clade. # - this tree is featured in Fig. 7. Other abbreviations: mt- mitochondria, Ana- *Anaplasmataceae*, Rick- *Rickettsiaceae*, Rhodosp- *Rhodospirillales*, Sph- *Sphingomonadales*, Ca- *Caulobacterales*, Pa- *Parvularculales*, OG- outgroup, BS- bootstrap, OCs- orthologous clusters.

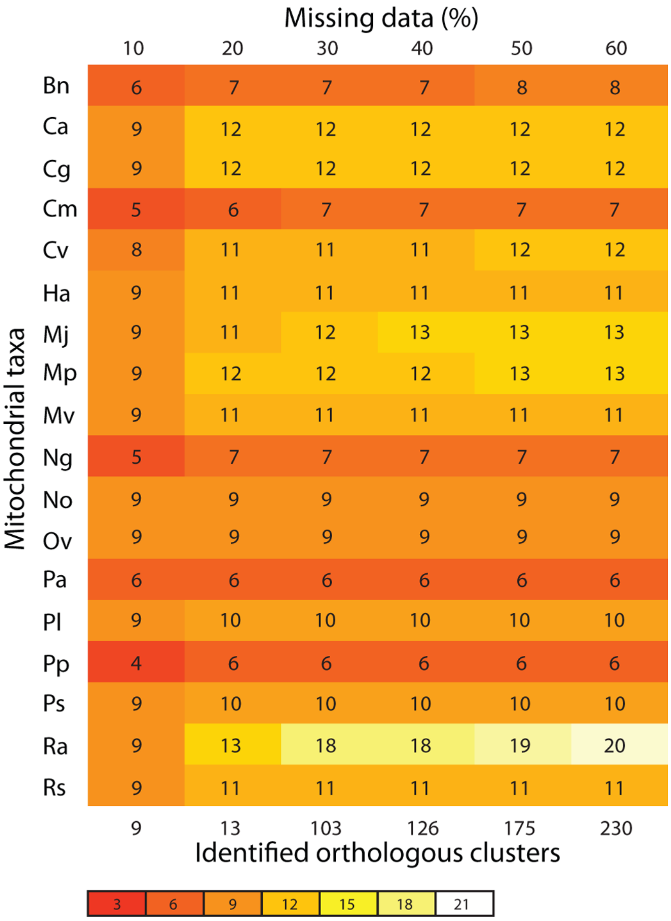
increments, using available genome sequences for each order and the same outgroups from the *Rickettsiales* analysis. We chose mitochondrial genome sequences (Table S3) as a function of the phylogenetic affiliation of their eukaryotic hosts. We included mitochondria from the Excavata, Chromalveolata, and Archaeplastida, as these clades have mitochondrial members considered "ancestral"[33,1,34], and are phylogenetically distinct from the "Unikonts," which include Metazoans[35,36], a group with highly variable mitochondrial evolutionary rates[37,38]. Further, in order to maximize the number of included OCs, we restricted our analyses to those mitochondria within these groups that contained more than 30 protein-coding genes (Table S3).

Figure 2 summarizes the results from the five datasets with a majority of trees that had branching patterns consistent among all datasets considered in this study. A sixth dataset, which included the *Rhodobacterales*, was in almost all cases incongruent in non-mitochondrial topology and/or had very poor bootstrap support at the major nodes or the mitochondrial node (SI tree files). The results from this dataset were therefore not included. Total OCs, the number of aligned sites, and bootstrap values for the mitochondrial node are reported in the figure. Trees that were incongruent with monophyletic relationships established without the mitochondrial sequences are noted as "Inc"; these topologies were usually poorly supported at major nodes where the branching order differed (SI tree files). 18 mitochondrial sequences (Table S3a) were tested with increasing numbers of alphaproteobacterial sequences (Fig. 2a, c–e), and a group of 48 mitochondrial sequences (Table S3b) was tested with just the *Rickettsiales*, including SAR11 (Fig. 2b). Two alternative topologies were observed with the largest dataset that was consistent with the monophyletic relationships we had established (Fig. 2e). In general, with the exception of the dataset in Fig. 2c, in all trees and all datasets with the median outgroup, the SAR11 were monophyletic with the mitochondria, or else were the outgroup to a clade joining mitochondria, the *Anaplasmataceae*, and the *Rickettsiaceae*, in either case linking them closely to mitochondrial origins (Fig. 2a, b, d, e).

When *Rickettsiales* were the only additional sequences in the tree, support for the monophyly of mitochondria and the SAR11 clade was very high and consistent across variation in missing data, gap-removal strategy, and the variation in taxon sampling within the mitochondria (Fig. 2a, b). One tree showed an anomalous pattern in which two mitochondrial sequences grouped with the *Ehrlichia* sequences (Fig. 2b asterisk), and two trees were incongruent with the trees that did not include mitochondria (Fig. 2b- "Inc"). Addition of the *Rhodospirillales* sequences changed the branching pattern and caused the mitochondria to group with the *Rickettsiaceae* instead of the SAR11 clade in all but one tree (Fig. 2c). In this dataset and those in Fig. 2d–e, trees with the fewest OCs were incongruent. However, inclusion of missing data resulted in considerably more signal and consistency in topology across datasets. Additionally, inclusion of more taxa from the orders *Sphingomonadales*, *Caulobacterales*, and *Parvularculales* reversed the topological change in Fig. 2c and again grouped the mitochondria as either monophyletic with SAR11 (Fig. 2d, e) or in a paraphyletic relationship with SAR11 as the outgroup to a clade joining mitochondria, the *Anaplasmataceae*, and the *Rickettsiaceae* (Fig. 2e). More trees placed the mitochondria and SAR11 as monophyletic (7) than as paraphyletic (5), and were better supported. In several trees, the deepest branching SAR11 strain, HIMB59, grouped with the mitochondria (Figs. 2, 7 daggers), thus making the monophyly of these groups more pronounced.

The addition of more taxa did not have a significant effect on the total numbers of OCs used in the analyses, except in the case where additional mitochondria sequences were used (Fig. 2b). While the relative contribution of each taxon decreased as the allowed missing data parameter was increased and more OCs were included, the absolute number of OCs increased for most taxa, thereby improving the net signal contributed by a given taxon. Figure 3 shows this effect for the mitochondrial taxa in Fig. 2d in trees with liberal gap removal (column L), where an increase in OCs was observed for every taxon; in some the increase was 70% or greater across the tested range.

Effects of outgroup G+C content bias. The deep-branching position of the *Rickettsiales* and SAR11 make them particularly vulnerable to compositional bias-based attraction artifacts depending on the outgroup taxa. The initial outgroup selection was the same as that of Williams et al.[6] (less two taxa) and covered a range of G+C content (40–67%) which we utilized specifically to avoid such artifacts. However, since the monophyly of SAR11 and the *Rickettsiales* is essential to examining the placement of the mitochondria in that group we tested the *Alphaproteobacteria* dataset in the presence of biased outgroups of both low and high G+C content. We also tested the stability of the datasets in Fig. 2 with biased outgroups. The number, and taxonomic affiliation at the class level, of outgroup taxa was kept constant to avoid significantly changing the number of included OCs or introducing additional artifacts based on branch distances. The biased *Alphaproteobacteria* outgroups spanned G+C ranges of 32–46%, mean ~40%, and 65–75%, mean ~67%, for the "low" and "high" groups, respectively (Table S1). The biased outgroups for the datasets including mitochondria were chosen with similar means: 41% (low) and 66% (high) (Table S2). All datasets were tested with 40 and 60% included missing data levels, as these values provided very robust signal and provided a large difference in OCs (more than 2× for most datsets). All gap-removal strategies were also tested.



Mitochondrial taxa	10	20	30	40	50	60
Bn	6	7	7	7	8	8
Ca	9	12	12	12	12	12
Cg	9	12	12	12	12	12
Cm	5	6	7	7	7	7
Cv	8	11	11	11	12	12
Ha	9	11	11	11	11	11
Mj	9	11	12	13	13	13
Mp	9	12	12	12	13	13
Mv	9	11	11	11	11	11
Ng	5	7	7	7	7	7
No	9	9	9	9	9	9
Ov	9	9	9	9	9	9
Pa	6	6	6	6	6	6
Pl	9	10	10	10	10	10
Pp	4	6	6	6	6	6
Ps	9	10	10	10	10	10
Ra	9	13	18	18	19	20
Rs	9	11	11	11	11	11
Identified orthologous clusters	9	13	103	126	175	230

Figure 3 | Mitochondrial OCs identified in the 18 mitochondria phylogeny alignments as a function of included missing data and total identified OCs for the liberal gap-removal trees in Fig. 2d. Key: **Bn**, *Brassica napus* mt; **Ca**, *Chlorokybus atmophyticus* mt; **Cg**, *Chaetosphaeridium globosum* mt; **Cm**, *Cyanidioschyzon merolae* mt; **Cv**, *Chara vulgaris* mt; **Ha**, *Hemiselmis andersenii* mt; **Mj**, *Malawimonas jakobiformis* mt; **Mp**, *Marchantia polymorpha* mt; **Mv**, *Mesostigma viride* mt; **Ng**, *Naegleria gruberi* mt; **No**, *Nephroselmis olivacea* mt; **Ov**, *Ostreococcus tauri* mt; **Pa**, *Pseudendoclonium akinetum* mt; **Pl**, *Pylaiella littoralis* mt; **Pp**, *Porphyra purpurea* mt; **Ps**, *Phytophthora soja* mt; **Ra**, *Reclinomonas americana* mt; **Rs**, *Rhodomonas salina* mt.

The main effect of biased outgroups on the *Alphaproteobacteria* was to alter the topological relationships of the major clades established using the median outgroup. However, the major clades remained intact, with the exception of two trees using the low G+C biased outgroup (Fig. 4). These placed SAR11 outside of the *Rickettsiales*, although this topology had much weaker bootstrap support than trees which placed SAR11 as monophyletic with the *Rickettsiales* (Fig. 4a, SI tree files), and SAR11 was still placed in the

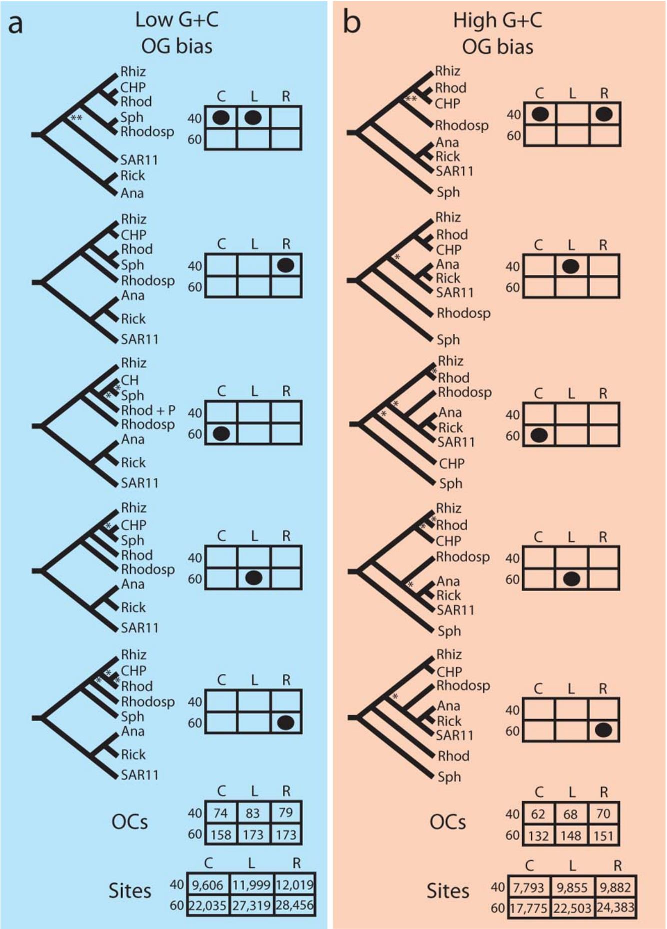


Figure 4 | *Alphaproteobacteria* tree topologies determined with outgroups biased by low (a) and high (b) G+C content. Each observed tree is shown with an associated table for the corresponding missing data percentage(s) (rows) and gap-removal parameter(s) (columns: C- conservative, L- liberal, R- remgaps). Overall included orthologous clusters (OCs) and amino acid sites for each dataset are specified at the bottom. * - bootstrap support for that node < 60%. OG- outgroup, Ana- *Anaplasmataceae*, Rick- *Rickettsiaceae*, Rhodosp- *Rhodospirillales*, Sph- *Sphingomonadales*, CHP- *Caulobacterales*, *Hyphomonadaceae*, *Parvularculales*, Rhod- *Rhodobacteraceae*.

Rickettsiales by the majority of trees. Trees with the high G+C biased outgroup no longer placed the *Rickettsiales* as the basal group to the tree, and instead pulled higher G+C content clades towards the outgroup as would be expected from such biasing. (Fig. 4b, SI tree files). While this bias disrupted the general placement of the *Rickettsiales* as the deepest branch of the *Alphaproteobacteria*, SAR11 still formed a monophyletic group with the *Rickettsiales*.

For the analyses which included mitochondrial sequences and varying numbers of alphaproteobacterial orders (Fig. 5), the results were mostly similar to those using the median outgroup, with the exception of the dataset in Fig. 5d, which did not agree with Fig. 2d. Including both high and low G+C bias, 75% of the trees grouped mitochondria with the *Rickettsiaceae*, instead of SAR11 as was shown with the median outgroup. Three of the trees did still group SAR11 and mitochondria together as in Fig. 2d. Fig. 5b low outgroup trees with 40% missing data were incongruent ("Inc") to the whole *Alphaproteobacteria* because they placed the *Anaplasmataceae* as basal to the *Rickettsiaceae*/SAR11/mitochondria group, but still supported the monophyly of SAR11 and mitochondria (SI tree files). The largest dataset including the *Rickettsiales*, *Rhodospirillales*, *Sphingomonadales*, *Caulobacterales*, and *Parvularculales* with the mitochondria again showed two separate tree topologies, distributed across both biases, and again those which supported the monophyly of SAR11 and mitochondria were better supported (Fig. 5e). Thus, the datasets in Fig. 5a,b, c, and e generally agreed with the median outgroup analyses shown in Fig. 2 regardless of the outgroup bias.

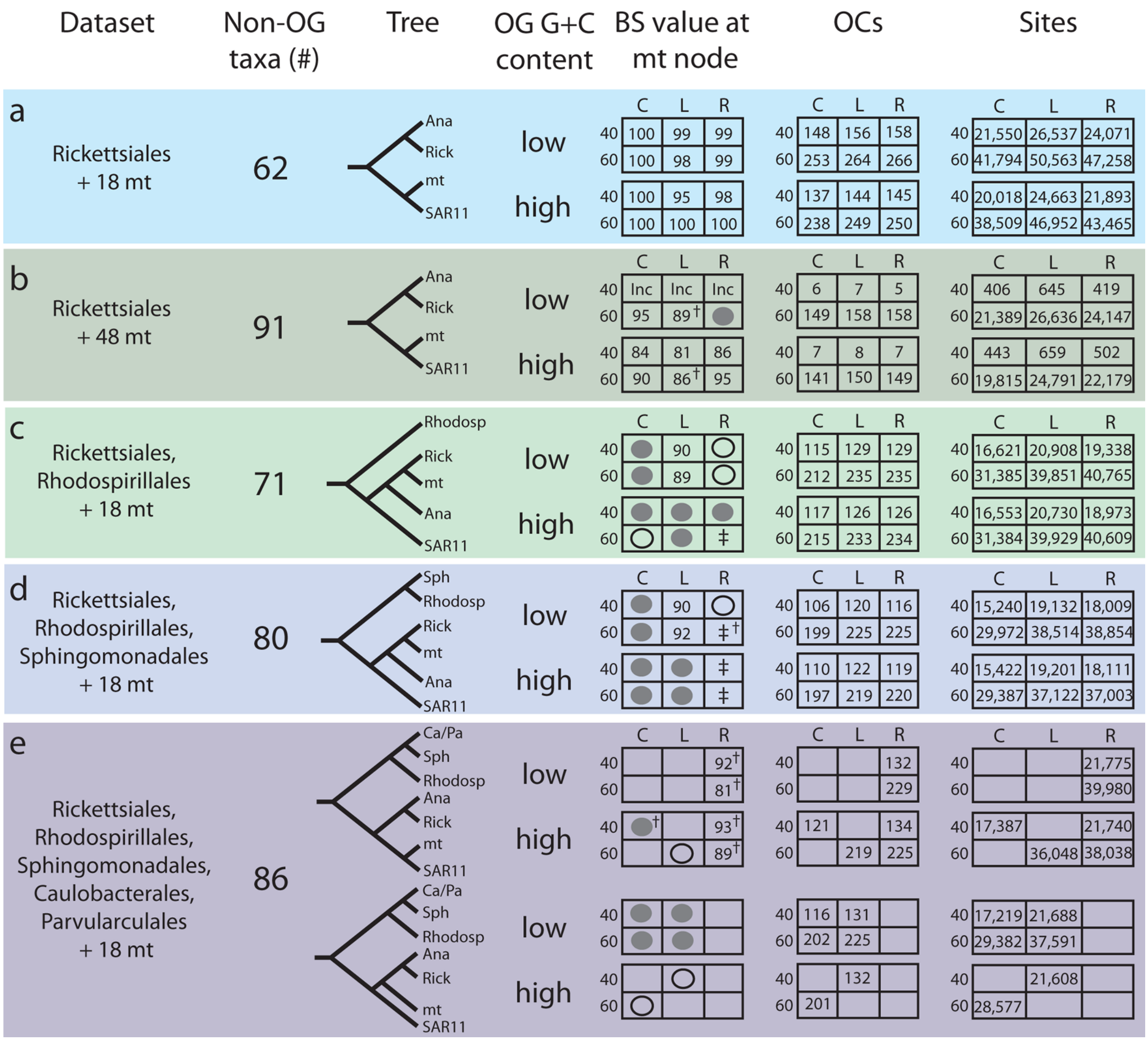
Discussion

We systematically tested phylogenomic relationships of SAR11 to *Alphaproteobacteria* and mitochondria using new, deeply branching SAR11 genomes we sequenced and over 60 additional genomes that were not available to previous studies[6]. The automated pipeline Hal allowed for methodic consistency across 216 separate phylogenomic analyses. There were some limitations to the pipeline: at low levels of allowed missing data, it found very few acceptable clusters across all taxa (data not shown), and there was an upper limit to the number of taxa that could be included with consistent results if the genomes were substantially different. Nevertheless, by combining the results of a variety of sampling and analysis perturbations, we have compiled a consistent view of the SAR11 clade within the *Alphaproteobacteria* and the relationship of mitochondria to this clade.

The observed position of SAR11 within the *Rickettsiales* and the *Alphaproteobacteria* agreed with a previous phylogenomic analysis of the *Alphaproteobacteria*[6], as well as other phylogenomic and 16S rRNA gene taxonomy studies[18,31,32,39]. Also, the consensus topology we observed for the entire *Alphaproteobacteria* supported the findings of Williams et al.[6] (Figs. 1a, S1). When taxon sampling was limited to the *Rickettsiales*, the observed relationships agreed with those of the complete *Alphaproteobacteria* tree (Fig. 1b), and all trees, regardless of gap-removal strategy, confirmed a *Rickettsia* phylogeny by Gillespie et al. created from 731 core genes[18] (SI tree files). Importantly, the use of biased outgroups also supported the placement of SAR11 in a monophyletic relationship with the *Rickettsiales* (Fig. 4). Thus, SAR11 placement within the *Rickettsiales* was mostly stable across perturbations of missing data, gap-removal strategy, taxon sampling, and outgroup G+C bias, making this topology the most likely given the currently available genomes. Further, based on comparisons of branch lengths between the SAR11 clade, *Rickettsiaceae*, and the *Anaplasmataceae* across the various trees (see an example in Figs. 6 and 7), and the 80% 16S rRNA gene sequence identity between HIMB59 (Group V) and HTCC1062 (Group I), SAR11 comprises enough phylogenetic diversity to be considered a distinct family, proposed here as *Pelagibacteraceae,* fam. nov.

When mitochondrial sequences were included in the analysis, the relationship of SAR11 to the *Rickettsiales* did not change, and most trees supported the monophyly of SAR11 and the mitochondria (Figs. 2, 5). The trees in Fig. 2c and Fig. 5c, d demonstrate the need for alteration of taxon sampling (see below) and attention to outgroup G+C bias when testing for phylogenetic signal. Although many of the trees from these two datasets provided support for the alternative monophyletic relationship of SAR11 and the *Rickettsiaceae*, they did not agree with the overall majority of trees. The dataset in Fig. 5d was not wholly consistent with itself, with only 25% of the trees in agreement when the same dataset was used without a biased outgroup (Fig. 2d). Also, the incongruence between Fig. 5d and Fig. 2d demonstrated that this selection of taxa was particularly sensitive to outgroup G+C bias. However, the remaining datasets were in good agreement regardless of outgroup G+C content (Figs. 2, 5). Overall, the analyses more frequently placed the mitochondria as monophyletic with SAR11 (85/128 trees), than either between SAR11, the *Rickettsiaceae*, and the *Anaplasmataceae* (11/128 trees), or monophyletic with the *Rickettsiaceae* (31/128 trees).

We chose our approach to test the robustness of tree topology to perturbations and thereby minimize the effects of horizontal gene transfer (HGT) and systematic error. The consistency of topologies and the high average bootstrap values across datasets, as well as the



Dataset	Non-OG taxa (#)	Tree	OG G+C content		BS value at mt node C	BS value at mt node L	BS value at mt node R	OCs C	OCs L	OCs R	Sites C	Sites L	Sites R
a Rickettsiales + 18 mt	62		low	40	100	99	99	148	156	158	21,550	26,537	24,071
				60	100	98	99	253	264	266	41,794	50,563	47,258
			high	40	100	95	98	137	144	145	20,018	24,663	21,893
				60	100	100	100	238	249	250	38,509	46,952	43,465
b Rickettsiales + 48 mt	91		low	40	Inc	Inc	Inc	6	7	5	406	645	419
				60	95	89†	●	149	158	158	21,389	26,636	24,147
			high	40	84	81	86	7	8	7	443	659	502
				60	90	86†	95	141	150	149	19,815	24,791	22,179
c Rickettsiales, Rhodospirillales + 18 mt	71		low	40	●	90	○	115	129	129	16,621	20,908	19,338
				60	●	89	○	212	235	235	31,385	39,851	40,765
			high	40	●	●	●	117	126	126	16,553	20,730	18,973
				60	○	●	‡	215	233	234	31,384	39,929	40,609
d Rickettsiales, Rhodospirillales, Sphingomonadales + 18 mt	80		low	40	●	90	○	106	120	116	15,240	19,132	18,009
				60	●	92	‡†	199	225	225	29,972	38,514	38,854
			high	40	●	●	‡	110	122	119	15,422	19,201	18,111
				60	●	●	‡	197	219	220	29,387	37,122	37,003
e Rickettsiales, Rhodospirillales, Sphingomonadales, Caulobacterales, Parvularculales + 18 mt	86		low	40			92†			132			21,775
				60			81†			229			39,980
			high	40	●†		93†	121		134	17,387		21,740
				60		○	89†		219	225		36,048	38,038
			low	40	●	●		116	131		17,219	21,688	
				60	●	●		202	225		29,382	37,591	
			high	40		○			132			21,608	
				60	○			201			28,577		

Figure 5 | Summary of phylogenomic analyses with mitochondrial sequences using biased outgroups. Data is divided by each dataset and corresponding tree topology (a–f). Internal tables for each dataset show the relevant information for each individual tree output, grouped according to G+C bias of the outgroup taxa. Rows are allowed missing data (40 and 60%), columns are gap-removal strategy (C- conservative, L- liberal, R- remgaps). Bootstrap support values ≥ 80% are reported, grey circles indicate a value between 79 and 51%, open circles ≤ 50%, Inc- incongruence in non-mitochondrial branch order compared with trees built without mitochondrial sequences and other studies. † - HIMB59 grouped with the mitochondria. ‡ - mitochondria grouped with the SAR11 clade. Other abbreviations: mt- mitochondria, Ana- *Anaplasmataceae*, Rick- *Rickettsiaceae*, Rhodosp- *Rhodospirillales*, Sph- *Sphingomonadales*, Ca- *Caulobacterales*, Pa- *Parvularculales*, OG- outgroup, BS- bootstrap, OCs- orthologous clusters.

large number of included OCs in most of the trees means that the effect of HGT was probably overcome with increasing amounts of sequence signal[38]. However, systematic error cannot be overcome by such means. While phylogenomic analysis can dramatically reduce stochastic error, the increase in characters does not relieve the problems of systematic errors, which remain constant[26,27,40,30]. Using the terminology of Philippe et al.[26], systematic errors (non-phylogenetic signal) can be broken into three categories: rate signal, compositional signal, and heterotachous signal (for detailed reviews see Philippe et al.[26] and Delsuc et al.[27]). Briefly, rate signal arises from variation in the evolutionary rates of different taxa, which can lead to long-branch attraction (LBA) artifacts of highly divergent taxa. Compositional signal is the result of G+C content bias between organisms in the analysis and can result in organisms of similar G+C content being artificially grouped together. Heterotachous signal comes from site and/or gene-specific rate variation within and between organisms.

All of these signal biases were addressed in our analysis. We made use of the CAT model of rate heterogeneity in RAxML specifically to deal with LBA[29], and created separate amino acid substitution models for each OC that was included in the tree-building process. Alteration of allowed missing data, which changed both the number and the type of OCs used throughout the analysis, was expected to limit the effects of heterotachous signal. Another potential problem for the phylogenetic placement of mitochondria and SAR11 was that in general, *Rickettsiales*, including SAR11, have much lower G+C content than other *Alphaproteobacteria* (Fig. 6). In our analyses, organisms with low G+C content were scattered throughout the alphaproteobacterial tree, and did not consistently branch together (Fig. 6), even in the face of substantial outgroup bias (Figs. 4, 5, SI tree files). Also, the monophyly of the mitochondria across datasets is noteworthy considering the wide G+C range among mitochondria, especially considering the presence of the *Anaplasma*, which have a much higher G+C content than the other *Rickettsiales* (Fig. 7). Analysis of the mitochondrial sequences with the *Rickettsiales* alone addressed both LBA and compositional signal error by removing higher G+C content organisms and those with shorter branches.

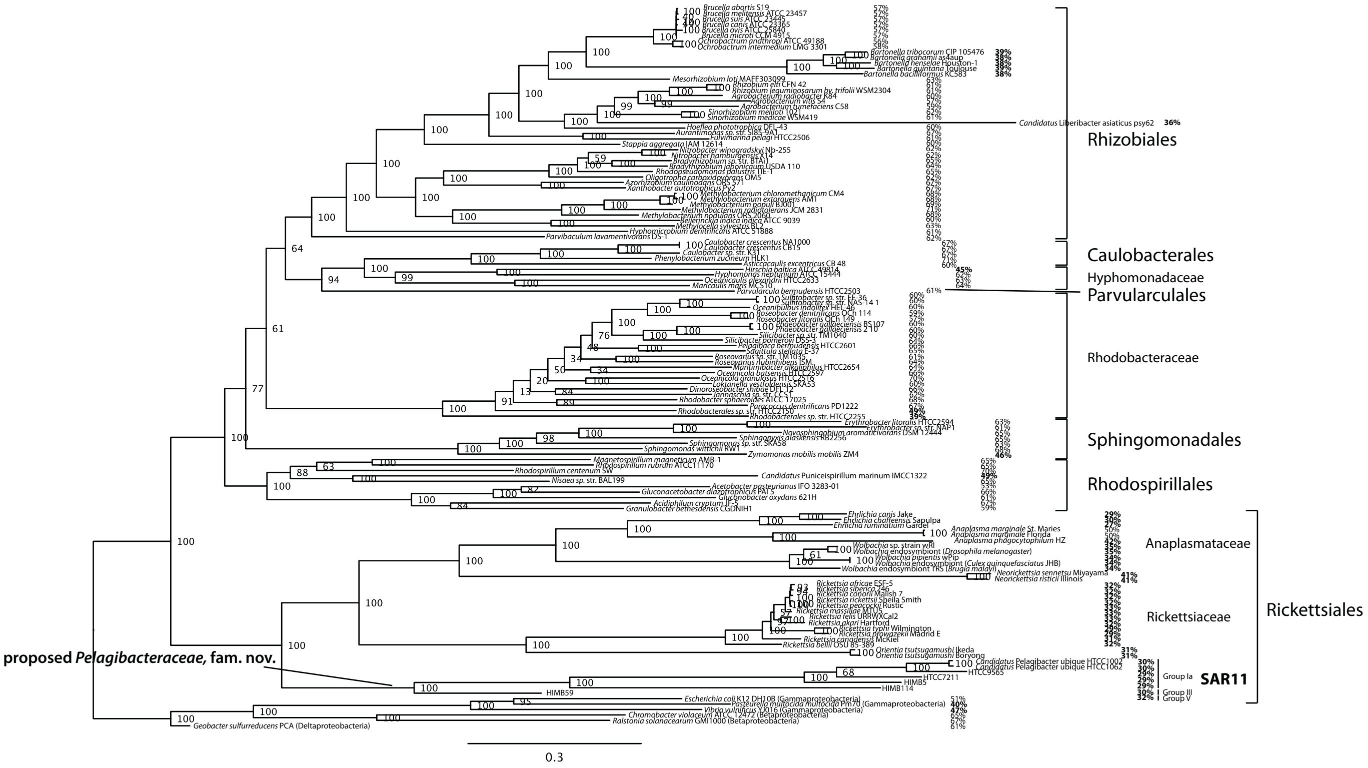

Rhizobiales
Caulobacterales
Hyphomonadaceae
Parvularculales
Rhodobacteraceae
Sphingomonadales
Rhodospirillales
Anaplasmataceae
Rickettsiaceae
Rickettsiales
SAR11
Group Ia
Group III
Group V
proposed Pelagibacteraceae, fam. nov.
Candidatus Liberibacter asiaticus psy62 36%
Candidatus Pelagibacter ubique HTCC1002
Candidatus Pelagibacter ubique HTCC1062
HTCC7211
HTCC9565
HIMB5
HIMB114
HIMB59
Escherichia coli K12 DH10B (Gammaproteobacteria)
Pasteurella multocida multocida Pm70 (Gammaproteobacteria)
Vibrio vulnificus YJ016 (Gammaproteobacteria)
Chromobacter violaceum ATCC 12472 (Betaproteobacteria)
Ralstonia solanacearum GMI1000 (Betaproteobacteria)
Geobacter sulfurreducens PCA (Deltaproteobacteria)
0.3

Figure 6 | Example tree of the *Alphaproteobacteria* using the median outgroup, 60% missing data included, liberal gap-removal. G+C % is indicated for each taxon, numbers below 50% are in bold. Scale bar indicates 0.3 changes per position. Bootstrap values are indicated at nodes.

The observation that biased outgroups and the addition of higher G+C *Alphaproteobacteria* sequences generally did not disrupt the topology, across a variety of included orthologous clusters, provides further support for the monophyly of SAR11 and mitochondria.

While the various techniques employed here are helpful to uncovering true phylogenetic signal, taxon sampling is perhaps the most important countermeasure to systematic errors[26,27,30].The effect of taxon sampling was clearly shown by the aberrant trees in Fig. 2c, which had high bootstrap support but were incongruent with the remaining trees. We were able to demonstrate consistency across large sampling differences, both in taxa and included OCs. The effect of allowing more missing data was to increase the number of OCs used in both the overall analyses as well as per taxa, which resulted in phylogenies that were more robust, in agreement with previous work[24,25] (and references therein). This had significant practical impact on our study because inclusion of missing data also allowed the inclusion of taxa with less coding sequence - specifically, mitochondria.

Yet, mitochondrial genome size is the result of massive gene transfer from the original endosymbiont to the nuclear genome, and therefore the signal which we made use of for this study has to be viewed as a subset of available information. Many of the original proto-mitochondrion genes retain enough phylogenetic signal to identify them as having come from ancestral *Alphaproteobacteria*, in spite of the fact that Eukaryotic substitution of analogous genes has occurred[13,41,42] (and references within). An initial study of these nuclear encoded mitochondrial genes showed close relationships to organisms throughout the *Alphproteobacteria*[22]. In light of the results we present here, close phylogenetic scrutiny of nuclear encoded genes that originated from mitochondria, in the context of new *Alphaproteobacteria* sequences such as SAR11, would appear to be a promising direction for future research on this topic. Also, obtaining sequences from deeper branching *Alphaproteobacteria*, and examination of these relationships with alternative methods, such as supertrees, or sequence-independent features like gene order, rare genomic changes, and other technologies currently being developed and tested (see Delsuc et al.[27] for a review), could provide additional perspective on the evolutionary links between SAR11, mitochondria, *Rickettsiales*, and other *Alphaproteobacteria*.

While the variety of issues discussed above warrant caution, our findings nevertheless provide support for an intriguing new hypothesis: that mitochondria evolved from free living ancestors of the SAR11 clade, rather than from a lineage of intracellular organisms, as previously suggested[10,12]. Prior work indicated that the ancestral mitochondrion was most likely an aerobic heterotroph[2,41] with respiratory genes closely related to those in *Rickettsiaceae* (reviewed in[10]). The physiology of SAR11 organisms is consistent with the proposed proto-mitochondria: all strains are heterotrophic, obligate aerobes, with predicted aerobic electron transport chains like those in mitochondria, composed of NADH dehydrogenases, fumarate reductases, ubiquinones, cytochrome bc_1 complexes, cytochrome *c*, cytochrome *c* oxidases, and ATP synthases. Several of these genes were among the orthologous clusters used to construct the phylogenies, along with ribosomal proteins (e.g., Table S4), and thus support the close evolutionary relationship between these pathways/genes in SAR11 organisms and the mitochondria.

Modern SAR11 cells have small genomes imposed by selection for efficient replication in nutrient limited systems[17], which have evolved unusual nutritional requirements that would pre-adapt them to closer dependence on other organisms[43–45]. While many explanations have been offered for the singular origin of mitochondria[2] (and references therein), the findings we present support a simple but plausible scenario: mitochondria originated in the aerobic ocean, involving a lineage that repeatedly colonized other species, and ultimately gave rise to the largest free living bacterial populations known. Contemporary marine systems contain abundant Archaea and Eukarya, both theorized hosts for the original mitochondrial symbiosis[2]. We cannot know if the current conditions are similar to those at the time of the mitochondrial and SAR11 ancestor, but we can add to the scientific discussion by postulating that it was likely a free-living marine organism, and focus interest broadly on the

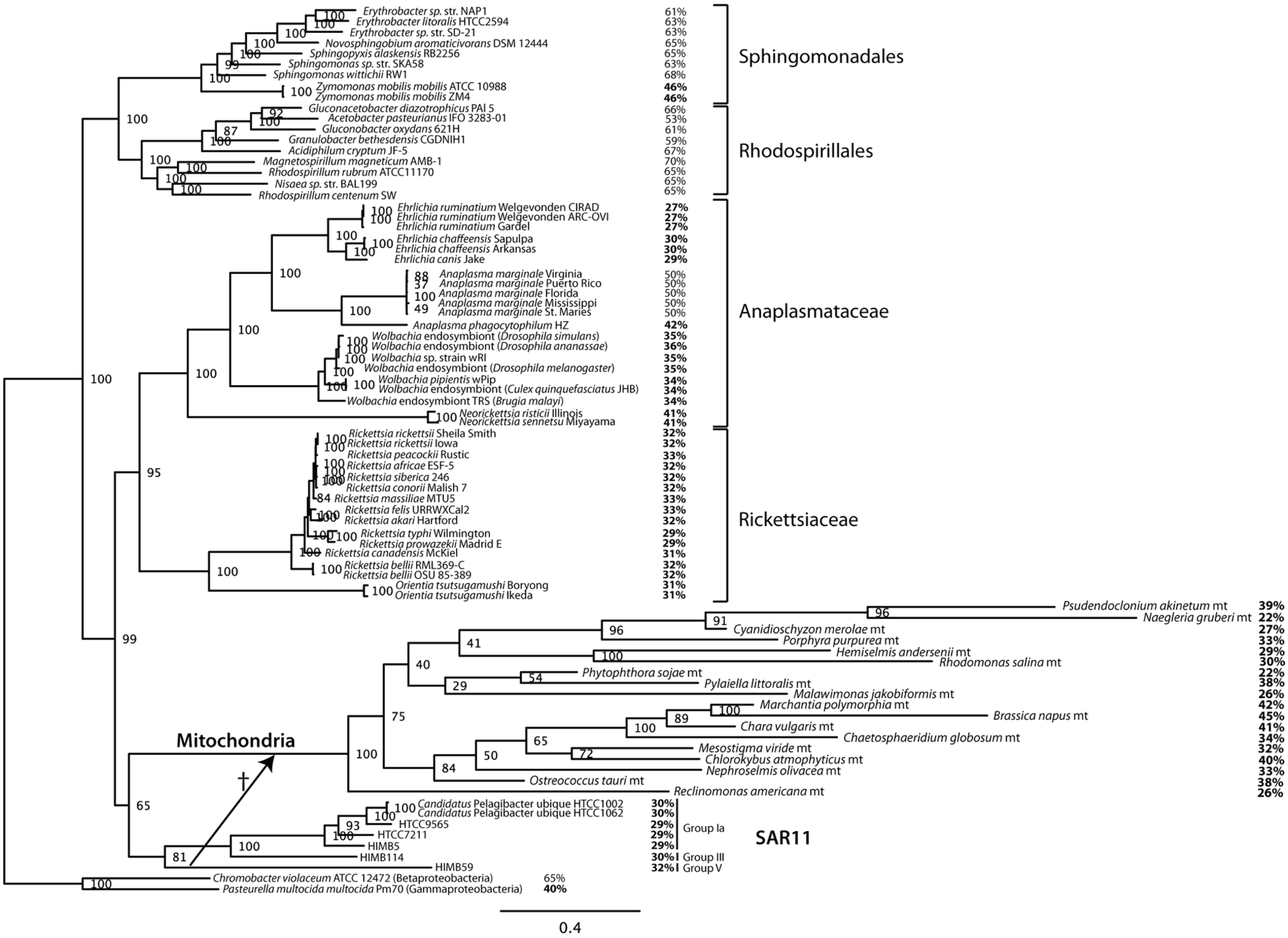

Erythrobacter sp. str. NAP1 61%
Erythrobacter litoralis HTCC2594 63%
Erythrobacter sp. str. SD-21 63%
Novosphingobium aromaticivorans DSM 12444 65%
Sphingopyxis alaskensis RB2256 65%
Sphingomonas sp. str. SKA58 63%
Sphingomonas wittichii RW1 68%
Zymomonas mobilis mobilis ATCC 10988 46%
Zymomonas mobilis mobilis ZM4 46%
Sphingomonadales
Gluconacetobacter diazotrophicus PAl 5 66%
Acetobacter pasteurianus IFO 3283-01 53%
Gluconobacter oxydans 621H 61%
Granulobacter bethesdensis CGDNIH1 59%
Acidiphilum cryptum JF-5 67%
Magnetospirillum magneticum AMB-1 70%
Rhodospirillum rubrum ATCC11170 65%
Nisaea sp. str. BAL199 65%
Rhodospirillum centenum SW 65%
Rhodospirillales
Ehrlichia ruminatium Welgevonden CIRAD 27%
Ehrlichia ruminatium Welgevonden ARC-OVI 27%
Ehrlichia ruminatium Gardel 27%
Ehrlichia chaffeensis Sapulpa 30%
Ehrlichia chaffeensis Arkansas 30%
Ehrlichia canis Jake 29%
Anaplasma marginale Virginia 50%
Anaplasma marginale Puerto Rico 50%
Anaplasma marginale Florida 50%
Anaplasma marginale Mississippi 50%
Anaplasma marginale St. Maries 50%
Anaplasma phagocytophilum HZ 42%
Wolbachia endosymbiont (Drosophila simulans) 35%
Wolbachia endosymbiont (Drosophila ananassae) 36%
Wolbachia sp. strain wRI 35%
Wolbachia endosymbiont (Drosophila melanogaster) 35%
Wolbachia pipientis wPip 34%
Wolbachia endosymbiont (Culex quinquefasciatus JHB) 34%
Wolbachia endosymbiont TRS (Brugia malayi) 34%
Neorickettsia risticii Illinois 41%
Neorickettsia sennetsu Miyayama 41%
Anaplasmataceae
Rickettsia rickettsii Sheila Smith 32%
Rickettsia rickettsii Iowa 32%
Rickettsia peacockii Rustic 33%
Rickettsia africae ESF-5 32%
Rickettsia siberica 246 32%
Rickettsia conorii Malish 7 32%
Rickettsia massiliae MTU5 33%
Rickettsia felis URRWXCal2 33%
Rickettsia akari Hartford 32%
Rickettsia typhi Wilmington 29%
Rickettsia prowazekii Madrid E 29%
Rickettsia canadensis McKiel 31%
Rickettsia bellii RML369-C 32%
Rickettsia bellii OSU 85-389 32%
Orientia tsutsugamushi Boryong 31%
Orientia tsutsugamushi Ikeda 31%
Rickettsiaceae
Psudendoclonium akinetum mt 39%
Naegleria gruberi mt 22%
Cyanidioschyzon merolae mt 27%
Porphyra purpurea mt 33%
Hemiselmis andersenii mt 29%
Rhodomonas salina mt 30%
Phytophthora sojae mt 22%
Pylaiella littoralis mt 38%
Malawimonas jakobiformis mt 26%
Marchantia polymorphia mt 42%
Brassica napus mt 45%
Chara vulgaris mt 41%
Chaetosphaeridium globosum mt 34%
Mesostigma viride mt 32%
Chlorokybus atmophyticus mt 40%
Nephroselmis olivacea mt 33%
Ostreococcus tauri mt 38%
Reclinomonas americana mt 26%
Mitochondria
†
Candidatus Pelagibacter ubique HTCC1002 30%
Candidatus Pelagibacter ubique HTCC1062 30%
HTCC9565 29%
HTCC7211 29%
HIMB5 29%
Group Ia
HIMB114 30% Group III
HIMB59 32% Group V
SAR11
Chromobacter violaceum ATCC 12472 (Betaproteobacteria) 65%
Pasteurella multocida multocida Pm70 (Gammaproteobacteria) 40%
0.4

Figure 7 | Example tree of mitochondrial taxa included with the *Rickettsiales*, *Rhodospirillales*, and *Sphingomonadales* using the median outgroup (Fig. 2d) showing all taxa, 50% missing data included, liberal gap removal. Dagger indicates the single taxon instability of HIMB59, observed in several trees throughout the datasets in Fig. 2. G+C % is indicated for each taxon, numbers below 50% are in bold. Scale bar indicates 0.4 changes per position. Bootstrap values are indicated at nodes.

mitochondria, the SAR11 clade, and *Rickettsiales* to identify shared properties that led to their successful divergence.

Methods

We recently sequenced five additional SAR11 genomes from three phylogenetically distinct clades, corresponding to the subgroups Ia, III, and a new, distantly related subgroup (Group V)[46] (Figs. 6 & 7), which were annotated consistently with others used in this study by IMG[47] (http://img.jgi.doe.gov/cgi-bin/m/main.cgi). We made exclusive use of publicly available bacterial genome sequences from the IMG database to control for gene-calling methodological differences across databases. The 127 taxa used in the whole alphaproteobacterial dataset (Table S1) were a subset of the total available sequences chosen to represent all sequenced genera. Given that some species are highly overrepresented in the database, we omitted several taxa from these overrepresented groups. Datasets with mitochondrial sequences included most available genome sequences from the respective *Alphaproteobacteria* orders (Table S2). Mitochondrial sequences (Table S3) were obtained from NCBI (http://www.ncbi.nlm.nih.gov/). The sequences were analyzed using Hal, an automated pipeline for phylogenetic analysis of genomic protein sequence data.

The Hal pipeline[23,48] (http://aftol.org/pages/Halweb3.htm; http://sourceforge.net/projects/bio-hal/) consists of a set of Perl scripts that automates a series of phylogenomic analyses using existing software and sequence analysis programs, and was executed on a 64-bit Linux cluster operating Red Hat Linux 3.2.3, Linux version 2.4.21. For this study, the Hal pipeline directed the following analyses: Protein sequences were imported in fasta format from sequenced genomes listed in Table S1–3 and subjected to an all-vs-all BLASTP with the output E-values provided to the program MCL[49]. Using a Markov Clustering algorithm, MCL grouped proteins into orthologous clusters (OCs) as function of the all-vs-all BLASTP E-values. Clustering was executed across a range of stringencies (inflation parameters 1.1 – 5.0) and OCs were filtered for any redundant clusters (clusters found at more than one inflation parameter), clusters containing more than one protein per genome (multi-copy OC), and clusters containing proteins whose best reciprocal BLAST hit was outside of the cluster (clusters more likely to contain paralogs). Clusters were also filtered across a range of missing values settings (10–60%) whereby a cluster may contain only one protein per genome but a defined percentage of the genomes may be missing from the cluster. This function allows for the incorporation for more single-copy clusters when such proteins are either missing from a given genome or missing from a genome annotation.

Protein sequences for accepted OCs were extracted from their respective genome fasta sequence files and aligned using MUSCLE[50] with default settings. To accommodate for problematic regions of the alignments, three separate alignments were created for each protein alignment: one removing all gap-containing columns (remgaps) and two removing problematic regions of the alignments based on the default conservative (Glocks-con) and liberal (Gblocks-lib) options of the program GBlocks[51]. Best models of amino acid substitution for each protein alignment were estimated using ProtTest[52]. Individual protein alignments were concatenated into one super-alignment and analyzed using RAxML[53] with the PROTCAT setting for the rate model and with each protein partition of the super-alignment assigned its best model of amino acid substitution. Nodal support was estimated based on 100 bootstrap replications using the rapid bootstrapping option as implemented in RAxML. For each analysis, three phylogenetic trees were generated, representing the three super-alignments (remgaps, GBlocks-con and GBlocks-lib) produced as part of the automated alignment routine.

1. Gray, M. W., Lang, B. F. & Burge, G. Mitochondria of Protists. *Annu. Rev. Genet.* **38**, 477–524 (2004).
2. Embley, T. M. & Martin, W. Eukaryotic evolution, changes and challenges. *Nature* **440**, 623–630 (2006).
3. Koonin, E. V. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol.* **11**, 209–221 (2010).

4. Fitzpatrick, D. A., Creevey, C. J. & McInerney, J. O. Genome Phylogenies Indicate a Meaningful α-Proteobacterial Phylogeny and Support a Grouping of the Mitochondria with the *Rickettsiales*. *Mol. Biol. Evol.* **23**, 74–85 (2005).
5. Andersson, S. G. E. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–143 (1998).
6. Williams, K. P., Sobral, B. W. & Dickerman, A. W. A Robust Species Tree for the *Alphaproteobacteria*. *J. Bacteriol.* **189**, 4578–4586 (2007).
7. Wu, M. *et al.* Phylogenomics of the Reproductive Parasite *Wolbachia pipientisw* Mel: A Streamlined Genome Overrun by Mobile Genetic Elements. *PLoS Biol.* **2**, 0327–0341 (2004).
8. Esser, C., Martin, W. & Dagan, T. The origin of mitochondria in light of a fluid prokaryotic chromosome model. *Biol. Letters* **3**, 180–184 (2007).
9. Ettema, T. J. G. & Andersson, S. G. E. The α-proteobacteria: the Darwin finches of the bacterial world. *Biol. Letters* **5**, 429–432 (2009).
10. Andersson, S. G. E., Karlberg, O., Canback, B. & Kurland, C. On the origin of mitochondria: a genomics perspective. *Philos. T. R. Soc. B.* **358**, 165–179 (2003).
11. Gray, M. W., Burger, G. & Lang, B. F. The origin and early evolution of mitochondria. *Genome Biol.* **2**, 1–5 (2001).
12. Emelyanov, V. V. Mitochondrial connection to the origin of the eukaryotic cell. *Eur J Biochem* **270**, 1599–1618 (2003).
13. Kurland, C. G. & Andersson, S. G. E. Origin and Evolution of the Mitochondrial Proteome. *Microbiol. Mol. Biol. Rev.* **64**, 786–820 (2000).
14. Esser, C. *et al.* A Genome Phylogeny for Mitochondria Among α-Proteobacteria and a Predominantly Eubacterial Ancestry of Yeast Nuclear Genes. *Mol. Biol. Evol.* **21**, 1643–1660 (2004).
15. Morris, R. *et al.* SAR 11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810 (2002).
16. Stingl, U., Tripp, H. J. & Giovannoni, S. J. Improvements of high-throughput culturing yielded novel SAR11 strains and other abundant marine bacteria from the Oregon coast and the Bermuda Atlantic Time Series study site. *ISME J.* **1**, 361–371 (2007).
17. Giovannoni, S. J. *et al.* Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science* **309**, 1242–1245 (2005).
18. Gillespie, J. J. *et al. Rickettsia* Phylogenomics: Unwinding the Intricacies of Obligate Intracellular Life. PLoS ONE **3**, 1–34 (2008).
19. Wilhelm, L., Tripp, H. J., Givan, S., Smith, D. P. & Giovannoni, S. J. Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biol Direct* **2**, (2007).
20. Brayton, K. A. *et al.* Complete genome sequencing of *Anaplasma marginale* reveals that the surface is skewed to two superfamilies of outer membrane proteins. *Proc. Nat. Acad. Sci. U.S.A.* **102**, 844–849 (2005).
21. Collins, N. E. *et al.* The genome of the heartwater agent *Ehrlichia ruminantium* contains multiple tandem repeats of actively variable copy number. *Proc. Nat. Acad. Sci. U.S.A.* **102**, 838–843 (2005).
22. Atteia, A. *et al.* A Proteomic Survey of *Chlamydomonas reinhardtii* Mitochondria Sheds New Light on the Metabolic Plasticity of the Organelle and on the Nature of the α-Proteobacterial Mitochondrial Ancestor. *Mol. Biol. Evol.* **26**, 1533–1548 (2009).
23. Robbertse, B., Yoder, R. J., Boyd, A., Reeves, J. B. & Spatafora, J. W. Hal: an Automated Pipeline for Phylogenetic Analyses of Genomic Data. *PLoS Curr. Tree of Life* **4**, (2011).
24. Philippe, H. *et al.* Phylogenomics of Eukaryotes: Impact of Missing Data on Large Alignments. *Mol Biol Evol* **21**, 1740–1752 (2004).
25. Wiens, J. J. Missing data and the design of phylogenetic analyses. *J. Biomed. Inform.* **39**, 34–42 (2006).
26. Philippe, H., Delsuc, F., Brinkmann, H. & Lartillot, N. PHYLOGENOMICS. *Annu. Rev. Ecol. Evol. Syst.* **36**, 541–562 (2005).
27. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the Reconstruction of the Tree of Life. *Nat Rev Genet* **6**, 361–375 (2005).
28. Lartillot, N. & Philippe, H. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
29. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7**, 1–14 (2007).
30. Philippe, H. *et al.* Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biol* **9**, e1000602 (2011).
31. Lee, K. The hierarchical system of the '*Alphaproteobacteria*': description of *Hyphomonadaceae* fam. nov., *Xanthobacteraceae* fam. nov. and *Erythrobacteraceae* fam. nov. *Int. J. Syst. Evol. Micr.* **55**, 1907–1919 (2005).
32. Logares, R., Brate, J., Heinrich, F., Shalchian-Tabrizi, K. & Bertilsson, S. Infrequent Transitions between Saline and Fresh Waters in One of the Most Abundant Microbial Lineages (SAR11). *Mol. Biol. Evol.* **27**, 347–357 (2010).
33. Gray, M. W., Burger, G. & Lang, B. F. Mitochondrial Evolution. *Science* **283**, 1476–1481 (1999).
34. Lang, B. R. *et al.* An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387**, 493–497 (1997).
35. Hampl, V. *et al.* Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups" *PNAS* **106**, 3859–3864 (2009).
36. Burki, F., Shalchian-Tabrizi, K. & Pawlowski, J. Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. *Biol. Letters* **4**, 366–369 (2008).
37. Brown, W. M., George Jr., M. & Wilson, A. C. Rapid evolution of animal mitochondrial DNA. *Proc. Nat. Acad. Sci. U.S.A.* **76**, 1967–1971 (1979).
38. Shearer, T. L., Oppen, M. V., Romano, S. & Worheide, G. Slow mitochondrial DNA sequence evolution in the Anthozoa (Cnidaria). *Mol. Ecol.* **11**, 2475–2487 (2002).
39. Rappe, M. S., Connon, S. A., Vergin, K. L. & Giovannoni, S. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**, 630–633 (2002).
40. Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. Phylogenomics: the beginning of incongruence? *Trends Genet.* **22**, 225–231 (2006).
41. Gabaldon, T. & Huynen, M. A. Reconstruction of the Proto-Mitochondrial Metabolism. *Science* **301**, 609 (2003).
42. Gross, J. & Bhattacharya, D. Mitochondrial and plastid evolution in eukaryotes: an outsiders' perspective. *Nat. Rev. Genet.* **10**, 495–505 (2009).
43. Tripp, H. J. *et al.* SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452**, 741–744 (2008).
44. Tripp, H. J. *et al.* Unique glycine-activated riboswitch linked to glycine-serine auxotrophy in SAR11. *Environ. Microbiol.* **11**, 230–238 (2009).
45. Schwalbach, M. S., Tripp, H. J., Steindler, L., Smith, D. P. & Giovannoni, S. J. The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. *Environ. Microbiol.* **12**, 490–500 (2010).
46. Moeseneder, M. M., Arrieta, J. M. & Herndl, G. J. A comparison of DNA- and RNA-based clone libraries from the same marine bacterioplankton community. *FEMS Microbiol. Ecol.* **51**, 341–352 (2005).
47. Markowitz, V. M. *et al.* IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* **36**, D534–D538 (2008).
48. Robbertse, B., Reeves, J. B., Schoch, C. L. & Spatafora, J. W. A phylogenomic analysis of the Ascomycota. *Fungal Gent. and Biol.* **43**, 715–725 (2006).
49. van Dongen, S. Graph Clustering by Flow Simulation. (2000).
50. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).
51. Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
52. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).
53. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).

Acknowledgements

We thank Christopher M. Sullivan and the Oregon State University Center for Genome Research and Biocomputing (CGRB) for management of and troubleshooting for the bioinformatics infrastructure. Sequencing, assembly, and annotation of SAR11 genomes was supported by the Gordon and Betty Moore Foundation Marine Microbiology Initiative, as part of its Marine Microbial Sequencing Project (http://www.moore.org/marinemicro), by an individual investigator award to SJG, and a grant from the National Science Foundation Science and Technology Center, Center for Microbial Oceanography: Research and Education(Award No. EF0424599)to MSR. This material is based upon work supported by the National Science Foundation under Award No. DBI-1003269 to JCT.

Author contributions

AB assisted with bioinformatics analysis, program writing, and created the SI website, MJH, JG, and PC cultivated cells, extracted DNA for genome sequencing, and aided in annotation, MSR provided three SAR11 genome sequences and assisted in study design, RJY wrote the Hal scripts and assisted with bioinformatics analysis, BR oversaw Hal development and implementation, JWS oversaw Hal development and implementation and assisted in study design, JCT and SJG designed the study, SJG and JWS assisted in writing the paper, JCT wrote the paper/created the figures, compiled the data, and performed the bioinformatics analysis. All authors discussed the results and commented on the paper.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/scientificreports

All tree files, super alignments, and model files used in the study are available at: http://sar11-phylogenomics.cgrb.oregonstate.edu.

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Thrash, J.C. *et al.* Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci. Rep.* **1**, 13; DOI:10.1038/srep00013 (2011).