Bacterial Genes Outnumber Archaeal Genes in Eukaryotic Genomes

Julia Brueckner and William F. Martin*

Institute for Molecular Evolution, Heinrich Heine University Düsseldorf, Germany

*Corresponding author: E-mail: bill@hhu.de.

Abstract

Eukaryotes are typically depicted as descendants of archaea, but their genomes are evolutionary chimeras with genes stemming from archaea and bacteria. Which prokaryotic heritage predominates? Here, we have clustered 19,050,992 protein sequences from 5,443 bacteria and 212 archaea with 3,420,731 protein sequences from 150 eukaryotes spanning six eukaryotic supergroups. By downsampling, we obtain estimates for the bacterial and archaeal proportions. Eukaryotic genomes possess a bacterial majority of genes. On average, the majority of bacterial genes is 56% overall, 53% in eukaryotes that never possessed plastids, and 61% in photosynthetic eukaryotic lineages, where the cyanobacterial ancestor of plastids contributed additional genes to the eukaryotic lineage. Intracellular parasites, which undergo reductive evolution in adaptation to the nutrient rich environment of the cells that they infect, relinquish bacterial genes for metabolic processes. Such adaptive gene loss is most pronounced in the human parasite *Encephalitozoon intestinalis* with 86% archaeal and 14% bacterial derived genes. The most bacterial eukaryote genome sampled is rice, with 67% bacterial and 33% archaeal genes. The functional dichotomy, initially described for yeast, of archaeal genes being involved in genetic information processing and bacterial genes being involved in metabolic processes is conserved across all eukaryotic supergroups.

Key words: eukaryote origin, endosymbiosis, archaeal host, last eukaryote common ancestor, symbiogenesis, classification.

Introduction

Biologists recognize three kinds of cells in nature: Bacteria, archaea, and eukaryotes. The bacteria and archaea are prokaryotic in organization, having generally small cells on the order of 0.5–5 μm in size and ribosomes that translate nascent mRNA molecules as they are synthesized on DNA (cotranscriptional translation) (Whitman 2009). Eukaryotic cells are generally much larger in size, more complex in organization, and have larger genomes possessing introns that are removed (spliced) from the mRNA on spliceosomes (Collins and Penny 2005). Eukaryotic cells always harbor a system of internal membranes (Gould et al. 2016; Barlow et al. 2018) that form the endoplasmic reticulum and the cell nucleus, where splicing takes place (Vosseberg and Snel 2017). Furthermore, eukaryotes typically possess double membrane bounded bioenergetic organelles, mitochondria, which were present in the eukaryote common ancestor (LECA) (Embley and Martin 2006; Roger et al. 2017), but have undergone severe reduction in some lineages (van der Giezen 2009; Shiflett and Johnson 2010). In terms of timing during Earth history, it is generally agreed that the first forms of life on Earth were prokaryotes, with isotopic evidence for the existence of bacterial and archaeal metabolic processes tracing back to rocks 3.5 Gy of age (Ueno et al. 2006; Arndt and Nisbet 2012) or older (Tashiro et al. 2017). The microfossil record indicates that eukaryotes arose later, ~1.4–1.6 Ga (Javaux and Lepot 2018), hence that eukaryotes arose from prokaryotes. Though eukaryotes are younger than prokaryotes, the nature of their phylogenetic relationship(s) to bacteria and archaea remains debated because of differing views about the evolutionary origin of eukaryotic cells.

In the traditional three domain tree of life, eukaryotes are seen as a sister group to archaea (Woese et al. 1990; Da Cunha et al. 2017, 2018) (fig. 1a). In newer two-domain
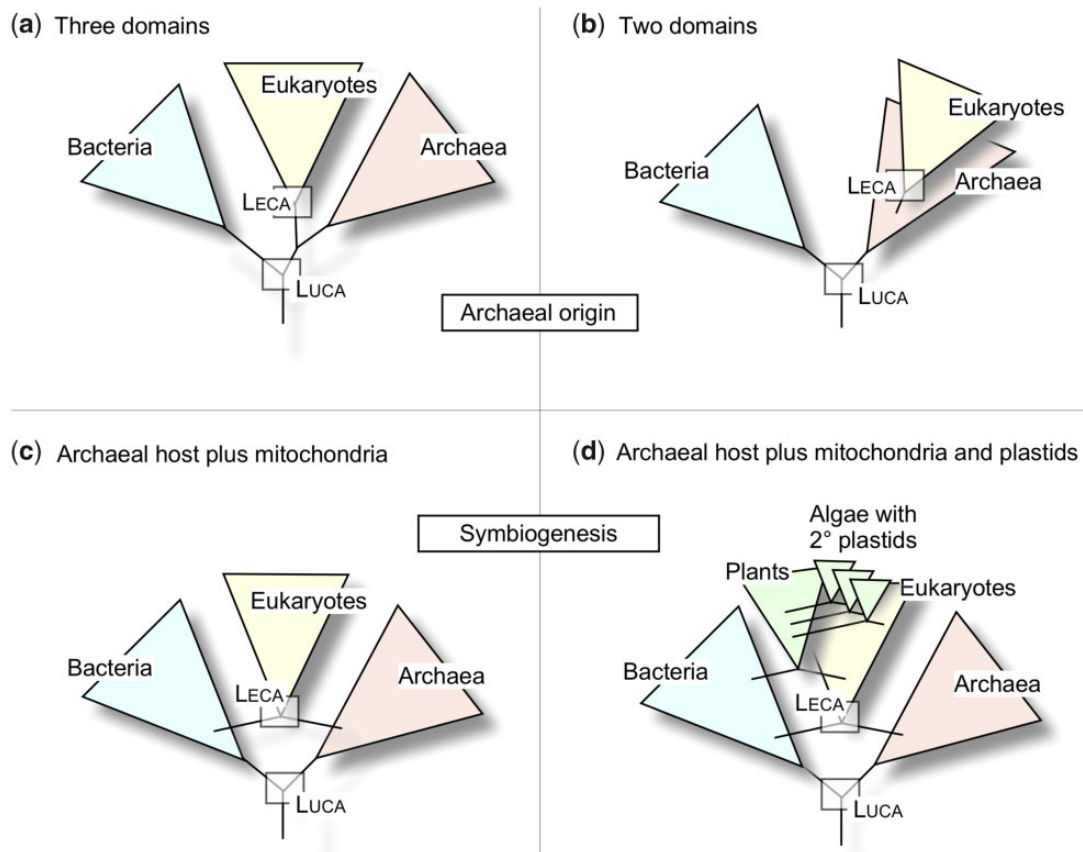
Fig. 1.—Differing views on the relationships of eukaryotes to prokaryotes. (a) The three domain tree. (b) The two-domain tree with an archaeal origin of eukaryotes. (c) Symbiogenesis at the origin of eukaryotes. (d) Symbiogenesis at the origin of eukaryotes plus plastids at the origin of the plant kingdom and secondary symbiotic events among algae (see Embley and Martin 2006; Gould et al. 2008; McInerney et al. 2014; Martin 2017).

trees, eukaryotes are viewed as branching from within the archaea (Cox et al. 2008; Williams et al. 2013) (fig. 1b). In both the two domain and the three domain hypotheses, this is often seen as evidence for "an archaeal origin" of eukaryotes (Cox et al. 2008; Williams et al. 2013) (fig. 1a, b). Germane to an archaeal origin is the view that eukaryotes are archaea that became more complex by gradualist evolutionary processes, such as point mutation and gene duplication (Field et al. 2011; Schlacht et al. 2014). Countering that view are two sets of observations relating to symbiogenesis (origin through symbiosis) for eukaryotes (fig. 1c, d). First, the archaea that branch closest to eukaryotes in the most recent phylogenies are very small in size (0.5 μm), they lack any semblance of eukaryote-like cellular complexity, and they live in obligate association with bacteria (Imachi et al. 2020), clearly implicating symbiosis (Imachi et al. 2020) rather than point mutation as the driving force at the origin of the eukaryotic clade (fig. 1c). Second, and with a longer history in the literature, are the findings that mitochondria trace to the LECA (Embley and Hirt 1998; van der Giezen 2009; McInerney et al. 2014) and that many genes in eukaryote genomes trace to

gene transfers from endosymbiotic organelles (Martin and Herrmann 1998; Timmis et al. 2004; Ku et al. 2015). A symbiogenic origin of eukaryotes would run counter to one of the key goals of phylogenetics, namely to place eukaryotes in a natural system of phylogenetic classification where all groups are named according to their position in a bifurcating tree. If eukaryotes arose via symbiosis of an archaeon (the host) and a bacterium (the mitochondrion), then eukaryotes would reside simultaneously on both the archaeal and the bacterial branches in phylogenetic schemes (Brunk and Martin 2019; Newman et al. 2019), whereby plants and algae that stem from secondary symbioses (Gould et al. 2008) would reside on recurrently anastomosing branches as in figure 1d.

Even though it is uncontested that symbiotic mergers lie at the root of modern eukaryotic groups via the single origin of mitochondria, plants via the single origin of plastids, and at least three groups of algae with complex plastids via secondary symbiosis (Archibald 2015), anastomosing structures such as those depicted in figure 1c and d do not mesh well with established principles of phylogenetic classification, because the classification of groups that arise by symbiosis is not

unique. One could rightly argue that plants are descended from cyanobacteria, which is in part true because many genes in plants were acquired from the cyanobacterial antecedent of plastids (Martin et al. 2002). Or one could save phylogenetic classification of eukaryotes from symbiogenic corruption by a democratic argument that eukaryotes are, by majority, archaeal based on the assumption that their genomes contain a majority of archaeal genes, making them archaea in the classificatory sense.

But what if eukaryotes are actually bacteria in terms of their genomic majority? The trees that molecular phylogeneticists use to classify eukaryotes are based on rRNA or proteins associated with ribosomes—cytosolic ribosomes in the case of eukaryotes. Ribosomes make up ∼40% of a prokaryotic cell's substance by dry weight, so they certainly are important for the object of classification. No one would doubt that eukaryotes have archaeal ribosomes in their cytosol. Archaeal ribosomes in the cytosol could, however, equally be the result of a gradualist origin of eukaryotes from archaea (Martijn and Ettema 2013; Booth and Doolittle 2015) or symbiogenesis involving an archaeal host for the origin of mitochondria (Martin et al. 2017; Martin 2017; Imachi et al. 2020). Ribosomes only comprise ∼50 proteins and three RNAs, whereas the proteins used for phylogenetic classification are only ∼30 in number, or roughly 1% of an average prokaryotic genome (Dagan and Martin 2006). The other 99% of the genome are more difficult to analyze, bringing us back to the question: At the level of whole genomes, are eukaryotes fundamentally archaeal?

Because the availability of complete genome sequences, there have been investigations to determine the proportion of archaeal-related and bacterial-related genes in eukaryotic genomes. Such an undertaking is straightforward for an individual eukaryotic genome, and previous investigations have focused on yeast (Esser et al. 2004; Cotton and McInerney 2010). These indicated that yeast harbors an excess of bacterial genes relative to archaeal genes, conclusions that we borne out in a subsequent, sequence similarity-based investigation for a larger genome sample (Alvarez-Ponce et al. 2013). Genome-wide phylogenetic analyses including plants, animals, and fungi (Pisani et al. 2007; Thiergart et al. 2012), two eukaryotic groups (Rochette et al. 2014), or six eukaryotic supergroups (Ku et al. 2015) reported trees for genes present in eukaryotes and prokaryotes, but fell short of reporting estimates for the proportion of genes in eukaryotic genomes that stem from bacteria and archaea, respectively, whereby all previous estimates have been limited by the small archaeal sample of sequenced genomes for comparison. Here, we have clustered genes from sequenced genomes of 150 eukaryotes, 5,443 bacteria, and 212 archaea. By normalizing for the large bacterial sample through downsampling, we obtain estimates for the proportion of genes in each eukaryote genome that identify prokaryotic homologs, but that only occur in archaea or bacteria, respectively.

Materials and Methods

Sequence Clustering

A total of 19,050,992 protein sequences from 5,655 complete prokaryotic genomes were downloaded from the NCBI RefSeq genomes database Release 78, September 2016 (O'Leary et al. 2016), encompassing 5,443 bacteria and 212 archaea (supplementary table 1a and b, Supplementary Material online). For eukaryotes 3,420,731 protein sequences from 150 sequenced genomes covering a phylogenetically diverse sample were downloaded from NCBI RefSeq (O'Leary et al. 2016), Ensembl Protists (Kersey et al. 2018), JGI (Nordberg et al. 2014), and GenBank (Benson et al. 2015) (supplementary table 1a and c, Supplementary Material online) as appropriate. Protein sequences from the three domains were each clustered separately and homologous clusters were combined as described previously (Carlton et al. 2007; Nelson-Sathi et al. 2015). The reciprocal best BLAST hits (rBBH) (Tatusov et al. 1997) of an all-versus-all BLAST (v. 2.5.0) (Altschul et al. 1997) were calculated for each domain (cut-off: expectation (E) value $\leq$ 1e-10). Pairwise global sequence identities were then generated for each sequence pair with the Needleman–Wunsch algorithm using the program "needle" of the EMBOSS package v. 6.6.0.0 (Rice et al. 2000) with a global identity cut-off $\geq$ 25% for bacterial and archaeal sequence pairs and $\geq$40% global identity for eukaryotic sequence pairs. Protein families were reconstructed applying the domain-specific rBBH to the Markov Chain clustering algorithm (MCL) v. 12-068 (Enright et al. 2002) on the basis of the global pairwise sequence identities, respectively. Due to the large bacterial data set, pruning parameters of MCL were adjusted until no relevant split/join distance between consecutive clusterings was calculated by the "clm dist" application of the MCL program family (-P 180,000 -S 19,800 -R 25,200). MCL default settings were applied for the archaeal and eukaryotic protein clustering. This yielded 16,875 archaeal protein families (422,054 sequences) and 214,519 bacterial protein families (17,384,437 sequences) with at least five sequences each and 239,813 eukaryotic protein families (1,545,316 sequences) with sequences present in at least two species (supplementary table 6, Supplementary Material online). To combine eukaryotic clusters with bacterial or archaeal clusters, the reciprocal best cluster approach (Ku et al. 2015) was applied with 50% best-hit correspondence and 30% BLAST local pairwise sequence identity of the interdomain hits between eukaryote and prokaryote sequences. Eukaryotic clusters having homologs in both bacterial and archaeal clusters were merged with their prokaryotic homologs as described (Ku et al. 2015). The cluster merging procedure left 752 eukaryotic clusters that had ambiguous (multiple) prokaryote cluster assignment, these were excluded from further analysis and 236,474 eukaryote clusters connected to no homologous prokaryotic cluster (eukaryote-specific, ESC, supplementary table 2,

Supplementary Material online) at the cut-offs employed here.

Assignment of Bacterial or Archaeal Origin

Because the number of prokaryotic sequences clustered was large, the 2,368 EPCs that were assigned one bacterial or one archaeal cluster exclusively were rechecked for homologs from the remaining prokaryotic domain at the E value $\leq$ 1e−10, global identity $\geq$ 25% threshold. The 266 cases so detected were excluded from bacterial–archaeal origin assignment, yielding 2,102 EPCs (supplementary table 2, Supplementary Material online, indicated by asterisks). The clusters generated from rBBH (E value $\leq$ 1e−10, global identity $\geq$ 25%) of all-versus-all BLAST of the 19,050,992 prokaryotic protein sequences are provided as supplementary material (supplementary table 6, Supplementary Material online). Downsampling to adjust for the overrepresentation of bacterial strains in the prokaryotic data set compared with the number of archaeal organisms was performed by generating 1,000 data sets with 212 bacterial taxa selected randomly according to the distribution of genera in the whole data set (supplementary table 7, Supplementary Material online). The sequences of the examined 212 archaeal and bacterial taxa were located in the 2,102 EPCs and each eukaryotic organism in the identified clusters was assigned to "bacterial," or "archaeal" depending on the domain of the prokaryotic cluster in the EPC. Each eukaryotic genome was only counted once per EPC and assigned the respective prokaryotic label to prevent overrepresentation of duplication rich organisms. This procedure was performed for all 1,000 downsized bacterial data sets for each EPC, the mean of 1,000 samples was scored (supplementary table 3, Supplementary Material online).

Cluster Annotation

Protein annotation information according to the BRITE (Biomolecular Reaction pathways for Information Transfer and Expression) hierarchy was downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG v. September 2017) website (Kanehisa et al. 2016), including protein sequences and their assigned function according to the KO numbers (Suppl. Material 8a, b). The sequences of each protein family from the 2,587 EPCs were locally aligned with "blastp" to the KEGG database to identify the annotation for each protein. In order to assign each protein to a KEGG function, only the best BLAST hit of the given protein with an E value $\leq$ 1e−10 and alignment coverage of 80% was selected. After assigning a function based on the KO numbers of KEGG for each protein in the EPCs, the majority rule was applied to identify the function for each cluster. The occurrence of the function of each protein was added and the most prevalent function was assigned for each cluster (supplementary table 4, Supplementary Material online). Poorly

characterized sequences or sequences with no assigned function were ignored, resulting in 1,836 clusters with annotations.

Presence and Absence of EPCs across Genomes

Presence of absence of genes in a cluster for each genome were plotted as a 2,587 × 5,805 binary matrix, rows were sorted taxonomically, columns were sorted in ascending order left to right according to density of distribution within eukaryotic groups. Hacrobia and SAR were treated as a eukaryotic group for clusters they shared with Archaeplastida only; these clusters reflect secondary symbioses (41).

Results

Using the MCL algorithm, we generated clusters for 19,050,992 protein sequences from 5,443 bacteria and 212 archaea with 3,420,731 protein sequences from 150 eukaryotes (see Materials and Methods) (supplementary table 1a–c, Supplementary Material online) spanning six eukaryotic supergroups (fig. 2a). This yielded 239,813 clusters containing eukaryotic sequences: 236,474 eukaryote-specific clusters and 2,587 clusters (1% of all eukaryote clusters) that contained prokaryotic homologs at the stringency levels employed here, as well as 752 eukaryotic clusters that were excluded from the analysis as they were assigned multiple prokaryote clusters. Of the 2,587 eukaryote–prokaryote clusters (EPCs), 1,853 contained only eukaryotes and bacteria, 515 of which contained only eukaryotes and archaea. Among the 2,587 EPC clusters, 8% (219) contained sequences from at least two eukaryotes and at least five prokaryotes spanning bacteria and archaea (see supplementary table 2, Supplementary Material online), which were not considered further for our estimates because here we sought estimates where the decision regarding bacterial or archaeal origin was independent of phylogenetic inference, which is possible for 92% of eukaryotic clusters that contain prokaryotic sequences. All sequences had unique cluster assignments, no sequences occurred in more than one cluster. That 1,853 clusters contained only eukaryotes and bacteria whereas 515 contained only eukaryotes and archaea appears to suggest a 3.6-fold excess of bacterial genes in eukaryotes, but bacterial genes are 25-fold more abundant in the data. For those genes that each eukaryote shares with prokaryotes, we estimated the proportion and number of genes having homologs only in archaea and only in bacteria, respectively, by downsampling the 25-fold excess of bacterial genomes in the sample in 1,000 subsamples of 212 bacteria and 212 archaea.

The proportion of bacterial and archaeal genes for each eukaryote is shown in figure 2b. Overall, 44% of eukaryotic sequences are archaeal in origin and 56% are bacterial. Across 150 genomes, eukaryotes possess 12% more bacterial genes than archaeal genes. There are evident group specific
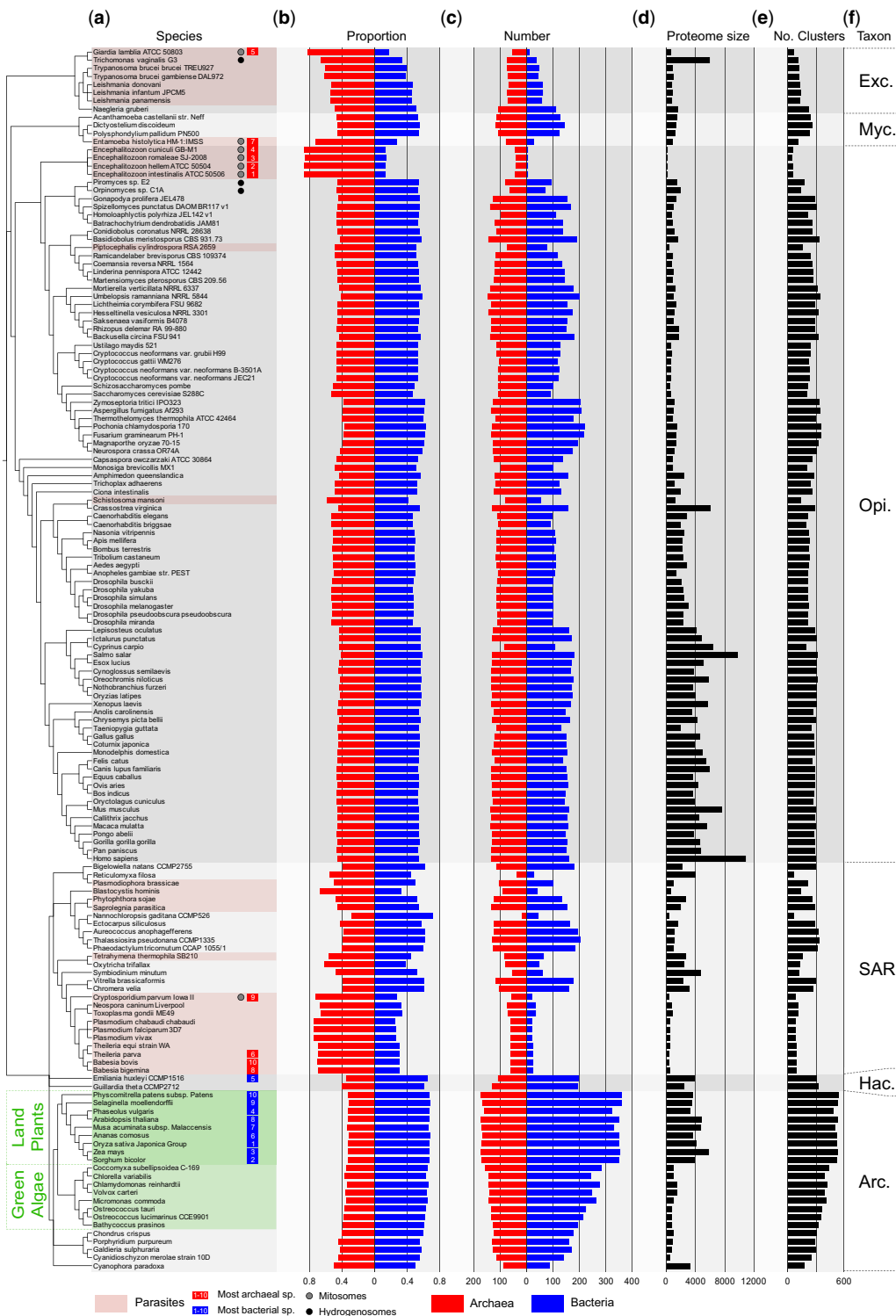
Fig. 2.—Bacterial and archaeal genes in eukaryotic genomes. Protein sequences from 150 eukaryotic genomes and 5,655 prokaryotic genomes (5,433 bacteria and 212 archaea) were clustered into eukaryote–prokaryote clusters (EPC) using the MCL algorithm (Enright et al. 2002) as described (Ku et al. 2015). To account for overrepresentation of bacterial sequences in the clusters, bacterial genomes were downsampled in 1,000 data sets of 212 randomly selected bacterial organisms, the means were plotted. The eukaryotic sequences in the EPCs that cluster exclusively with bacterial or archaeal homologs were labeled bacterial (blue) or archaeal (red) accordingly. (*a*) Eukaryotic lineages and genomes were grouped by taxonomy. Numbers next to the species name on the left side indicate the ten most bacterial (blue) and archaeal (red) genomes, respectively. (*b*) The avg. relative proportion of bacterial and archaeal genes per genome. (*c*) The number of eukaryotic clusters with bacterial or archaeal homologs is shown. (*d*) The proteome size for the genome. (*e*) The sum of all

Table 1

Proportion of Bacterial and Archaeal Derived Genes in Eukaryotic Genomes

Group	Archaeal	Bacterial
All eukaryotes	0.44	0.56
All without plastids[a]	0.47	0.53
All with plastids[b]	0.39	0.61
Land plants	0.33	0.67
Opisthokonts	0.46	0.54
Hacrobia	0.38	0.62
SAR	0.50	0.50
Archaeplastida	0.36	0.64
Mycetozoa	0.50	0.50
Excavata	0.58	0.42
Parasites[c]	0.62	0.38

[a]All except members of SAR, Hacrobia, and Archaeplastida as designated in supplementary table 3, Supplementary Material online.

[b]All members of SAR, Archaeplastida, and Hacrobia as designated in supplementary table 3, Supplementary Material online.

[c]Eukaryotes scored as parasites are designated in figure 2. Among 239,813 clusters containing eukaryote sequences 2,587 clusters (1%) contained prokaryotic homologs at the stringency levels employed here.

differences (fig. 2b). If we look only at organisms that never harbored a plastid, the excess of bacteria genes drops from 56% to 53%. If we look only at groups that possess plastids the proportions of bacterial homologs increases to 61% versus 39% archaeal (table 1, supplementary table 3, Supplementary Material online). Note that our estimates are based on the number of clusters, meaning that gene duplications do not figure into the estimates. A bacterial derived gene that was amplified by duplication to 100 copies in each land plant genome is counted as one bacterial derived gene. This is seen in figure 2 for *Trichomonas*, where a large number on gene families have expanded in the *Trichomonas* lineage (Carlton et al. 2007), reflected in a conspicuously large proteome size (fig. 2d), but a similar number of clusters (fig. 2e) as neighboring taxa.

The proportions for different eukaryotic groups are shown in table 1. Land plants have the highest proportion of bacterial derived genes at 67%, or a 2:1 ratio of bacterial genes relative to archaeal. The eukaryote with the highest proportion of bacterial genes in our sample is rice, with 67.1% bacterial and 32.9% archaeal genes. The higher proportion of bacterial genes in plastid containing eukaryotes relative to other groups corresponds with the origin of the plastid and gene transfers to the nucleus (Ku et al. 2015). The eukaryote with the highest proportion of archaeal genes in our sample are the human parasite *Encephalitozoon intestinalis* and the rabbit parasite *Encephalitozoon cuniculi*, with 86% archaeal and 14%

bacterial derived genes. Parasitic eukaryotes have the largest proportions of archaeal genes, but not by novel acquisitions, rather by having lost large numbers of bacterial genes as a result of reductive evolution in adaptation to nutrient rich environments. This is evident in figure 2c, where the numbers of archaeal and bacterial genes per genome are shown. Parasites, with their reduced genomes, such as *Giardia lamblia*, *Trichomonas vaginalis*, or *Encephalitozoon* species, appear more archaeal. The number of archaeal, or bacterial genes in an organism does not correlate with genome size (supplementary fig. 1, Supplementary Material online, Pearson correlation coefficient: archaeal $r^2 = 0.38$, bacterial $r^2 = 0.33$).

Opisthokonts generally have a more even distribution of bacterial and archaeal homologs in their genomes but are still slightly more bacterial (54%, table 1 and supplementary table 3, Supplementary Material online). The black and gray dots in figure 2a indicate organisms that possess reduced forms of mitochondria, hydrogenosomes (black) or mitosomes (gray) (van der Giezen et al. 2005). The ten most archaeal or bacterial organisms are indicated by a red or blue rectangle, respectively. The most archaeal eukaryotes are all parasites (highlighted in red) and have undergone reductive evolution, also with respect to their mitochondria, which are often reduced to mitosomes (fig. 2a). Nine of the ten most bacterial organisms in the sample are plants (highlighted in green) with the fifth most bacterial organism being one of the only two Hacrobia in the data set.

The functional distinction that eukaryotic genes involved in the eukaryotic genetic apparatus and information processing tend to reflect an archaeal origin whereas genes involved in eukaryotic biochemical and metabolic processes tend to reflect bacterial origins (Martin and Müller 1998; Rivera et al. 1998) has been borne out for yeast (Esser et al. 2004; Cotton and McInerney 2010) and small genome samples (Thiergart et al. 2012; Alvarez-Ponce et al. 2013; Rochette et al. 2014). The distributions of eukaryotic genes per genome that have archaeal or bacterial homologs across the respective KEGG function category at the first level (metabolism, genetic information processing, environmental information processing, cellular processes, and organismal systems) are shown in figure 3. The category human diseases is not shown, as only very few proteins in the EPCs were so annotated. The categories genetic information processing (information) and metabolism account for 90% of all annotated eukaryotic sequences in the EPCs (supplementary table 4, Supplementary Material online). In the category metabolism, 67.6% of eukaryotic genes are bacterial

Fig. 2.—Continued

eukaryotic sequences in the eukaryote–prokaryote clusters. (f) Taxonomic groups are labeled on the far right panel (Arc.—Archaeplastida, Exc.—Excavata, Hac.—Hacrobia, Myc.—Mycetozoa, Opi.—Opisthokonts). Highlighted in green is the branch with the taxa of plants and green algae, parasites are highlighted in red. The black dots indicate organisms with hydrogenosomes, the gray dot indicates organisms with mitosomes.
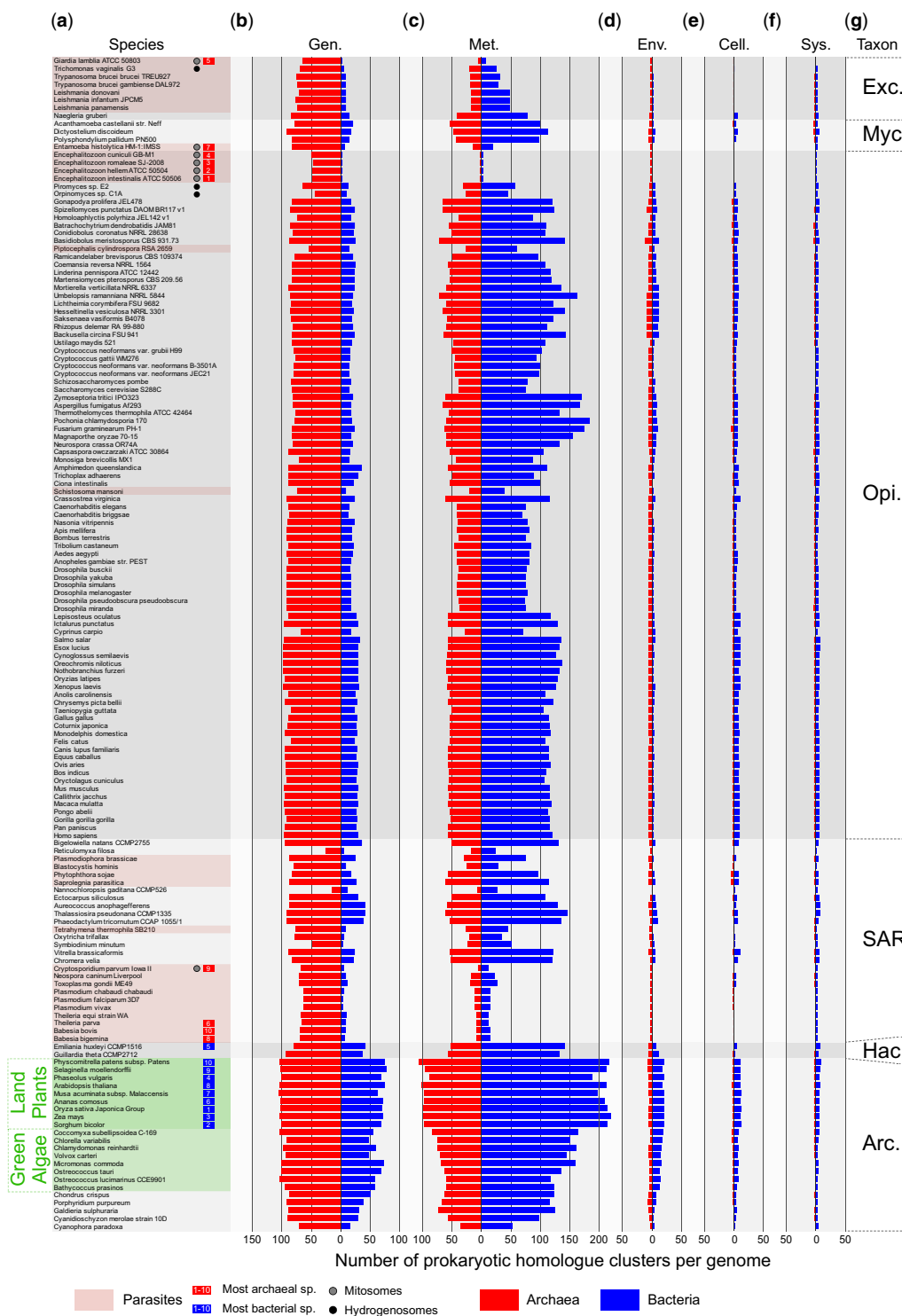
FIG. 3.—Functional categories. Protein sequences from 150 eukaryotic genomes and 5,655 prokaryotic genomes were clustered into 2,587 eukaryote–prokaryote clusters (EPC) (Ku et al. 2015). Sorted according to a reference tree for eukaryotic lineages generated from the literature and taxonomic groups are labeled. The red bars indicate eukaryotic gene families that are archaeal in origin, blue indicates a bacterial origin of the gene family. Functional annotations according to the KEGG BRITE hierarchy on the level A was assigned for each EPC, identifying the function for each sequence in the protein cluster by performing a protein BLAST against the KEGG database and then applying the most prevalent function per protein family. Only the categories (b) "Genetic Information Processing" (Gen.), (c) "Metabolism" (Met.), (d) "Environmental Information Processing" (Env.), (e) "Cellular Processes" (Cell.), and (f) "Organismal Systems" (Sys.) are depicted, as the label "Human Diseases" was hardly represented. Species names are indicated in column (a) and taxonomic groups (f) are labeled on the far right panel (Arc.—Archaeplastida, Exc.—Excavata, Hac.—Hacrobia, Myc.—Mycetozoa, Opi.—Opisthokonts). Highlighted in green is the branch uniting land plants and green algae; the black and gray dots indicate organisms with hydrogenosomes or mitosomes, respectively.
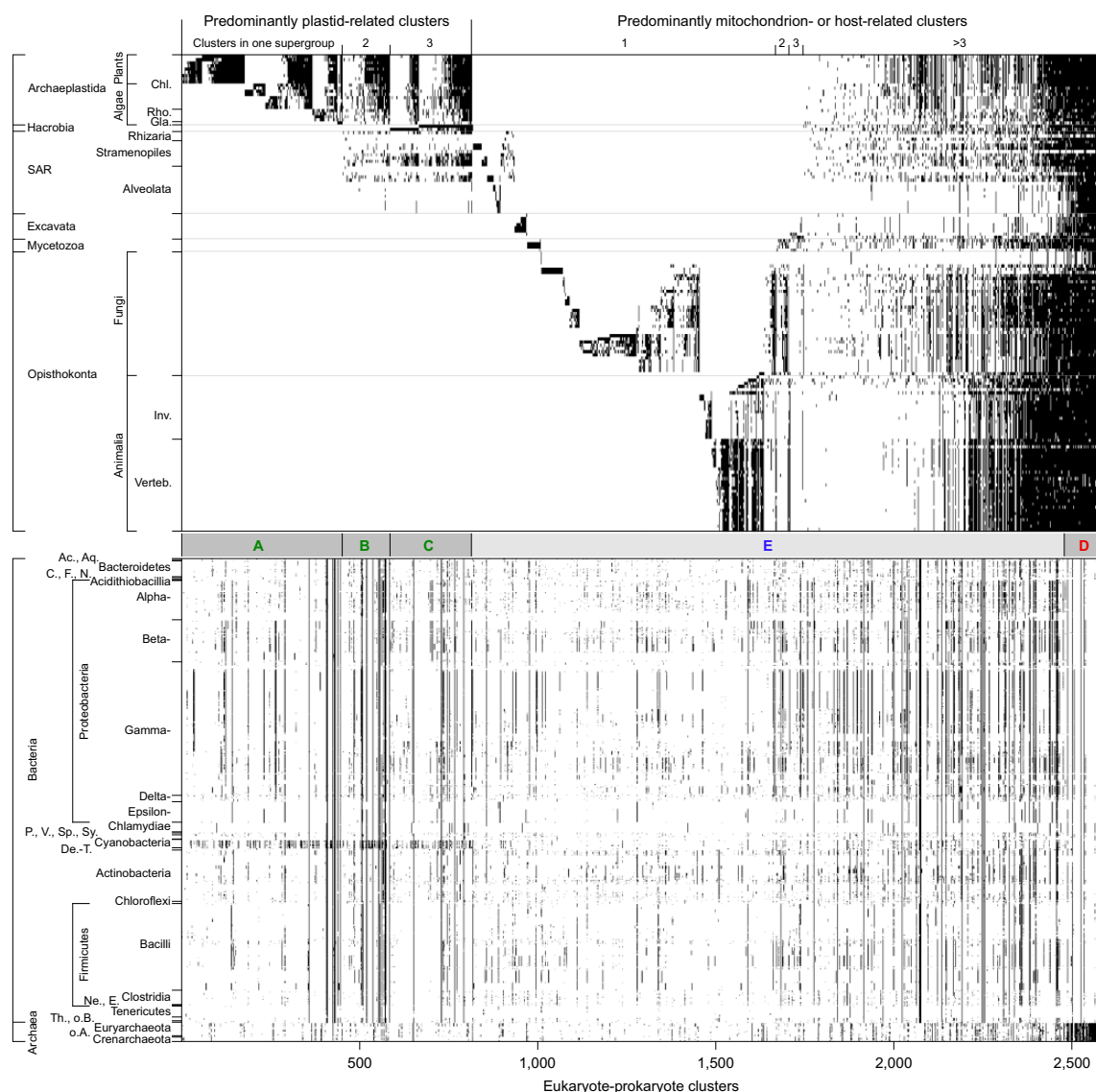
FIG. 4.—Gene sharing matrix. Each black tick represents the presence of a gene in the respective taxon. First, the 2,587 EPCs (*x* axis) were sorted according to their distribution across the six eukaryotic supergroups with the photosynthetic lineages on the left (block *A–C*). Host- or mitochondrion-related genes distributed across the six supergroups are depicted in block *E*. Clusters with mostly archaeal homologs are indicated in block *D* (Chl.—Chloroplastida, Rho.—Rhodophyta, Gla.—Glaucophyta, Inv.—Invertebrates, Verteb.—Vertebrates; Ac.—Acidobacteria, Aq.—Aquificiae, C.—Chlorobi, F.—Fusobacteria, N.—Nitrospirae, P.—Planctomycetes, V.—Verrucomicrobia, Sp.—Spirochaetes, Sy.—Synergistetes, De.-T.—Deinococcus-Thermus, Ne.—Negativicutes, E.—Erysipelotrichia, Th.—Thermotogae, o. B.—other Bacteria, o. A.—other Archaea).

whereas 76.9% of EPCs involved in information are archaeal. The distinction between informational and metabolic genes first described for yeast appears to be valid across all eukaryotic genomes.

The distribution of the genes in the 2,587 EPCs across genomes for six supergroups is depicted in figure 4. The order of eukaryotic and prokaryotic organisms (rows) can be found in supplementary table 5, Supplementary Material online. Block A represents only Archaeplastida, block B depicts genes found in Archaeplastida and SAR, block C encompasses all

genes that are distributed across the three taxa that contain plastids; Archaeplastida, SAR, and Hacrobia. The lower part of the figure shows the prokaryotic homologous genes. Cyanobacterial genes are especially densely distributed across blocks A–C. Genes that are predominantly mitochondrion- or host-related are indicated in blocks D and E. Eukaryotic genes that are universally distributed across the six supergroups are mainly archaeal in origin (block D). Especially organisms with reduced genomes, such as parasites (marked with asterisks on the right), have lost genes associated with metabolism,

leaving them mainly archaeal (fig. 4). In the wake of symbiogenic mergers, which are very rare in evolution, gene loss sets in, whereby gene loss is very common in eukaryote genome evolution, one of its main underlying themes (Ku et al. 2015; Deutekom et al. 2019).

The estimates we obtain are based on a sample of genes that meet the clustering thresholds employed here. Many eukaryotic genes are inventions of the eukaryotic lineage in terms of domain structure and sequence identity. Those genes either arose in eukaryotes de novo from noncoding DNA, or they arose through sequence divergence, recombination, and duplication involving preexisting coding sequences, the bacterial and archaeal components of which should reflect that demonstrable in the conserved fraction of genes analyzed here. It is possible that archaeal genes and domains are more prone to recombination and rapid sequence divergence than bacterial domains are, but the converse could also be true and there is no a priori evidence to indicate that either assumption applies across eukaryotic supergroups. Hence with some caution, our estimates, which are based on the conserved fraction of sequences only, should in principle apply for the archaeal and bacterial components of the genome as a whole.

Discussion

Guided by endosymbiotic theory, evidence for genomic chimaerism in eukaryotes emerged in the days before there were sequenced genomes to analyze (Martin and Cerff 1986; Brinkmann et al. 1987; Zillig et al. 1989; Martin et al. 1993; Golding and Gupta 1995; Martin and Schnarrenberger 1997). The excess of bacterial genes in eukaryotic genomes we observe here has been observed before, but with smaller samples and with different values. In a sample of 15 archaeal and 45 bacterial genomes using sequence comparisons, Esser et al. (2004) found that ~75% of yeast genes that have prokaryotic homologs are bacterial in origin. Cotton and McInerney (2010) used 22 archaea and 197 bacteria to investigate the yeast genome and also found an excess of bacterial genes. Using 14 eukaryotic genomes, 52 bacteria and 52 archaea, Alvarez-Ponce et al. (2013) found a 3:1 excess of bacterial to archaeal genes in many eukaryotes, similar to the result of Esser et al. (2004), but they also observed an archaeal majority of genes in intracellular parasitic protists including *Giardia* and *Entamoeba*, as we observe here. It was, however, unknown if the genes studied by Alvarez-Ponce et al. (2013) traced to the LECA, hence it was unknown whether the archaeal excess in parasites was due to loss (as opposed to gain in nonparasitic lineages), and phylogenetic trends of gain or loss could not be observed.

Rivera and Lake (2004) constructed trees from two eukaryotes, three archaea, and three bacteria with homologs detected by searches with a bacterial and an archaeal query ("conditioning") genome, they detected trees indicating a bacterial origin and trees indicating an archaeal origin for the eukaryotic gene; the conflicting signals were combined into a ring. Thiergart et al. (2012) generated alignments and trees for homologs from 27 eukaryotes and 994 prokaryotes, they found an excess of bacterial genes and 571 eukaryotic genes with prokaryotic homologs that trace to the LECA based on monophyly. Rochette et al. (2014) generated trees and alignments for homologs from 64 eukaryotes, 62 archaea, and 820 bacteria, they found 434 eukaryote genes with prokaryote homologs that trace to the LECA. Ku et al. (2015) generated alignments and trees for genes shared among 55 eukaryotes, 134 archaea, and 1,847 bacteria using similar clustering methods and clustering thresholds as used here, they found that ~90% of 2,585 genes shared by prokaryotes and eukaryotes indicate monophyly, hence a single acquisition corresponding to the origin of mitochondria (eukaryotes) or the cyanobacterial origin of plastids. That observation, together with the phylogenetic pattern of lineage-specific distributions observed here (figs. 2 and figs. 3), indicates that gene gains at eukaryote origin and at the origin of primary and secondary plastids were followed by lineage-specific differential loss, which was also noted by Ku et al. (2015), but for a smaller genome sample than that investigated here. That we observe a smaller excess of bacterial genes than that reported by Esser et al. (2004) or Alvarez-Ponce et al. (2013) is probably due to our larger archaeal sample and the use of downsampling to reduce bacterial bias.

Using a sample of 5,655 prokaryotic and 150 eukaryotic genomes and downsampling procedures to correct for the overabundance of bacterial genomes versus archaeal genomes for comparisons, we have obtained estimates for the proportion of archaeal and bacterial genes per genome in eukaryotes based on gene distributions. We found that the members of six eukaryotic supergroups possess a majority of bacterial genes over archaeal genes. If eukaryotes were to be classified by genome-based democratic principle, they would be have to be grouped with bacteria, not archaea. The excess of bacterial genes disappears in the genomes of intracellular parasites with highly reduced genomes, because the bacterial genes in eukaryotes underpin metabolic functions that can be replaced by metabolites present in the nutrient rich cytosol of the eukaryotic cells that parasites infect. The functions of the ribosome and genetic information processing cannot be replaced by nutrients, hence reductive genome evolution in parasites leads to preferential loss of bacterial genes and leaves archaeal genes remaining. In photosynthetic eukaryote lineages, the genetic contribution of plastids to the collection of nuclear genomes is evident in our analyses, both in lineages with primary plastids descended directly from cyanobacteria and in lineages with plastids of secondary symbiotic origin. The available sample of archaeal genomes is still limiting for comparisons of the kind presented here.

As improved culturing and sequencing of complete archaeal genomes progresses, new lineages are being characterized at the level of scanning electron microscopy that
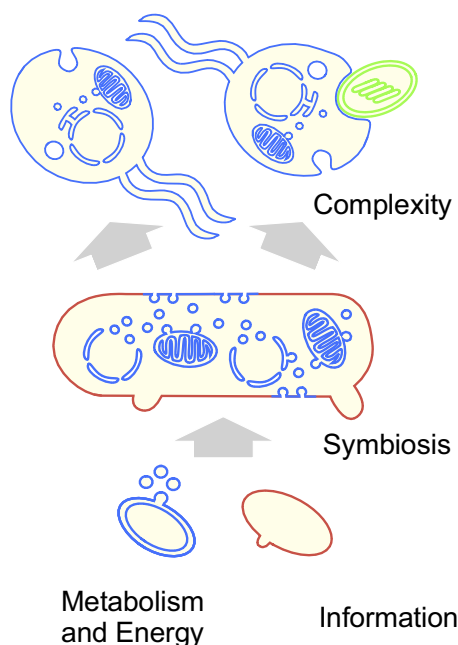
Fig. 5.—Bacterial and archaeal contributions to eukaryotes. Schematic representation of eukaryote origin involving an archaeal host and a mitochondrial symbiont that transforms the host via gene transfer from the endosymbiont (Martin and Müller 1998; Imachi et al. 2020). The model combines elements of different proposals: Bacterial outer membrane vesicles at the origin of the eukaryotic endomembrane system (Gould et al. 2016); archaeal outer membrane vesicles at the origin of host membrane protrusions enabling endosymbiosis without phagocytosis (Imachi et al. 2020); a syncytial eukaryote common ancestor (Garg and Martin 2016); eukaryote origin starting an archaeal host and a bacterial symbiont brought into physical symbiotic interaction by anaerobic syntrophic interactions (Martin and Müller 1998; Imachi et al. 2020); a combination of information (host) plus metabolism and energy (symbiont) (Martin 2017; Brunk and Martin 2019) at eukaryote origin.

branch, in ribosomal trees, as sisters to the host lineage at eukaryote origin (Imachi et al. 2020). These archaea are, however, not complex like eukaryotes, rather they are prokaryotic in size and shape and unmistakably prokaryotic in organization (Imachi et al. 2020). That is, the closer microbiologists hone in on the host lineage for the origin of mitochondria, the steeper the evolutionary grade between prokaryotes and eukaryotes becomes, in agreement with the predictions of symbiotic theory (Imachi et al. 2020) (fig. 5) and in contrast to the expectations of gradualist theories for eukaryote origin (Martin 2017). At the same time, the analyses presented here uncover a bacterial majority of genes in eukaryotic genomes, a majority that traces to the LECA (Ku et al. 2015), which is also in line with the predictions of symbiotic theory. The most likely biological source of the bacterial majority of genes in the LECA is the mitochondrial endosymbiont (Ku et al. 2015). Genomes record their own history. Eukaryotic genomes testify to the role of endosymbiosis in evolution.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

Author Contributions

Data processing and analysis: J.B.; Manuscript composition: J.B. and W.F.M.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.

Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO. 2013. Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc Natl Acad Sci USA. 110(17):E1594–E1603.

Archibald JM. 2015. Endosymbiosis and eukaryotic cell evolution. Curr Biol. 25(19):R911–R921.

Arndt N, Nisbet E. 2012. Processes on the young Earth and the habitats of early life. Annu Rev Earth Planet Sci. 40(1):521–549.

Barlow LD, Nývltová E, Aguilar M, Tachezy J, Dacks JB. 2018. A sophisticated, differentiated Golgi in the ancestor of eukaryotes. BMC Biol. 16(1):27.

Benson DA, et al. 2015. GenBank. Nucleic Acids Res. 43(D1):D30–D35.

Booth A, Doolittle WF. 2015. Eukaryogenesis, how special really? Proc Natl Acad Sci USA. 112(33):10278–10285.

Brinkmann H, Martinez P, Quigley F, Martin W, Cerff R. 1987. Endosymbiotic origin and codon bias of the nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize. J Mol Evol. 26(4):320–328.

Brunk CF, Martin WF. 2019. Archaeal histone contributions to the origin of eukaryotes. Trends Microbiol. 27(8):703–714.

Carlton JM, et al. 2007. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. Science 315(5809):207–212.

Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. Mol Biol Evol. 22(4):1053–1066.

Cotton JA, McInerney JO. 2010. Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. Proc Natl Acad Sci USA. 107(40):17252–17255.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci USA. 105(51):20356–20361.

Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P. 2017. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLoS Genet. 13(6):e1006810.

Da Cunha V, Gaia M, Nasir A, Forterre P. 2018. Asgard archaea do not close the debate about the universal tree of life topology. PLoS Genet. 14(3):e1007215.

Dagan T, Martin W. 2006. The tree of one percent. Genome Biol. 7(10):118.

Deutekom ES, Vosseberg J, van Dam TJP, Snel B. 2019. Measuring the impact of gene prediction on gene loss estimates in Eukaryotes by quantifying falsely inferred absences. PLoS Comput Biol. 15(8):e1007301.

Embley TM, Hirt RP. 1998. Early branching eukaryotes? Curr Opinion Genet Dev. 8(6):624–629.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. Nature 440(7084):623–630.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30(7):1575–1584.

Esser C, et al. 2004. A genome phylogeny for mitochondria among alphaproteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol. 21(9):1643–1660.

Field MC, Sali A, Rout MP. 2011. On a bender—BARs, ESCRTs, COPs, and finally getting your coat. J Cell Biol. 193(6):963–972.

Garg SG, Martin WF. 2016. Mitochondria, the cell cycle, and the origin of sex via a syncytial eukaryote common ancestor. Genome Biol Evol. 8(6):1950–1970.

Golding GB, Gupta RS. 1995. Protein based phylogenies support a chimeric origin of the eukaryotic genome. Mol Biol Evol. 12(1):1–6.

Gould SB, Garg SG, Martin WF. 2016. Bacterial vesicle secretion and the evolutionary origin of the eukaryotic endomembrane system. Trends Microbiol. 24(7):525–534.

Gould SB, Waller RF, McFadden GI. 2008. Plastid evolution. Annu Rev Plant Biol. 59(1):491–517.

Imachi H, et al. 2020. Isolation of an archaeon at the prokaryote-eukaryote interface. Nature 577(7791):519–525.

Javaux EJ, Lepot K. 2018. The Paleoproterozoic fossil record: implications for the evolution of the biosphere during Earth's middle-age. Earth Sci Rev. 176:68–86.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44(D1):D457–D462.

Kersey PJ, et al. 2018. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res. 46(D1):D802–D808.

Ku C, et al. 2015. Endosymbiotic origin and differential loss of eukaryotic genes. Nature 524(7566):427–432.

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

Martin W, Brinkmann H, Savona C, Cerff R. 1993. Evidence for a chimaeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. Proc Natl Acad Sci USA. 90(18):8692–8696.

Martin W, Cerff R. 1986. Prokaryotic features of a nucleus-encoded enzyme: cDNA sequences for chloroplast and cytosolic glyceraldehyde-3-phosphate dehydrogenases from mustard (Sinapis alba). Eur J Biochem. 159(2):323–331.

Martin W, et al. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA. 99(19):12246–12251.

Martin W, Herrmann RG. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol. 118(1):9–17.

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

Martin W, Schnarrenberger C. 1997. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. Curr Genet. 32(1):1–18.

Martin WF. 2017. Symbiogenesis, gradualism, and mitochondrial energy in eukaryote origin. Period Biol. 119(3):141–158.

Martin WF, Tielens AGM, Mentel M, Garg SG, Gould SB. 2017. The physiology of phagocytosis in the context of mitochondrial origin. Microbiol Mol Biol Rev. 81(3):e00008–e00017.

McInerney JO, O'Connell M, Pisani D. 2014. The hybrid nature of the eukaryota and a consilient view of life on Earth. Nat Rev Microbiol. 12(6):449–455.

Nelson-Sathi S, et al. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517(7532):77–80.

Newman D, Whelan F, Moore M, Rusilowicz M, McInerney JO. 2019. Reconstructing and analysing the genome of the Last Eukaryote Common Ancestor to better understand the transition from FECA to LECA. bioRxiv 538264. https://doi.org/10.1101/538264

Nordberg H, et al. 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucl Acids Res. 42(D1):D26–D31.

O'Leary NA, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44(D1):D733–D745.

Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol. 24(8):1752–1760.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16(6):276–277.

Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci USA. 95(11):6239–6244.

Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431(7005):152–155.

Rochette NC, Brochier-Armanet C, Gouy M. 2014. Phylogenomic test of the hypotheses for the evolutionary origin of eukaryotes. Mol Biol Evol. 31(4):832–845.

Roger AJ, Muñoz-Gómez SA, Kamikawa R. 2017. The origin and diversification of mitochondria. Curr Biol. 27(21):R1177–R1192.

Schlacht A, Herman EK, Klute MJ, Field MC, Dacks JB. 2014. Missing pieces of an ancient puzzle: evolution of the eukaryotic membrane-trafficking system. Cold Spring Harb Perspect Biol. 6(10):a016048.

Shiflett AM, Johnson PJ. 2010. Mitochondrion-related organelles in eukaryotic protists. Annu Rev Microbiol. 64(1):409–429.

Tashiro T, et al. 2017. Early trace of life from 3.95 Ga sedimentary rocks in Labrador, Canada. Nature 549(7673):516–518.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278(5338):631–637.

Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. 2012. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol. 4(4):466–485.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 5(2):123–135.

Ueno Y, Yamada K, Yoshida N, Maruyama S, Isozaki Y. 2006. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. Nature 440(7083):516–519.

van der Giezen M. 2009. Hydrogenosomes and mitosomes: conservation and evolution of functions. J Eukaryot Microbiol. 56(3):221–231.

van der Giezen M, Tovar J, Clark CG. 2005. Mitochondrion-derived organelles in protists and fungi. Int Rev Cytol. 244:177–227.

Vosseberg J, Snel B. 2017. Domestication of self-splicing introns during eukaryogenesis: the rise of the complex spliceosomal machinery. Biol Direct. 12(1):30.

Whitman WB. 2009. The modern concept of the procaryote. J Bacteriol. 191(7):2000–2005.

Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504(7479):231–236.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA. 87(12):4576–4579.

Zillig W, et al. 1989. Did eukaryotes originate by a fusion event? Endocyt Cell Res. 6:1–25.

Associate editor: Davide Pisani