An ectosymbiosis-based mechanism of eukaryogenesis

J. Tze-Fei Wong*, Xi Long and Hong Xue

Division of Life Science, Hong Kong University of Science and Technology, Hong Kong, China.

*Corresponding author: J. Tze-Fei Wong, bcjtw@ust.hk

Abstract

The findings of a deep branching Microsporidia clade on the SSU rRNA tree, and diversity of sequence motifs in eukaryotic Hsp70s rendered invalid the endosymbiosis-first theory that mitosome- and hydrogenosome-containing amitochondriate eukaryotes (AMIs) arose from mitochondriate eukaryotes (MTEs) via reductive evolution. Instead, evidence of widespread ectosymbioses indicated that eukaryogenesis was started by an archaeal parent via its acquisition of archaeal proteins through 'accelerated gene adoption', and bacterial proteins from ectosymbionts including a clostridial ectosymbiont that supplied its [Fe] hydrogenase and pyruvate:ferredoxin oxidoreductase genes to the AMIs. Subsequent endosymbiosis with Tistrella gave rise to mitochondria with the participation of other alphaproteobacteria. The high frequencies of top similarity bitscores displayed by Giardia, Edhazardia and Trichomonas toward *Aciduliprofundum boonei* (Abo) pertaining to the enzymes of DNA biology, far surpassing the frequencies toward any Asgard or TACK archaeon, established Abo as the source of these enzymes in eukaryotes, and the archaeal parent of Eukarya.

For over a century endosymbiosis has been regarded as the basis for eukaryogenesis, with recent focus placed on an archaeal parent-alphaproteobacterial endosymbiosis[1]. It was pointed out based on the SSU rRNA sequence of *Vairimorpha necatrix* that Microsporidia branched very early from the other eukaryotes on the SSU rRNA tree[2]. The evolution of MTEs from the AMIs was postulated by the 'Archezoa' theory based on the autogenous evolution of a primitive phagotrophic archezoan[3], and by the 'crown' theory based on the locations of MTE branches above the AMI branches on the SSU rRNA tree[4]. However, the discoveries of typical mitochondrial proteins in the AMIs gave rise to the 'endosymbiosis first' theory that the presence of mitochondrial proteins in AMIs implies the harboring of mitochondria or their endosymbiotic ancestors by the AMIs in the past[5,6]. As a result, the proposal of Lokiarchaeota being a suitable archaeal parent was held up by the missing mechanism for importing an alphaproteobacterial endosymbiont[7], and the proposal of the Aciduliprofundum archaeons being appropriate archaeal parents would not be implementable if these small archaeons are required to accommodate an upfront endosymbiont[8]. In addition, searches have not yielded a consensual phylogenetic location for the Last Eukaryotic Common Ancestor (LECA)[9-12], and autogenous development of eukaryotes has not given a better fit to omics data than endosymbiosis[13]. In view of these difficulties, the present study has been directed to an investigation of the possibility of LECA formation through ectosymbiosis instead of endosymbiosis.

Primitivity of Microsporidia

Rejection of the primitivity of AMIs despite their low branchings on SSU rRNA trees stemmed from the supposition that their long branches may be due to an accelerated rate of evolution, making them susceptible to the artifact of long branch attraction[14]. However, the tree in Figure 1, built using the maximum likelihood method that minimized long branch attraction artifacts[15], not only positioned the AMIs as low branchers, but also grouped the Microsporidia species into a clade.

The branches of the Amoebozoa AMI species Ein and Enh (see Table 1 for abbreviations of species names), which reductively evolved from Amoebozoa MTE species[16], were indeed much longer than the MTE branches of Asu, Ddi and Dpu. However, despite their long branches, Ein and Enh were included in the same Amoebozoa clade as Asu, Ddi and Dpu on the tree. Both Mon which lost the mitochondria or mitochondria-related organelles (MROs) it once possessed[17], and Cpa which has degenerate mitochondria[18] also did not move to low-branching positions on the tree on account of their experience of reductive evolution. The nucleomorphs with their vestigial genomes likewise stayed with their ancestral red or green algae on the tree[19]. Accordingly, there was little basis to assume that the clade of nine AMI Microsporidia species had moved to a low branching position on the tree because of their reductive evolution from fungi[20], away from the modern fungal Sce, Spo, Lbi, Ncr, Mci, Bde and Spu sitting at the top of the tree. Among the Microsporidia, Ecu with its 2.9 Mb genome was much more reduced in size compared to some of the other microsporidians[21], but it was located within the same clade as other microsporidians on either the SSU rRNA tree (Figure 1) or a tree for 53 concatenated proteins[22]. Therefore, the evidence for Microsporidia being an independently evolved AMI group rather than products of reductive evolution, in contradiction to any theory that requires the derivation of all AMIs from MTEs, was strong.

The motifs of eukaryotic Hsp70 sequences also rendered uncertain the evidence for the endosymbiosis-first theory founded on the affinities between mitochondrial proteins and their counterparts in the AMIs[23]. Figure 2 shows the Group-I EKP/DKP/GKP and Group-II NGDAWV motifs, which were both absent from the archaeal sequences of Abo, Lok and Hgi but resembled bacterial sequences: the EKP-motif resembled that of Aba, the DKP-motif resembled those of Cur and Kae, and the GKP-motif resembled those of Mca and Eco. The NGDAWV-motif was found in not only the alphaproteobacteria Tis, Atu and Rru, but also the gammaproteobacterium Eco and the deltaproteobacterium Pel. Such diversified motifs in both AMIs and MTEs were in keeping with the profuse variety of prokaryotic proteins found in the proteomes of eukaryotes[8,24-27], but neither the diversity nor the apparent paucity or lack of the EKP motif in the MTEs was consistent with the origin of all MTE and AMI Hsp70s from a single alphaproteobacterial endosymbiont.

Ectosymbiotic mechanism of eukaryogenesis

In view of the unlikelihood of endosymbiosis-first, and the intrinsic compatibility of ectosymbiosis with AMI primitivity and Hsp70 diversity, an ectosymbiosis-based mechanism of eukaryogenesis was examined in Figure 3.

In *Stage 1* of this mechanism, bacterial ectosymbionts were attached to an archaeal parent. Enwrapment of part or whole of the ectosymbionts by the archaeal parent led to the influx of bacterial genes, proteins and metabolic products, whereupon bacterial ester-lipids

(blue) infiltrated the archaeal ether-lipid (red) membranes to form initially chimeric and later fully bacterial membranes. The archaeal host and the ectosymbionts worked together in *Stage 2* to develop the mitosome, which contributed protein presequences for organellar targeting[28] and took part in Fe-S cluster assembly[29]. Upon departure of the ectosymbionts from the archaeal host, a mitosome-containing LECA/MSO emerged. It was suggested that an early membrane-curving module might play a role in the origin of nuclear pore complex and coated vesicles[30], in which case extra curvable prokaryotic membranes such as those of Abo[31] would merit extensive investigation. One of the incentives for the formation of cell nucleus might be that archaeal introns were spliced by an endoribonuclease at RNA bulge-helix-bulge motifs. Without a nucleus, some mRNAs from bacterial ectosymbionts would risk misrecognition as introns by the archaeal enzyme and erroneously excised.

At *Stage 3* the archaeal host collaborated further with the ectosymbionts to replace the mitosome with hydrogenosome, and generated hydrogenosome-containing AMIs. The bioenergetic resourcefulness of the hydrogenosome was illustrated by the amitochondriate hydrogenosome-containing animal phylum *Loricifera*[32]. Further archaeal-bacterial cooperation produced alphaproteobacterial endosymbionts at *Stage 4*, which led to the formation of mitochondria and the MTEs. Thus the ectosymbiotic mechanism gives rise to an LECA that is AMI in nature at *Stage 2*, and the many ectosymbionts participating in *Stages 1-4*, far exceeding the ones illustrated in Figure 3, would readily account for the multiple bacteria-derived Hsp70 motifs in eukaryotes in Figure 2.

Prokaryotic proteins in eukaryotic proteomes

The eukaryotic proteomes were enriched with prokaryote-derived proteins. The bacterial/archaeal protein ratios of 0.73/0.85/1.09/1.26/1.32 for Eae/Gla/Trv/Sce/Dme respectively pointed to the diminishing proportion of archaeal proteins as evolution of the MTEs proceeded, possibly due to the greater functional scope offered by bacterial proteins relative to archaeal proteins. The proteomes of Per, Bbo, Pfa and Bho were more primitive in character than those of other MTEs based on their higher proportions of archaeal proteins (Figure 4).

The entry of prokaryotic proteins into eukaryotes could be mediated by a number of different mechanisms. While horizontal gene transfers (HGTs) might be limited by low rates of acceptance by the gene recipients[33], it was found recently that some archaea displayed an 'accelerated gene adoption' (AGA) phenotype detectible by the elevated levels of bacteria-derived proteins in their proteomes[8]. Moreover, the genes of the bacterial species Rpr, Bap, Ctr, Mpn, and Tpa furnished relatively few protein coding genes that were adopted by the AGA-active archaeons, and their proteins were also depleted in the proteomes of both Gla and Trv, demonstrating the key role played by AGA in controlling influx of bacterial genes into the eukaryogenic lineage. Also the paucity of AGA activity in bacteria would also render them much less effective than archaea in hosting eukaryogenesis owing to the importance of collection of exogenous genes by the host in this regard. Anaerobic Abo and its facultatively anaerobic companion species Acf, on account of their top average similarity bitscores among archaeons toward Gla and Trv proteomes and the PEP-utilizing enzyme mobile domain of Gla, along with their outstanding AGA activity, emerged as leading candidate archaeal parents[8]. When the DNA-biology enzymes/proteins belonging to the Cluster of Ortholog Groups (COG) for replication, recombination and repair[34] were analyzed with respect to the similarity bitscores displayed by primitive AMIs toward Abo and various Asgard and TACK archaeons

which included some attractive candidate archaeal parents[8,35],the frequencies of top similarity bitscores received by Abo from the primitive AMIs pertaining to the DNA-biology enzymes, viz. for 8/21 enzymes from Gla, 6/20 from Eae, 7/29 from Trv and 7/20 from Gin, far surpassed those of any Asgard or TACK archaeon tested (Figure 5a and Supplementary Figure S1). These results identified Abo as the major source of the DNA-biology enzymes in the genomes of all of these primitive AMIs, and therefore Abo as the archaeal parent of Eukarya.

It has been pointed out that, because different groups of present-day eukaryotes share few common features, the vertical inheritance between them remains an assumption[35]. However, Figure 5b panels 1-8 show that a series of archaeal proteins including DNA-biology enzymes and eukaryotic signature proteins[36] (ESPs) maintained fairly steady similarity bitscores toward both AMIs and MTEs ranging from Microsporidia to Hsa, which attested to the vertical transmission of these proteins between eukaryotes, and their significance as part of the core of eukaryotic proteomes. Bho and its subtypes Bs1, Bs4, and Bs3 (no reported proteome and therefore not included in Figure 4), which were in possession of both hydrogenosome components and partial mitochondria[37], shared with the mitosome- and hydrogenosome-containing AMIs as well as the MTEs Bbo and Pfa the presence of Ssp-Tte-Hth proteins in their proteomes, thus pointing to a relationship between these firmicute species and the MRO-containing AMIs in keeping with the metabolic resemblance between hydrogenosome and Clostridium[38]. This was confirmed in Figure 5b panels 9-10, where the MRO-containing AMIs other than Microsporidia exhibited substantial bitscores toward the [Fe] hydrogenase and pyruvate:ferredoxin oxidoreductase (PFO) of clostridial Tte. The low or zero bitscores exhibited by the Microsporidia could be the result of similarity erosion due to the extensive mutation of microsporidial proteins: Nbo was the only microsporidian in the 9-member clade on the SSU rRNA tree with any residual bitscore toward the PFO of Tte.

Previously, Paracoccus was found to branch near eukaryotes on the trees for the subunits of succinate dehydrogenase/fumarate reductase[39]. The similarity bitscores of these subunits from Pde toward eukaryotes (Figure 5b panels 11-12) showed the expected absence of these subunits from the AMIs in contrast to their presence in a variety of MTEs. Comparison of panels 9-12 indicated that Ein and Enh exchanged their mitochondrial FrdB and SdhA for mitosomal [Fe] hydrogenase and PFO in their adaptation to anaerobiosis. The source of their [Fe] hydrogenase and PFO, be it AGA from Clostridium, or some ancestral AMI or Amoebozoa remains an open question.

When 151 different alphaproteobacteria together with some other bacteria were compared with respect to their similarity bitscores toward the mtDNA-encoded proteins of different eukaryotes including the bacteria-like Ram[40] (Figure 5c top panel; Supplementary Figure S2), the bitscores of Tis ranked the highest, followed by another Rhodospirillales Azo, suggesting that the endosymbiont precursor of mitochondria was either Tis or a close relative, in accord with the placement of the Tis branch next to the mitochondrial branch on the alphaproteobacterial tree[41]. However, a range of other alphaproteobacteria either contributed mtDNA-encoded proteins or showed overlapping similarities with Tis proteins; among the non-alphaproteobacteria, only Rso and Xca exhibited some modest bitscores. These findings were indicative of a shared genetic propensity of the alphaproteobacterial group for symbiosis with the eukaryotes, as in the case of the graded propensities of different strains of Nostoc for ectosymbiosis with Anthoceros[42]. Such shared propensity promoted symbioses between the eukaryotic lineage and multiple alphaproteobacterial species, thereby allowing entry of their genes into some eukaryotes to replace or recombine with the Tis-genes in mtDNA. Among human mtDNA-encoded proteins HM1-11, Ehrlichia and Rpr yielded the highest bitscores

177 toward HM2 and HM3 respectively (middle panel), suggesting that their proteins had replaced
178 original Tis sequences. Since Ehrlichia and Rpr were both infectious for humans, the
179 replacements might have occurred during infection. Replacements/recombinations also could
180 be a factor in the exceptionally diverse sources of Blastocystis mtDNA-encoded proteins and
181 host-related SSU rRNA subtypes[43].

182 Discussion

183 Because the 1.5 Mb archaeal parent genome of Abo was much smaller than for example
184 the 11.2 Mb eukaryotic genome of Gla, recruitment of exogenous prokaryotic genes into the
185 Abo genome was a top priority right from the start of eukaryogenesis. Too small to house an
186 upfront endosymbiont, the archaal host had to depend on AGA and ectosymbiosis for
187 acquisition of archaeal proteins from such prominent sources as Mco, Pfu, Mac, TACKs and
188 Asgards into the AMIs (Figure 4). The modest numbers of bacterial proteins from Tis, Rru, Bja,
189 Atu, Cvo, Cur, Rso, Aba, Xca, Pel, Dra,Tht, Bsu, Syn, Ssp, Tte, Hth, Cpo, Mau and Pmo
190 detected in various AMIs pointed to gene acquisitions by the AMIs from these bacteria through
191 ectosymbioses.The Eco proteins was a special case: they were present in modest numbers in
192 the microsporidial Slo-Oco, but increased sharply from Ebi onward to Hsa. Thus Eco was likely
193 an ectosymbiont with modest contribution of proteins at first, but became later either an
194 endosymbiont, or an especially effective or sustained ectosymbiont with an enhanced influx.
195 In any event, the presence of Eco proteins in both the AMIs and MTEs was consistent with the
196 Eco-like *str*-S10-*spc*-alpha operons, as well as clustering and organization of *rps* and *rpl* genes
197 in some mtDNAs[44].

198 The wide ranging influxes of prokaryotic genes through AGA and ectosymbiosis
199 enabled the AMIs to expand in genome size, and in cell volume akin to the development of
200 giant cell size by the archaeon *Giganthauma karukerense* through ectosymbiosis with a
201 gammaproteobacterium[45]. The convergence of enlarged cell volume, an expanded genome with
202 ESPs from Asgards, TACKs and other prokaryotes, and an exceptionally flexible cell
203 membrane that could enwrap symbionts with its blebs[31] equipped the AMIs for the
204 development of phagocytosis[46,47]. This resulted in endosymbioses with Tis marked by a flood
205 of Tis proteins that greatly exceeded the protein influxes from various ectosymbionts; and still
206 later with a cyanobacterium marked by a comparable flood of Syn-like proteins into the algae
207 Tps, Esi, Bpr, Gth and Cme. These endosymbioses gave the MTEs the twin biofuel engines of
208 oxidative phosphorylation and photosynthesis that drove their relentless expansion for the next
209 two giga years. In addition, the Rru, Bja, Atu, Cvo, Pde, Cur, Rso, Aba, and Xca proteins all
210 grew to pronounced levels in the MTEs that signaled their endosymbioses or highly
211 effective/sustained ectosymbioses with the eukaryotes. Notably, the mtDNA-encoded proteins
212 of numerous eukaryotes notched higher bitscores toward Tis than Pde, Bja, Bsu, Filomicrobium
213 and Halocynthiibacter (Figure 5c top panel), whereas the nuclear encoded PycA, NirB, AceF,
214 AHP1 and FMR2 of some eukaryotes notched higher bitscores toward Pde, Bja, Bsu,
215 Filomicrobium or Halocynthiibacter than Tis (bottom panel). Therefore the eukaryotes
216 plausibly formed separate symbioses with Tis, Pde, Bja, Bsu and others; while the Tis-
217 endosymbiont served as precursor to mitochondria, the symbioses with Pde, Bja, Bsu and
218 others introduced a variety of useful proteins into the eukaryotes.

219 With respect to the ectosymbiosis-based mechanism (Figure 3), the role of Abo as
220 archaeal parent and the presence of numerous ectosymbionts detected in Figure 4 confirmed
221 the postulated interactions of archaeal parent with ectosymbionts in various stages starting from

Stage 1. The contributions of [Fe] hydrogenase and PFO by Tte confirmed the postulate of archaeal-ectosymbiont collaborations to build mitosome and hydrogenosome in *Stage 2* and *Stage 3* respectively. Identification of a Tis-endosymbiont as precursor of mitochondria confirmed the postulate of archaeal-alphaproteobacterial endosymbiosis in *Stage 4*. Clearly, a horde of micron-size prokaryotes generated the Eukarya domain with their genes through accelerated gene adoption, ectosymbiosis and endosymbiosis, and all four stages of the ectosymbiotic mechanism were verified by fluxes of prokaryotic protein-coding genes into the eukaryotes[48].

The unique crown-capped structure of the SSU rRNA tree[4] stands as a monument to the cataclysmic effects of oxygen on eukaryotic evolution. That Abo and the lowest branching AMIs were anaerobic suggests that eukaryogenesis started in the anoxic Archean Eon. Oxygen-producing cyanobacteria evolved at or before 2.7 Gya, providing an atmosphere containing a trace amount of free oxygen that persisted up to 2.4 Gya. It made possible sterol biosynthesis in eukaryotes, leaving sterane footprints of eukaryotes in 2.7-Gya rocks even though it fell far short of supporting aerobic metabolism[49]. The increased levels of oxygen during the Great Oxidation Event at about 2.4 Gya in early Proterozoic Eon was a grave hazard to the anaerobic bacteria, archaea and eukaryotes, for the incongruence between enzyme distribution in aerobic metabolic pathways and organismal speciation also indicated that all three domains of organisms were already formed at that juncture. In response to the rising oxygen, all the surviving species from that era built oxic metabolic networks in addition to their earlier anoxic ones[50]. Among the species on the eukaryotic SSU rRNA tree, which was by necessity an all-AMI tree stemming from an amitochondriate LECA, the lowest branching species with their limited resources probably built oxic networks that were little more than 3-4% their anoxic networks as exemplified by the anaerobic prokaryotes Tma, Mka and Tte. As a result, these lowest branchers remained as AMIs, were saddled with reduced competitiveness, and had to choose between parasitism and confinement to anaerobic niches. Their long branches on the present-day tree could be indicative of their arduous evolutionary journeys to stay alive. Higher up on the tree, Pfa built an oxic network equal to about 5% of its anoxic one and managed to possess mitochondria, but still ended up as a parasite. In contrast, the elaborate species at the top of the tree acquired mitochondria and built large oxic networks greater than 20% of the anoxic network in Sce, 30% in Cel, 40% in *Arabidopsis thaliana*, or 50-60% in the mammals. Endowed with large oxic networks and plentiful ATP from oxidative phosphorylation, they launched into adaptive radiation to form the spikes on the crown of the tree. The accomplishments of the Hsa lineage began when it became one of the spikes.

Methods

Source of Data and Materials

Protein sequences were retrieved from NCBI GenBank release 231 (ftp://ftp.ncbi.nlm.nih.gov/genomes/)[51]. Assemblies of species analyzed were given in Supplementary Table S1. Quality checked SSU rRNA sequences were downloaded from the SILVA database release 132 (https://www.arb-silva.de/)[52]. The accession numbers of SSU rRNAs analyzed were available in Supplementary File S1. Eukaryotic mitochondrial DNA-encoded protein sequences were retrieved from the RefSeq mitochondrial reference genomes in the NCBI Protein database (https://www.ncbi.nlm.nih.gov/protein)[53].

Maximum-likelihood SSU rRNA tree

The SSU rRNAs of 47 eukaryotes and 3 archaeons (Abo, Mja, Lok) as outgroup were aligned in Clustal Omega[54], and the tree was constructed using RAxML with an GTR+GAMMA+I model, with 1000 bootstraps[55]. The final maximum-likelihood tree with bootstrap support values is available in Supplementary File S2.

Hsp70-type Sequences

The proteome of each species analyzed was searched against the database of Clusters of Orthologous Groups of proteins (COG)[34] using RPSBLAST[56] at a threshold set by the '-evalue' flag at 0.01. Potential Hsp70 sequences in each species were identified as those mapped to COG0443 (Molecular chaperone DnaK or Hsp70). Among the potential Hsp70 sequences, one sequence was manually selected, if available, for each of the Group I _KP motif and the Group II NGDAWV motif. See Supplementary Table S2 for sequence accession numbers.

Estimation of Nuclear or Mitochondrial Proteome Similarity Bitscores

When comparing proteome similarities for Figures 4 and 5c top panel, the proteomes of all subject species were used to construct a local BLAST database using makeblastdb[56], and every query proteome was searched against the local database using BLASTP with a BLOSUM62 matrix and thresholds setting to e-value < $1x10^{-5}$, percent identity > 25% and query coverage > 50%. Only the query and subject sequences that were the best-match of each other, viz. when query sequence *n* from species 1 exhibited the highest bitscore toward subject sequence *m* among all proteins of species 2 and vice versa, were included in the estimation of inter-proteome similarity, which was given by the sum of BLASTP bitscores of all such best-matched proteins between the two proteomes. The data for nuclear and mitochondrial proteome similarities is given in Supplementary Tables S3 and S4a respectively. The bacterial/archaeal protein ratio (R) was estimated as $R = \sum_1^i N_b / \sum_1^j N_a$, where $\boldsymbol{i}$ was the number of proteins in a eukaryotic proteome that could be blasted to the proteins of any of the 33 bacterial species analyzed in Figure 4; and $\boldsymbol{N_b}$ was the highest bitscore so obtained. Likewise, $\boldsymbol{j}$ was the number of proteins in the same eukaryotic proteome that could be blasted to any of the also 33 archaeal species; and $\boldsymbol{N_a}$ was the highest bitscore so obtained.

Similarity of COGs between eukaryotes and prokaryotes

Best-matched query and subject sequences passing BLASTP thresholds of e-value < $1x10^{-5}$, percent identity > 25% and query coverage > 50% that were mapped to the same COG using RPSBLAST at a threshold of e-value < 0.01 furnished data for Figures 5a, 5b and 5c middle and bottom panels. For Figure 5b, the bitscores of 10 out of 196 protein pairs that did not pass the percent-identity or query-coverage thresholds but showed BLASTP e-values < $1x10^{-10}$ were also included (see Supplementary Table S5 for the e-values, percent identities and query coverages for Figure 5b).

Data availability

All the data generated in this study were included in the supplementary information files.

303 # References

304 1. Martin, W. F., Garg, S. & Zimorski, V. Endosymbiotic theories for eukaryote origin. *Philos. Trans. R.*
305 *Soc. Lond., B, Biol. Sci.* **370**, 20140330, doi:10.1098/rstb.2014.0330 (2015).

306 2. Vossbrinck, C., Maddox, J., Friedman, S., Debrunner-Vossbrinck B. A. & Woese C. R. Ribosomal
307 RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature* **326**, 411–414 (1987).

308 3. Cavalier-Smith, T. The origin of eukaryotic and archaebacterial cells. *Ann. N. Y. Acad. Sci.* **503**, 17-
309 54 (1987).

310 4. Sogin, M. L. & Silberman, J. D. Evolution of the protists and protistan parasites from the perspective
311 of molecular systematics. *Int. J. Parasitol.* **28**, 11-20 (1998).

312 5. Roger, A. J. Reconstructing early events in eukaryotic evolution. *Am. Nat.* **154**, S146-S163 (1999).

313 6. Van der Giezen, M. Hydrogenosomes and mitosomes: conservation and evolution of functions. *J.*
314 *Eukaryot. Microbiol.* **56**, 221-231 (2009).

315 7. Dacks, J. B. et al. The changing view of eukaryogenesis-fossils, cells, lineages and how they all come
316 together. *J. Cell. Sci.* **129**, 3695-3703 (2016).

317 8. Long, X., Xue, H. & Wong, J. T. Descent of Bacteria and Eukarya from an archaeal root of life. *Evol.*
318 *Bioinform.* **16**, 1-11 (2020).

319 9. Rogozin, I. B., Basu, M. K., Csuros, M. & Koonin, E. V. Analysis of rare genomic changes does not
320 support the unikont-bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary
321 radiation of eukaryotes. *Genome Biol Evol* **1**, 99-113 (2009).

322 10. Cavalier-Smith, T. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biol.*
323 *Lett.* **6**, 342-345 (2010).

324 11. Katz, L. A., Grant, J. R., Parfrey, L. W. & Burleigh, J. G. Turning the crown upside down: gene tree
325 parsimony roots the eukaryotic tree of life. *Syst. Biol.* **61**, 653-660 (2012).

326 12. He, D. et al. An alternative root for the eukaryote tree of life. *Curr. Biol.* **24**, 465-470 (2014).

327 13. McInerney, J., Pisani, D. & O'Connell, M. J. The ring of life hypothesis for eukaryote origins is
328 supported by multiple kinds of data. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **370**, 20140323,
329 doi:10.1098/rstb.2014.0323 (2015).

330 14. Morin, L. Long branch attraction effects and the status of "basal eukaryotes": phylogeny and
331 structural analyses of the ribosomal RNA gene cluster of the free-living diplomonad Trepomonas
332 agiles. *J. Eukaryot. Microbiol.* **87**, 167-77 (2000).

333 15. Bergsten, J. A review of long-branch attraction. *Cladistics* **21**, 163-193 (2005).

334 16. Cavalier-Smith, T. Multigene phylogeny resolves deep branching of Amoebozoa. *Mol. Phylogenet.*
335 *Evol.* **83**, 293-304 (2015).

336 17. Karnkowska, A. et al. A eukaryote without a mitochondrial organelle. *Curr. Biol.* **26**, 1274-84 (2016).

337 18. Abrahamsen, M. S. et al. Complete genome sequence of the apicomplexan Cryptosporidium parvum.
338 *Science* **304**, 441-445 (2004).

339 19. Van de Peer, Y., Rensing, S. A., Maier, U. G. & De Wachter, R. Substitution rate calibration of small
340 subunit ribosomal RNA identifies Chlorarachniophyte endosymbionts as remnants of green algae.
341 *Proc. Natl. Acad. Sci. U.S.A.* **93**, 7732-36 (1996).

342 20. Germot, A., Philippe, H. & Le Guyader, H. Evidence for loss of mitochondria in Microsporidia from
343 a mitochondrial-type HSP70 in *Nosema locustae*. *Mol. Biochem. Parasitol.* **87**, 159-168 (1997).

344 21. Troemel, E. R. & Becnel, J. J. Genome analysis and polar tube firing dynamics of mosquito-infecting
345 microsporidia. *Fungal Genet. Biol.* **83**, 41-44 (2015).

346 22. Capella-Gutierrez, S., Marcet-Houben, M. & Gabaldon, T. Phylogenomics supports Microsporidia as
347 the earliest diverging clade of sequenced fungi. *BMC Biol.* **10**, 47 (2012).

348 23. Arisue, N., Sanchez, L. B., Weiss, L. M., Muller, M. & Hashimoto, T. Mitochondrial-type hsp70 genes
349 of the amitochindriate protists Giardia intestinalis, Entamoeba histolytica and two microsoporidia.
350 *Parasitol. Int.* **51**, 9-16 (2002).

351 24. Esser, C. et al. A genome phylogeny for mitochondria among alpha-proteobacteria and a
352 predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* **21**, 1643-60 (2004).

353 25. Rochette, N. C., Brochier-Armanet, C. & Gouy, M. Phylogenomic test of the hypotheses for the origin
354 of eukaryotes. *Mol. Biol. Evol.* **31**, 832-845 (2014).

355 26. Koonin, E. V. & Yutin, N. The dispersed archaeal eukaryome and the complex archaeal ancestor of
356 eukaryotes. *Cold Spring Harb Perspect Biol* **6**, a016188 (2014).

357 27. Akanni W. A. et al. Horizontal gene flow from eubacteria to archaebacteria and what it means for our
358 understanding of eukaryogenesis. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **370**, 20140337, doi:
359 10.1098/rstb.2014.0337(2015).

360 28. Clark, C. G. & Roger, A. J. Direct evidence for secondary loss of mitochondria in Entamoeba
361 histolytica. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 6518-21 (1995).

362 29. Pyrih, J. et al. Minimal cytosolic iron-sulfur cluster assembly machinery of Giardia intestinalis is
363 partially associated with mitosome. *Mol. Microbiol.* **102**, 701-714 (2016).

364 30. Devos, D. et al. Components of coated vesicles and nuclear pore complex share a common molecular
365 architecture. *PLoS Biol.* **2**, e380, doi: 10.1371/journal.pbio.0020380 (2004).

366 31. Reysenbach, A. L. et al. A ubiquitous thermoacidophilic archaeon from deep-sea hydrothermal vents.
367 *Nature* **442**, 444-447 (2006).

368 32. Danovaro, R. et al. The first metazoa living in permanently anoxic conditions. *BMC Biol.* **8**, 30 (2010).

369 33. Kurland, C. G., Canback, B. & Berg, O. G. Horizontal gene transfer: a critical view. *Proc. Natl. Acad.*
370 *Sci. U.S.A.* **100**, 9658-62 (2003).

371 34. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-
372 scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33-36 (2000).

373 35. Eme , L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes.
374 *Nat. Rev. Microbiol.* **15**, 711-723 (2017).

375 36. Hartman, H. & Fedorov, A. The origin of the eukaryotic cell: a genomic investigation. *Proc. Natl.*
376 *Acad. Sci. U.S.A.* **99**, 1420-25 (2002).

37. Gentekaki, E. et al. Extreme genome diversity in the hyper-prevalent parasitic eukaryote Blastocystis. *PLoS Biol.* **15**, e2003769, doi:10.1371/journal.pbio.2003769 (2017).

38. Whatley, J. M., John, P. & Whatley, F. R. From extracellular to intracellular: the establishment of mitochondria and chloroplasts. *Proc. R. Soc. Lond., B, Biol. Sci.* **204**, 165-187 (1979).

39. Tielens, A. G. M., Rotte, C., van Hellemond, J. J. & Martin, M. Mitochondria as we don't know them. *Trends Biochem. Sci.* **27**, 565-572 (2012).

40. Burger, G., Gray, M. W., Forget, L. & Lang, B. F. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol Evol* **5**, 418-438 (2013).

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

42. Enderlin, C. S. & Meeks, J. C. Pure culture and reconstitution of the Anthoceros-Nostoc symbiotic association. *Planta.* **158**, 157-165 (1983).

43. Stensvold, C. R. et al. Subtype distribution of Blastocystis isolates from synanthropic and zoo animals and identification of a new subtype. *Int. J. Parasitol.* **39**, 473-479 (2009).

44. Gray, M. W. Mitochondrial evolution. *Cold Spring Harb Perspect Biol* **4**, a011403, doi:10.1101/cshperspect.a011403 (2012).

45. Muller, F., Brissac, T., Le Bris, N., Felbeck, H. & Gros, O. First description of giant Archaea (Thaumarchaeota) associated with putative bacterial ectosymbionts in a sulfidic marine habitat. *Environ. Microbiol.* **12**, 2371-83 (2010).

46. Yutin, N., Wolf, M. Y., Wolf, Y. I. & Koonin, E. V. The origins of phagocytosis and eukaryogenesis. *Biol. Direct* **4**, 9, doi:10.1186/1745-6150-4-9 (2009).

47. Martin, W. F., Tielens, A. G. M., Mentel, M., Garg, S. G. & Gould, S. B. The physiology of phagocytosis in the context of mitochondrial origin. *Microbiol. Mol. Biol. Rev.* **81**, e0008-17 (2017).

48. Wong, J. T., Xue, H. & Long, X. An ectosymbiosis-based mechanism of eukaryogenesis. *Preprint* (2020).

49. Kasting, J. F. The primitive Earth. In *Prebiotic Evolution and Astrobiology* eds. Wong, J. T. & Lazcano, A. 57-63 Landes Bioscience (2009).

50. Raymond, J. & Segre, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311**, 1764-67 (2006).

51. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44**, D67-72, doi:10.1093/nar/gkv1276 (2016).

52. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-596, doi:10.1093/nar/gks1219 (2013).

53. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).

54. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135-145, doi:10.1002/pro.3290 (2018).

55. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

56. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

Acknowledgements

The authors would like to acknowledge the University Grant Council for funding support to Applied Genomics Center [No. 5112-703-0110D-42002].

Author contributions

J.T.W. and H.X. conceived the study; X.L. collected the data and performed computational analysis; and J.T.W., X.L. and H.X wrote the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare that there is no conflict of interest with respect to the research, authorship, and/or publication of this article.

Materials & Correspondence

Correspondence to J. Tze-Fei Wong.

Table 1. Partial list of species analysed.

ABBR.	SPECIES NAME
ARCHAEA	
Abo	*Aciduliprofundum boonei*
Acf	*Aciduliprofundum sp. MAR08-339*
Afu	*Archaeoglobus fulgidus*
Aia	*Acidilobus sp. 7A*
Alt	*C.Altiarchaeales archaeon*
Ape	*Aeropyrum pernix*
Bat	*C.Bathyarchaeota archaeon*
Csu	*C.Caldiarchaeum subterraneum*
Csy	*Cenarchaeum symbiosum*
Fac	*Ferroplasma acidiphilum*
Ffo	*Fervidicoccus fontis*
Hal	*Halobacterium salinarum*
Hei	*C.Heimdallarchaeota archaeon*
Hgi	*Haloferax gibbonsii*
Hla	*Halobiforma lacisalsi*
Kcr	*C.Korarchaeum cryptofilum*
Lok	*Lokiarchaeum sp. GC14_75*
Mac	*Methanosarcina acetivorans*
Mbo	*Methanoregula boonei*
Mco	*Methanocella conradii*
Mes	*C.Methanosuratus sp.*
Min	*C.Methanomassiliicoccus intestinalis*
Mja	*Methanocaldococcus jannaschii*
Mka	*Methanopyrus kandleri*
Mnt	*Methanonatronarchaeum thermophilum*
Mph	*Methanophagales archaeon*
Mte	*C.Methanoplasma termitum*
Nca	*C.Nitrosocaldus cavascurensis*
Nga	*C.Nitrososphaera gargensis*
Odi	*C.Odinarchaeota archaeon*
Pae	*Pyrobaculum aerophilum*
Pfu	*Pyrococcus furiosus*
Psy	*C.Prometheoarchaeum syntrophicum*
Sso	*Saccharolobus solfataricus*
Tac	*Thermoplasma acidophilum*
Tho	*C.Thorarchaeota archaeon*
Tvo	*Thermoplasma volcanium*
Woa	*C.Woesearchaeota archaeon*
BACTERIA	
Aae	*Aquifex aeolicus*
Aba	*Acinetobacter baumannii*
Atu	*Agrobacterium tumefaciens*
Azo	*Azospirillum sp. M2T2B2*
Bap	*Buchnera aphidicola*
Bja	*Bradyrhizobium japonicum*
Blo	*Bifidobacterium longum*
Bsu	*Bacillus subtilis*
Cex	*Caldisericum exile*
Cje	*Campylobacter jejuni*
Cpo	*Cloacibacillus porcorum*
Cur	*Curvibacter sp.*
Cvo	*Chelativorans sp. BNC1*
Dra	*Deinococcus radiodurans*
Dth	*Dictyoglomus thermophilum*
Eco	*Escherichia coli*
Hth	*Hungateiclostridium thermocellum*
Kae	*Klebsiella aerogenes*
Kol	*Kosmotoga olearia*
Mau	*Mahella australiensis*
Mca	*Macrococcus caseolyticus*
Mtu	*Mycobacterium tuberculosis*
Pde	*Paracoccus denitrificans*
Pel	*Pelobacter sp. SFB93*
Pmo	*Petrotoga mobilis*
Rpr	*Rickettsia prowazekii*
Rru	*Rhodospirillum rubrum*
Rso	*Ralstonia solanacearum*

ABBR.	SPECIES NAME
BACTERIA (CONTINUED)	
Spn	*Streptococcus pneumoniae*
Ssp	*Sporanaerobacter sp. NJN-17*
Syn	*Synechocystis sp. PCC 6803*
Tht	*Thermobaculum terrenum*
Tis	*Tistrella mobilis*
Tma	*Thermotoga maritima*
Tte	*Thermoanaerobacter tengcongensis*
Xca	*Xanthomonas campestris*
EUKARYA	
Aal	*Anncaliia algerae*
Acc	*Acanthamoeba castellanii*
Ago	*Andalucia godoyi*
Asu	*Acytostelium subglobosum*
Bbo	*Babesia bovis*
Bde	*Batrachochytrium dendrobatidis*
Bho	*Blastocystis hominis*
Bpr	*Bathycoccus prasinos*
Bs1	*Blastocystis sp. subtype 1*
Bs3	*Blastocystis sp. subtype 3*
Bs4	*Blastocystis sp. subtype 4*
Cel	*Caenorhabditis elegans*
Cme	*Cyanidioschyzon merolae*
Cne	*Cryptococcus neoformans*
Cpa	*Cryptosporidium parvum*
Ddi	*Dictyostelium discoideum*
Dme	*Drosophila melanogaster*
Dpu	*Dictyostelium purpureum*
Dre	*Danio rerio*
Eae	*Edhazardia aedis*
Ebi	*Enterocytozoon bieneusi*
Ecu	*Encephalitozoon cuniculi*
Ein	*Entamoeba invadens*
Enh	*Entamoeba histolytica*
Esi	*Ectocarpus siliculosus*
Gin	*Giardia intestinalis*
Gla	*Giardia lamblia*
Gth	*Guillardia theta*
Has	*Homo sapiens*
Imu	*Ichthyophthirius multifiliis*
Lbi	*Laccaria bicolor*
Mbr	*Monosiga brevicollis*
Mci	*Mucor circinelloides*
Mon	*Monocercomonoides sp. PA203*
Mpa	*Marchantia paleacea*
Nbo	*Nosema bombycis*
Ncr	*Neurospora crassa*
Ngr	*Naegleria gruberi*
Oco	*Ordospora colligate*
Per	*Perkinsela sp. CCAP 1560/4*
Pfa	*Plasmodium falciparum*
Pte	*Paramecium tetraurelia*
Ram	*Reclinomonas Americana*
Sap	*Saprolegnia parasitica*
Sce	*Saccharomyces cerevisiae*
Slo	*Spraguea lophii*
Spo	*Schizosaccharomyces pombe*
Spu	*Spizellomyces punctatus*
Sra	*Strongyloides ratti*
Ssa	*Spironucleus salmonicida*
Tbr	*Trypanosoma brucei*
Tgo	*Toxoplasma gondii*
Tps	*Thalassiosira pseudonana*
Trh	*Trachipleistophora hominis*
Trv	*Trichomonas vaginalis*
Tth	*Tetrahymena thermophila*
Ttr	*Thecamonas trahens*
Vcu	*Vavraia culicis*

Note: C. in front of species name stands for Candidatus. Detailed species information is given in Supplementary Table S1.

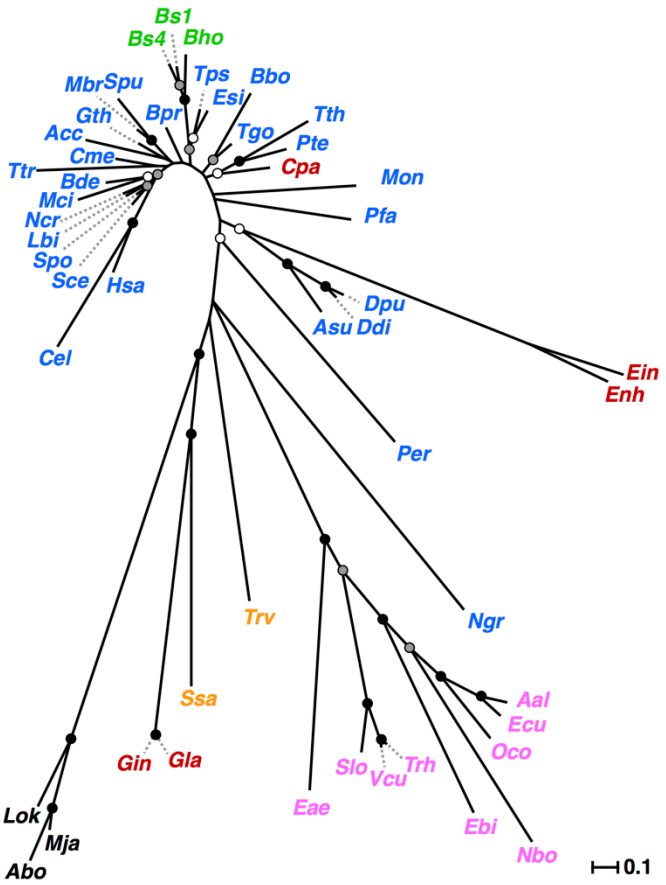

Figure 1. Maximum-likelihood SSU rRNA tree. Eukaryotes with mitochondria, and also Mon, are labeled in blue; Microsporidia in pink; non-Microsporidia eukaryotes with mitosome in red; eukaryotes with hydrogenosome in orange; eukaryotes with MRO and part of mitochondria in green; and an archaeal outgroup in black. Nodes with > 90% bootstrap support are represented by black circles, > 70% grey circles, > 50% open circles, and ≤ 50% without circle.

```
Hsp70 I    KRLIGR                          -KP                          KMK
Gin I      DAKRLIGRRFNDPEVQADLKHFPFKVICGPDEKPQIQVVYKGETKTFSPEEISSMVLTKMK  90-150
Gla I      DAKRLIGRRFNDPEVQADLKHFPFKVICGPDEKPQIQVVYKGETKTFTPEEISSMVLTKMK  69-129
Aba        DAKRLIGRRFSDPTVQNDMKLWPFKVISGPAEKPMIVVTYKGEEKQFAAEEISSMVLTKMK  3-63

Ein I      DAKRLIGRRFSDPAIQEDMKHWSFTVVNDGNDKPLIQVEYKGETKKFTPEEISSMVLTKMK  71-131
Enh I      DAKRLIGRRFSDPAIQNDMKHWSFKVIDDGHDKPLIEVEYKGEVKKFTPEEISSMVLTKMK  71-131
Cpa I      DAKRLIGRKFDDQAVQSDMTHWPFKVVRGPKDKPIISVNYLGEKKEFHAEEISAMVLQKMK  80-140
Ssa I      DAKRLIGRRFSEADVQDDMKHFPFTVVG-VDDKPMIEVVYKGETKRFTPEEISSMILLKMK  68-127
Ngr I      DAKRLIGRKFTDSSVQSDMKHWPFKVITKGDDKPYIQVEYKNETHTFTPEQVSSMVLTKMR  71-131
Ddi I      DAKRLIGRKFSDSVVQSDMKHWPFKVITNNDDKPLIQVDFKGETKTFSPEEISSMVLSKMK  66-127
Acc I      DAKRLIGRNFSDPIVQADMKHWPFKVVQKPGDKPFISVEYKGETKEFSPEEVSSMVLIKMK  69-129
Lma I      DAKRLIGRKFNDSVVQSDMKHWPFKVTTKGDDKPVISVQYRGEEKTFTPEEISSMVLLKMK  69-129
Bsl I      DAKRLIGRHFSDQIVQNDLKHWPFKLIADKNDKPIIEVTFMGKTQQYTCEQISSMVLGKMK  74-134
Bs4 I      DAKRLIGRKFSDAIVQNDLKHWPFRLIPDKFDKPIIEVQYLGKTKQFTCEQISSMVLGKMK  74-134
Pte I      DAKRLIGRKFSESTVQQDLKLWPFKVEAGADDKPIIVVKYKGENKKFHPEEISSMVLTKMK  69-129
Cur        DAKRLIGRKFSDSVVQADAKHWPFKVVTRGDDKPLIEVTFKGETKQFAPEEVSSMVLVKMK  74-134
Kae        DATRLIGRRFTDASVQSDMKLWPFKVLPGPGDKPMIVVNYKGEEKQFSAEEISSMVLIKMR  1-61
                                       ↑
Bho I      DAKRLIGRKFDDASVQSDMKHWPFTLKRGPEGKPVIEVEYKGERKQFSPEEISSMVLTKMK  72-132
Esi I      DAKRLIGRKFTDPAVASDKKHWPFSVVAGTGGKPTIEVSFKGDKKQFAPEEISSMVLVKMK  119-179
Hsa I      DAKRLIGRKFEDATVQSDMKHWPFRVVS-EGGKPKVQVEYKGETKTFFPEEISSMVLTKMK  70-129
Mca        DAKRLIGRKFEDATVQADMKHWPFEVVS-DGGKPKIKVSYKGEDKTFFPEEVSSMVLTKMK  71-130
Eco        DAKRLIGRKFDDGSVQSDMKHWPFKVVNAGGGKPKVQVEYKGETKTFTPEEISSMVLVKMK  68-128

Hsp70 II   KRLIGR                          NGDAWV                       KMK
Gin II     ASKRLIGRRFTDAEVKRAAKHVSYDIVQGPSGEAMINVPNL--QKAISPIEVGSEVLKYIK  77-135
Gla II     ASKRLIGRRFTDAEVKRAAKHVSYDIVQGPNGEAMINVPNL--HKSVSPIEVGSEILKYIK  75-133
Ecu II     ASKRLIGRKFGDPEVEKYVKGLPFDTMSHCNGDVWIRV-D---GKKYSPAQIGAFVLSKLK  78-134
Nce II     ASKRLLGRKFDDVELKDYLKSLPYDTTKHCNGDIWIKI-D---NKKYSPAQIGAFILSKMK  73-129
Enh II     GVKRMIGCQYKNVEQ----QRKPYKIVEGRNGEGWIYI-N---GKTYSPSEISSFILKKLK  74-126
Ein II     HAKRLIGIKYDQVTP---FMRQGYTIIKGPNNDAWVEV-N---GVKYSPSQISAEVLKKMK  74-127
Cpa II     ATKRLIGRRYEEEAIKKEQGILPYKIVRADNGDAWVEA-R---GERYSPSQIGAFILEKMK  117-173
Trv II     ATKRLIGRSFEDEMTKKCREMVPYQIIKAKNGDAWVKDEK---GNEYSPAQIGAFVLTKMK  88-145
Ngr II     AIKRLIGRRFDDPMTKRDMEIVPYDIIRGPNGDAWVRA-G---GKDLSPSQVGAFILQKMK  93-149
Ddi II     ATKRLIGRRFDDPMTKKDMTMVPYKIVKGPNGDAWFEV-K---GKMISPSEAGAMVLQKMK  96-152
Lma II     AVKRLIGRRFEDEHIQKDIKNVPYKIVRAGNGDAWVQDGN---GKQYSPSQIGAFVLEKMK  92-149
Acc II     AVKRLIGRDFNDPMTQKDLKMVPYKIVRHSNGDAWLEDSW---GKKYSPSEISAFTLTKMK  107-164
Bho II     AAKRLIGRSYDDPEVAKIKKLVPYEIIKGDNGDAWVSA-R---GQKMSPSQIGSMVLTKMK  73-129
Bsl II     AAKRLIGRSYDDPEVAKIKKLVPYEIVKGDNGDAWVQV-R---GQKMSPSQIGSMVLGKMK  98-154
Bs4 II     AAKRLIGRPYSDPEVEKIKKLVPYEIVKGDNGDAWVSV-R---GQKMSPSQIGSMVLTKMK  98-154
Esi II     ATKRLIGRRFKDKEVQGVQKLVPYKIVESQDGAAWVEV-Q---GNKMSPSQVGSMVLTKMK  113-169
Pte II     ATKRLIGRRFDDPNVQKDIKHLSYSVVRAQNGDAWVSLKS---GQTYSPSQIGAFVLIKMK  94-151
Hsa II     ATKRLIGRRYDDPEVQKDIKNVPFKIVRASNGDAWVEA-H---GKLYSPSQIGAFVLMKMK  119-175
Tis        AVKRLIGRTFNDPTTQKDAKMVPYRIVKADNGDAWVEV-R---GKKQSPSQISAQILKKMK  68-124
Atu        AVKRLIGRRYEDPTVEKDKALVPFEIVKGDNGDAWVKA-Q---DKNYSPSQISAMILQKMK  68-124
Rru        AIKRLIGRRYSDPTVEKDKGLVPYRIVKGDNGDAWVQT-R---DEKYAPSQLSAFILQKMK  68-124
Eco        AIKRLIGRRFQDEEVQRDVSIMPFKIIAADNGDAWVEV-K---GQKMAPPQISAEVLKKMK  68-124
Pel        AIKRLVGRKFDSDAVRKDIEISPFKIINADNGDAWVEA-R---GKKYSPPEISAMILQKMK  68-124

Archaea DnaK
Abo        AFKRKMGTDYKY-------------------------KIY---GKEYTPQQLSAFLLQKIK  73-105
Lok        AIKRKMGTSYKV-------------------------KID---GKEYTPQEISAMILRKIK  72-104
Hgi        SIKRHMGED-----------------------GYSVDIE---GEEYTPEQISAMILQKIK  70-103
```

Figure 2. Hsp70 sequence motifs. Group-I sequences with EKP/DKP/GKP motifs and Group-II sequences with NGDAWV motif in eukaryotic Hsp70 sequences and prokaryotic DnaK sequences were aligned using the KRLIGR and KMK motifs at the 5' and 3' ends of segment as positional markers. The numbers on the right indicate the positions of amino acid residues of each segment within its Hsp70 or DnaK sequence. Names of eukaryotes are colored as in Figure 1, and names of prokaryotes are colored black. The arrow indicates the position of an additional amino acid residue on the Ddi I sequence.

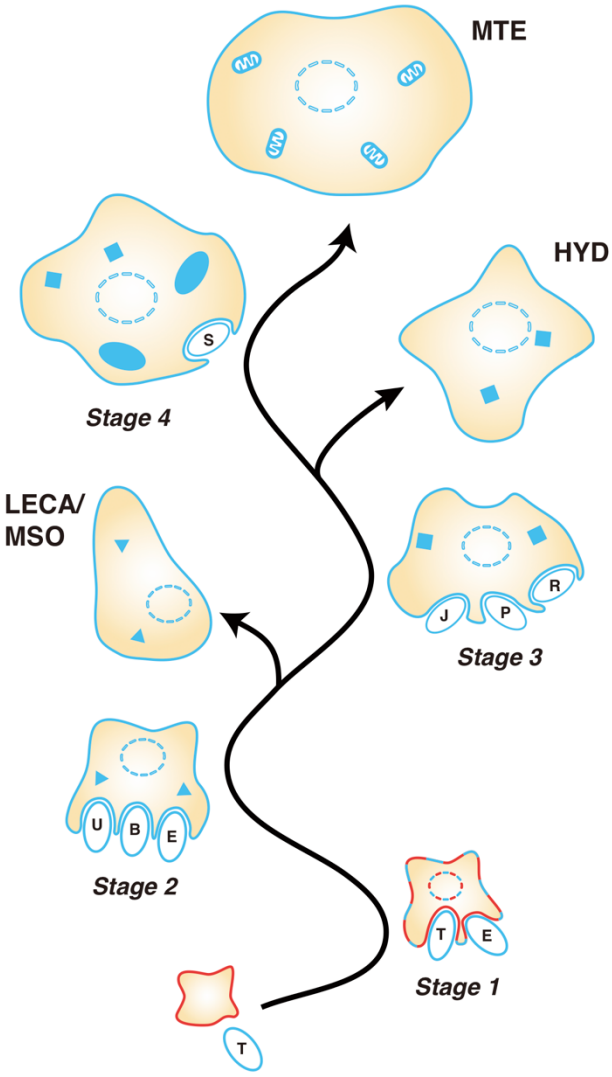

Figure 3. Ectosymbiotic mechanism of eurkaryogenesis. Stages 1-4 are developmental stages. Ellipse enclosed by dashed line represents cell nucleus. The membranes of archaeal parent are colored red and bacterial membranes colored blue. Solid triangles are mitosomes, and solid squares hydrogenosomes. MSO, HYD and MTE represent mitosome-containing, hydrogenosome-containing and mitochondria-containing eukaryotes. The letters T, E, U, B, J, P, R and S stand for Tte, Eco, Rru, Bsu, Bja, Pde, Rso and Tis respectively, representing a small fraction of the bacteria taking part as ectosymbionts in the mechanism. Solid ellipses denote endosymbionts, and ellipses with cristae mitochondria.

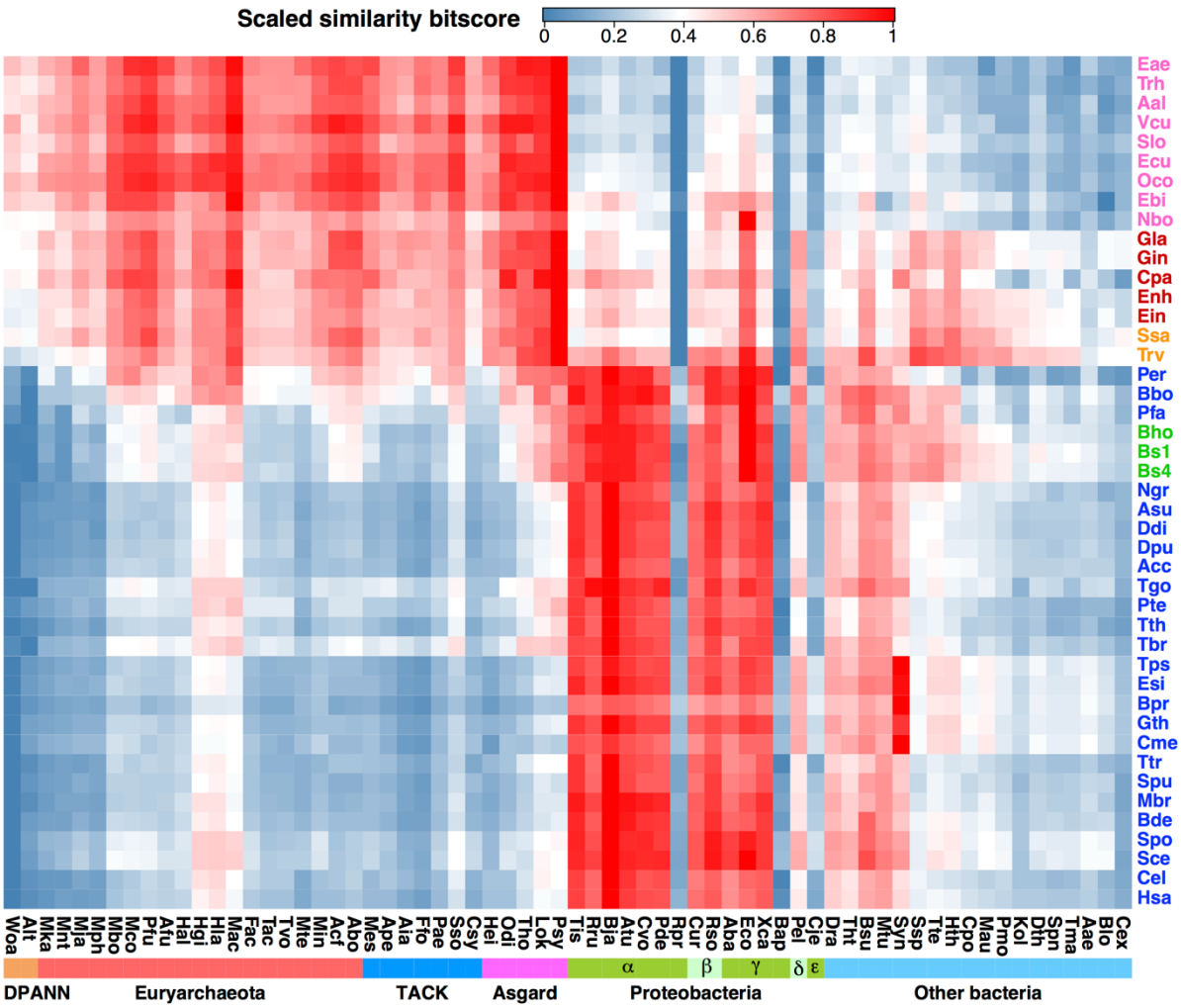

Figure 4. Heat map of inter-proteome similarity bitscores between eukaryotes (y-axis) and prokaryotes (x-axis). For each eukaryotic proteome, its similarity bitscores toward different prokaryotes were scaled linearly from 0 to 1.0 on the thermal scale, with zero (blue) marking the minimum, and 1.0 (red) the maximum bitscores. The names of different eukaryotes are colored as in Figure 1; and the colour bands on the x-axis indicate different groupings of prokaryotic species. The raw bitscores for the heat map are shown in Supplementary Table S3.

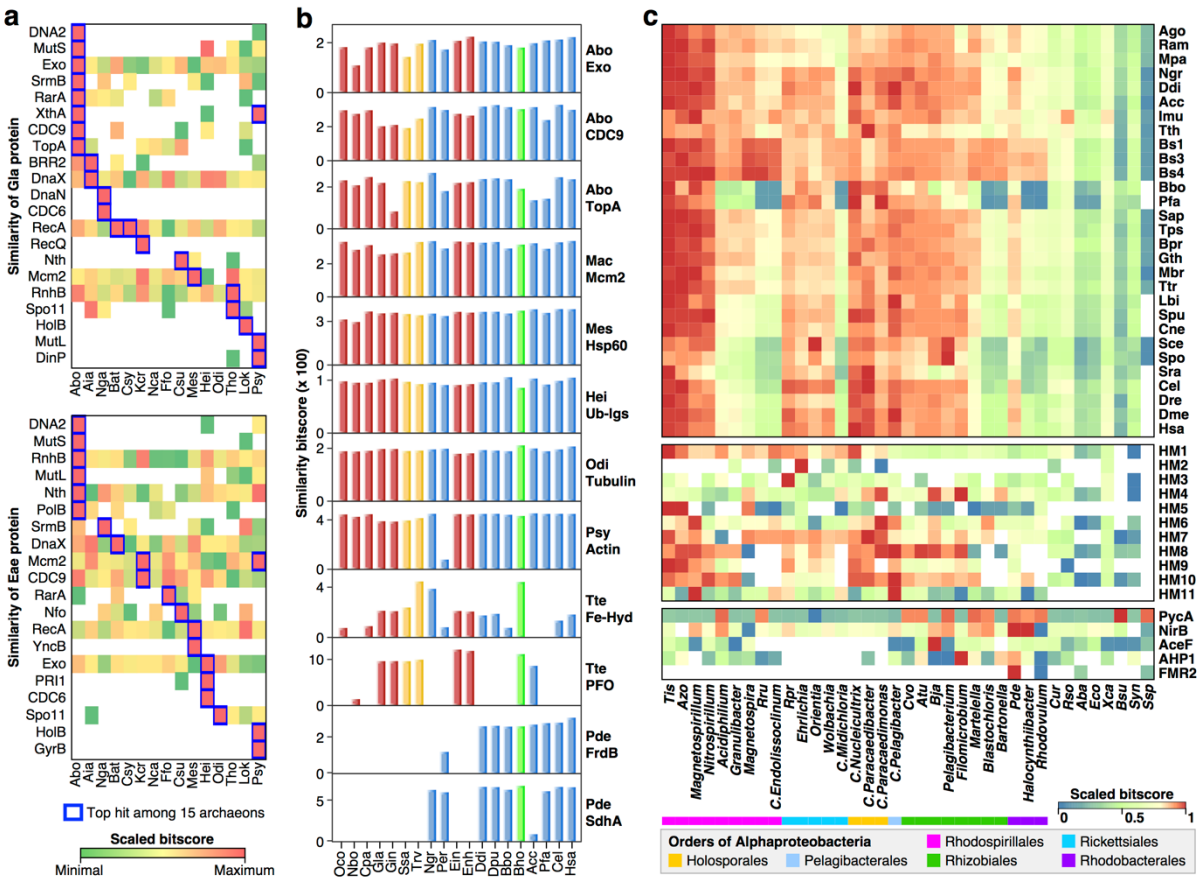

Figure 5. Protein similarities. (a) Similarity bitscores of 'Replication, recombination and repair' enzymes/proteins in Gla or Eae toward their counterparts belonging to the same COG in different archaeons. (b) Similarity bitscores of different eukaryotic proteins (x-axis) toward their prokaryotic counterparts employed as probes (y-axis, on right): Exo, 5'-3' exonuclease; CDC9, ATP-dependent DNA ligase; TopA, DNA topoisomerase IA; Mcm2, DNA replicative helicase MCM subunit Mcm2; Ub-lgs, ubiquitin-protein ligase; Fe-Hyd, iron only hydrogenase; FrdB the Fe-S protein subunit, and SdhA the flavoprotein subunit, of succinate dehydrogenase/fumarate reductase. (c) Top panel: similarity bitscores between bacterial proteomes (x-axis) and total mtDNA-encoded proteins in individual eukaryotes (y-axis). Middle panel: similarity bitscore between each of human mtDNA-encoded proteins HM1-11 (representing NADH dehydrogenase subunits 1-4, 4L and 5, cytochrome b, cytochrome oxidase subunits I-III, and ATP synthase F0 subunit-6 respectively) and its counterpart in different bacterial species. Bottom panel: similarity bitscore between nuclear DNA-encoded eukaryotic proteins PycA, pyruvate carboxylase from Sce; NirB, NAD(P)H-nitrite reductase from Pte; AceF, pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide acyltransferase (E2) component from Tbr; AHP1, peroxiredoxin from Pfa; or FMR2, fatty acid repression mutant protein from Tbr, and its counterpart in different bacterial species. (See similarity bitscores for Gla, Eae, Trv and Gin in an expansion of part a in Supplementary Figure S1; bitscore datasets for parts b and c in Supplementary Tables S4-S5; comparison of the probes for Fe-Hyd and PFO from Tte, Ssp and Hth in Supplementary Table S5b; and expanded bitscore datasets relating to mtDNA-encoded proteins in Supplementary Figure S2).