# Supertrees Disentangle the Chimerical Origin of Eukaryotic Genomes

*Davide Pisani, James A. Cotton, and James O. McInerney*

Department of Biology, The National University of Ireland, Maynooth, Maynooth, County Kildare, Ireland

Eukaryotes are traditionally considered to be one of the three natural divisions of the tree of life and the sister group of the Archaebacteria. However, eukaryotic genomes are replete with genes of eubacterial ancestry, and more than 20 mutually incompatible hypotheses have been proposed to account for eukaryote origins. Here we test the predictions of these hypotheses using a novel supertree-based phylogenetic signal-stripping method, and recover supertrees of life based on phylogenies for up to 5,741 single gene families distributed across 185 genomes. Using our signal-stripping method, we show that there are three distinct phylogenetic signals in eukaryotic genomes. In order of strength, these link eukaryotes with the Cyanobacteria, the Proteobacteria, and the Thermoplasmatales, an archaebacterial (euryarchaeotes) group. These signals correspond to distinct symbiotic partners involved in eukaryote evolution: plastids, mitochondria, and the elusive host lineage. According to our whole-genome data, eukaryotes are hardly the sister group of the Archaebacteria, because up to 83% of eukaryotic genes with a prokaryotic homolog have eubacterial, not archaebacterial, origins. The results reject all but two of the current hypotheses for the origin of eukaryotes: those assuming a sulfur-dependent or hydrogen-dependent syntrophy for the origin of mitochondria.

## Introduction

There is considerable interest in the deepest branches of the tree of life (Woese and Fox 1977; Gogarten et al. 1989; Iwabe et al. 1989; Searcy and Hixon 1991; Rivera and Lake 1992; Martin and Muller 1998; Doolittle 1999; Martin et al. 2001; Creevey et al. 2004; Rivera and Lake 2004; Beiko, Harlow, and Ragan 2005; Ciccarelli et al. 2006; Embley and Martin 2006; Kurland, Collins, and Penny 2006; Lopez-Garcia and Moreira 2006; Margulis et al. 2006; Pace 2006; de Duve 2007; Doolittle and Bapteste 2007; Goldenfeld and Woese 2007), but understanding their relationships is proving difficult. In particular, evidence provided by the study of morphology, biochemistry, and alternative molecular markers has led to more than 20 hypotheses for the origin of eukaryotes (Martin et al. 2001; Embley and Martin 2006). These hypotheses can be divided into two groups, proposing either that eukaryotes are a primary lineage of life, or that they are a chimeric lineage originating from the symbiosis of two prokaryotes. Because these hypotheses predict different sister group relationships for eukaryotic genes (see Table 1) they can be tested using phylogenetic methods on a genomic scale.

Recently, two studies (Rivera and Lake 2004; Ciccarelli et al. 2006) investigated the evolutionary history of early life using multiple genes, but they reached irreconcilable conclusions. Rivera and Lake (2004) found that eukaryotes are a chimeric lineage and that the “Tree of Life” is therefore a ring: a weakly connected network. Ciccarelli et al. (2006) concluded that the eukaryotes are a primary lineage of life, and that the “Tree of Life” is thus a tree. Both studies are problematic. Rivera and Lake (2004) had limited species sampling (seven species), and they used a genome content method that only uses patterns of gene presence and absence to reconstruct phylogenies (McInerney 2006). Ciccarelli et al. (2006) used a small set of (31) nonrandomly selected informational genes (Rivera et al. 1998), corresponding to 1% of the average bacterial genome, which could have resulted in a biased view of evolution (Dagan and Martin 2006). We used supertrees (see *Materials and Methods*) to perform phylogenomic analyses (Delsuc, Brinkmann, and Philippe 2005) of 168 prokaryotic genomes and 17 eukaryotes. Supertrees use more of the information in complete genomes than genome content methods, as they use the evolutionary history of genes, not simply their presence or absence. Here, they were used to derive a prokaryotic tree of life based on 5,741 genes, and to resolve the relationships of the eukaryotes using data sets of up to 2,807 genes.

If eukaryotes are chimeric (Martin et al. 2001; Rivera and Lake 2004; Embley and Martin 2006; Margulis et al. 2006) different genes will trace their ancestry to different prokaryotic groups (Esser et al. 2004). Their genomes will therefore contain multiple phylogenetic signals (Pisani and Wilkinson 2002) which will support different sister group relationships. By examining phylogenies for individual genes, we can assess the strength of each signal in terms of the number of genes with particular evolutionary histories and so assess the reported hypotheses concerning eukaryote origins (see Table 1). This approach is not without difficulties: random errors may cause disagreement between phylogenies with the same evolutionary history, and phylogenetic artifacts caused by, for example, long-branch attraction (e.g., Pisani 2004) and compositional heterogeneity (Delsuc, Brinkmann, and Philippe 2005) are expected to affect ancient phylogenies (Gribaldo and Philippe 2002). However, as pointed out by Lake (2007), it is only from the study of complete genomes that we may hope to understand the origin of the eukaryotes, and it is thus surprising that no systematic, gene-by-gene assessment of the deep evolutionary history of this group has previously been attempted. To further visualize these signals, we devised a supertree-based phylogenetic signal-stripping approach. Essentially, we recovered a supertree using our entire data set, and we identified a first (best supported) sister group of the eukaryotes. We then removed from the data all the gene trees supporting this sister group and recovered a new supertree. This identified a second, strongly supported, sister group of the Eukaryota. This process was repeated until all sister groups were identified (see *Materials and Methods* for details).

We found three different signals in eukaryotic genomes, linking them, in order of strength, with Cyanobacteria (the chloroplast endosymbiont), the Proteobacteria

Key words: supertrees, phylogenomics, tree of life, network of life, eukaryote origins.

E-mail: james.o.mcinerney@nuim.ie.

*Mol. Biol. Evol.* 24(8):1752–1760. 2007
doi:10.1093/molbev/msm095
Advance Access publication May 15, 2007

**Table 1**
**The most prominent hypotheses for the origin of Eukaryota. See refs (Martin et al. 2001; Embley and Martin 2006) for more comprehensive lists.**

| Hypothesis | Implied Relationships | Phylogenetic signals expected in genomic analyses |
|---|---|---|
| Tree of life[a] | Archaea and Eukaryota are sister groups. | Eukaryotic genes should show 3 monophyletic domains or Eukaryota Archaebacteria. |
| Eukaryota-first[b] | Eukaryota is the first diverging domain, while Eubacteria and Archaea are sister groups. | Most eukaryotic genes should not have a prokaryotic homologue. Others should show 3 monophyletic domains or Eukaryota with Archaebacteria. |
| Eocyte[c] | Eukaryota is the sister group of Crenarchaeota. | Eukaryotic genes with Crenarchaeota. |
| Phagotrophy[d] | Eukaryota and Archaea are sister groups. This group stemmed from Actinobacteria. | Eukaryotic genes with Archaebacteria, and these two with Actinobacteria. |
| Serial endosymbiosis[e] | Symbiosis of a Thermoplasma-like archaeon and a spirochete (Eubacteria). Mitochondria probably *via* symbiosis with an $\alpha$-proteobacterium. | Eukaryotic genes with Thermoplasma, spirochetes or $\alpha$-Proteobacteria. |
| Syntrophy-1[f] | Eukaryota originated through the symbiosis of a methanogen and a $\delta$-proteobacterium. | Eukaryotic genes with methanogenic Archaea (or within Euryarchaeota), $\delta$- or $\alpha$-Proteobacteria. |
| HydrogenHypothesis[g] | Eukaryota originated through the symbiosis of a methanogen and an $\alpha$-proteobacterium (the mitochondrion). | Eukaryotic genes with methanogenic Archaea (or within Euryarchaeota) or $\alpha$-Proteobacteria. |
| Syntrophy-2[h] | Eukaryota originated through the symbiosis of a sulfur-methabolising Thermoplasmatales-like euryarchaeote and an $\alpha$-proteobacterium (the mitochondrion). | Eukaryotic genes with Thermoplasmatales (or within Euryarchaeota) or $\alpha$-Proteobacteria. |
| Ring of life[i] | Eukaryota originated through the symbiosis of a Crenarchaeota and an $\alpha$-proteobacterium. | Eukaryotic genes with Crenarchaeota or $\alpha$-Proteobacteria. |

[a] e.g. (Woese and Fox 1977; Ciccarelli et al. 2006; Pace 2006)
[b] e.g. (Kurland, Collins, and Penny 2006)
[c] (Rivera and Lake 1992)
[d] (Cavalier-Smith 2002)
[e] (Margulis et al. 2006)
[f] (Lopez-Garcia and Moreira 2006)
[g] (Martin and Muller 1998)
[h] (Searcy and Hixon 1991)
[i] (Rivera and Lake 2004).

($\alpha$-Proteobacteria; the mitochondrial endosymbiont), and the Thermoplasmatales (the archaebacterial host). Assuming no systematic bias, the smaller signals from other eubacterial groups (see Table 2) can be explained by lateral gene transfers into either the bacterial symbionts (Esser, Martin, and Dagan 2006) or the nascent eukaryotic genome. We can thus reject most of the hypotheses in Table 1. Considering that the chloroplast entered the eukaryotic cell at a relatively recent time (Timmis et al. 2004), and that no amitochondriate eukaryote is known (Embley and Martin 2006; de Duve 2007), we conclude that the origin of the eukaryotes is most simply explained by a hydrogen-driven or sulfur-driven syntrophic symbiosis (Searcy and Hixon 1991; Martin and Muller 1998) between a Thermoplasmatales-like euryarchaeote and an $\alpha$-proteobacterium. Irrespective of the metabolic details of this ancient community, eukaryotes are derived from archaebacterial and eubacterial ancestors and are not a primary evolutionary lineage. Archaebacteria and Eubacteria are not natural groups, and a new paradigm is thus needed to replace the "tree of life" (Rivera and Lake 2004; McInerney and Wilkinson 2005; Doolittle and Bapteste 2007; Goldenfeld and Woese 2007).

## Materials and Methods

### Identification of Putative Single-Gene Families

One hundred sixty-eight prokaryotic genomes (including 21 archaeal genomes) and 18 eukaryotic genomes were downloaded from COGENT (http://cgg.ebi.ac.uk/services/cogent/); see Table S1 of the Supplementary Material online for a list. These were assembled in three data sets: the first including 144 eubacterial genomes and 21 archaebacterial genomes (168 in total), the second 8 eukaryotic, 21 archaebacterial, and 97 eubacterial genomes (126 in total), and the third 17 eukaryotic, 21 archaebacterial, and 102 eubacterial genomes (140 in total). The three data sets overlap, rather than being subsets of each other, providing a means (akin to jackknifing) by which to evaluate the stability of our results, and to investigate potential taxon-sampling related biases.

For all three data sets, all-versus-all BLAST searches were carried out implementing the same strategy used by Creevey et al. (2004) and Fitzpatrick et al. (2006) to identify clusters of homologous sequences (see Supplementary Material online). Clusters of putative orthologs were generated by selecting, from the set of gene families defined using the all-versus-all BLAST strategy described in the Supplementary Material online, only the single gene families—i.e., those with only a single sequence from any genome. Multigene families were not considered for this study. We made no attempt to exclude gene families including xenologs, because even universally distributed "core" genes, assuming that there is a universal core, can be laterally transferred, including rRNA genes (Charlebois and Doolittle 2004).

Every single gene family (putative orthologous family–see above) that included at least four taxa was aligned using ClustalW (Thompson, Higgins, and Gibson 1994),

**Table 2**
**Eukaryotic (single) gene families with a prokaryotic homologue and the sister group relationships they imply.**

| | 126 species data set | | | 140 species data set | | | Averages | | |
|---|---|---|---|---|---|---|---|---|---|
| Eukaryotic outgroup | Number of genes | Normalised Number of genes | Proportion (%) | Number of genes | Normalised Number of genes | Proportion (%) | Number of genes | Normalised Number of genes | Proportion (%) |
| Cyanobacteria | 81 | 39.32* | 20.98 | 63.5 | 30.82* | 15.30 | 72.25 | 35.07* | 18.03 |
| α-Proteobacteria | 41.5 | 7.89* | 10.75 | 35.5 | 6.33* | 8.55 | 38.5 | 7.08* | 9.61 |
| γ-Proteobacteria | 28.5 | 4.01 | 7.38 | 26 | 3.60 | 6.26 | 27.25 | 3.80 | 6.80 |
| β-Proteobacteria | 7 | 2.46 | 1.81 | 9.5 | 3.53 | 2.28 | 8.25 | 2.98 | 2.05 |
| δ-Proteobacteria | 7 | 5.08 | 1.81 | 6.5 | 4.71 | 1.56 | 6.75 | 4.89 | 1.68 |
| ϵ-Proteobacteria | 2.5 | 4.58 | 0.64 | 2.5 | 4.58 | 0.6 | 2.5 | 4.58 | 0.62 |
| Undetermined Proteobacteria | 18 | 1.05 | 4.66 | 26.5 | 1.52 | 6.38 | 22.25 | 1.28 | 5.55 |
| Archaebacteria | 53.5 | 10.37* | 13.86 | 40 | 7.75* | 9.63 | 46.75 | 9.06* | 11.6 |
| Spirochetes | 4.5 | 4.93 | 1.16 | 5.5 | 6.02 | 1.32 | 5 | 5.47 | 1.24 |
| Actinobacteria | 12 | 4.5 | 3.1 | 14 | 5.25 | 3.37 | 13 | 4.87 | 3.24 |
| Other Eubacteria | 41.5 | 5.42 | 10.75 | 33.5 | 4.09 | 8.07 | 37.5 | 4.73 | 9.36 |
| Unclear Support | 89 | 2.50 | 23.05 | 152 | 3.61 | 36.62 | 120.5 | 3.1 | 30.08 |

Proportions are calculated, for each data set, over the total number of single gene families with a prokaryotic homologue. Raw gene numbers were normalised dividing, for each group, the number of genes that originated from it by the total number of genes in the genomes of the considered group.

* indicates values that are above the third percentile and thus significantly different from the median of the considered sample. The results presented above are based on the optimal ML trees. Using 70% bootstrap consensus trees did not significantly change these results. Exclusion of trees derived from alignments scoring compositional heterogeneous sequences did not change our results significantly. The majority of eukaryotic genes with a prokaryotic homologue were still of: (1) Cyanobacterial, (2) Proteobacterial and (3) Archaebacterial origin. However, interestingly, the exclusion of trees obtained from compositional heterogeneous alignments resulted in a marked decrease of the genes with an Actinobacterial homologue, and in the disappearance of those with Spirochaetes, δ-, and ϵ-proteobacterial homologues, suggesting these groupings were probably artifactual.

curated using Gblocks (Castresana 2000), and screened for the presence of phylogenetic signal using the Permutation Tail Probability (PTP) test (Archie 1989) as implemented in PAUP4b10 (Swofford 1998). For the ClustalW analyses, the default penalty settings were used, except that we corrected for multiple substitutions when generating the guide trees. A total of 19,898 alignments were generated over the three data sets. Such a large number of alignments cannot be manually curated, and Gblocks (Castresana 2000) was used to remove highly variable, and so potentially poorly aligned or fast-evolving, positions. For the Gblocks analyses, we set the minimal length of a block to 8 amino acid positions, and the maximum number of allowed contiguous nonconserved amino acid positions to 15. Gapped sites were not systematically removed; rather they were treated as any other site in the alignment. Alignments that after the Gblocks analyses were less than 100 positions long were excluded from any further analysis. All remaining alignments were subjected to a permutation tail probability (PTP) test (performed using PAUP4b10), using 2,000 permutations of one random addition sequence each, with the MulTree option turned off. Only alignments that passed the PTP test ($p \leq 0.05$), and thus have significant clustering signal, were used for the phylogenetic analyses.

For every alignment that passed the PTP test, the best fitting amino acid substitution model was selected using the Akaike Information Criterion (AIC), and the $\chi^2$ test for amino acid composition bias was performed to identify alignments showing compositional heterogeneity. Phylogenetic trees were built for each alignment passing the PTP test (including the compositional heterogeneous ones) using Maximum Likelihood (ML) under the best fitting substitution model. Support for the nodes in these trees were estimated using the bootstrap (100 replicates). Maximum likelihood analyses, model selection analyses and the $\chi^2$ test for amino acid composition bias were performed using MultiPhyl (Keane et al. 2006). The ML gene trees obtained were used as input trees for subsequent supertree analyses. All alignments, phylogenetic trees, and the scripts used to automate these analyses are available from the authors upon request.

Because of the dimensions of our data sets, we could not implement methods to attempt countering Long Branch Attraction (LBA; e.g., Brinkmann and Philippe 1999; Pisani 2004). However, the Gblock analyses should have removed the most variable sites in our alignments, and all analyses were performed using ML under the best fitting substitution model. Analyses performed using ML under the best fitting model should be relatively robust to LBA (assuming the best fitting model is not too different from the correct model; see, for example, Lemmon and Moriarty 2004). These precautions cannot completely prevent the possibility that LBA artifacts may affect some of our results, but they should effectively minimize it.

## Supertree Analyses

Supertree reconstruction is a two-step procedure in which gene trees are combined into a single species tree using one of the available supertree methods (Wilkinson et al. 2005a). For each of our three data sets, we initially considered only gene trees obtained from alignments for which Gblocks removed less than 50% of the sites, leaving a total of 5,741, 2,807, and 2,504 alignments, respectively. To establish whether this 50% cutoff biased our results, we then performed analyses (for the 126 and the 140 species data sets) using trees obtained from alignments for which Gblocks removed up to 90% of the sites.

Supertree analyses were performed using two alternative methods: Matrix Representation with Parsimony (Baum 1992; Ragan 1992), and a Neighbor-Joining based implementation of the Average Consensus (Lapointe and Cucumel 1997) method (NJ-AC). Alternative supertree methods have different properties (Wilkinson et al. 2005a; 2007a). Within the limits of feasibility it is important to compare results obtained using different supertree methods, as this may indicate whether a certain solution is, to some extent, method-dependent. Matrix representation with parsimony (MRP) is a parsimony-based supertree method that uses only topological information to recover the supertree, whereas the NJ-AC procedure is a distance-based supertree method and also uses branch length information to derive supertrees. Estimating support in supertree analyses may be difficult (Wilkinson et al. 2005b). Here we used input tree bootstrapping (Creevey et al. 2004; Burleigh, Driskell, and Sanderson 2006). Input tree bootstrapping proportions are interpreted as standard bootstrap proportions. Low bootstrap supports may indicate the presence in the data of conflicting bona fide signals, such as signals supporting alternative placements of the eukaryotes, rather than lack of signal. This cannot be diagnosed using bootstrapping alone, so we used this method in conjunction with a second approach, introduced here, that we called "phylogenetic signal-stripping" (see below).

For the MRP analyses, matrix representations of the ML-derived single gene trees in each of our three data sets were generated using CLANN (Creevey and McInerney 2005). The matrices were then analyzed using PAUP4b10. For all MRP analyses 1,000 replicates with random addition sequences were performed with the multiple-tree option turned off, and swapping the trees using the Tree Bisection-Reconnection (TBR) algorithm. Trees obtained from these analyses were stored and used to start a second analysis that was performed with the multiple-tree option turned on. For all NJ-AC analyses, an AC distance matrix was generated using CLANN, after which PAUP4b10 was used to build a NJ tree from this matrix. Negative branch lengths were prohibited for the NJ analyses.

For all bootstrap analyses, 100 replicates were performed. For the NJ-AC method, 100 pseudoreplicate AC matrices were generated using CLANN. Bootstrap supertrees were then obtained for each pseudoreplicate AC matrix, using the NJ algorithm as implemented in PAUP4b10 and prohibiting negative branch lengths. For the MRP method, 100 pseudoreplicate MRP matrices were generated using CLANN, and these were then analyzed using parsimony in PAUP4b10. For each bootstrap replicate, 100 heuristic searches were performed with a random addition sequence, and the multiple-tree option turned off. This step was followed by branch swapping of the best trees using the TBR algorithm with the multiple-tree option on. These searches can be extremely long on bootstrapped data sets, so a time limit of 1 h was imposed on each TBR phase. The trees presented here are all 50% majority rule consensus trees with minority components obtained from the bootstrap supertree analyses.

Supertrees for the 168 species data set were built to disentangle the phylogenetic relationships within the prokaryotes. The 126 and 140 species data sets were used to understand the phylogenetic relationships of the Eukaryota. Assuming that there is sufficient phylogenetic signal left in eukaryotic genomes to disentangle their phylogenetic relationships, if eukaryotes are chimeras our data will contain multiple phylogenetic signals, and bootstrap analyses should provide low support for (1) the Archaebacteria-Eubacteria split and (2) the split separating the Eukaryota from their closer sister group. On the contrary, if Eukaryota are not chimeras, or if the contribution of Eubacteria to extant eukaryotic genomes is insignificant, as suggested, for example, by Rivera and Lake (1992), Kurland, Collins, and Penny (2006), and Pace (2006), these analyses should return a relatively well-supported tree.

### Phylogenetic Signal-Stripping.

Phylogenetic data sets typically convey multiple signals (Pisani and Wilkinson 2002), only a subset of which may be represented in the optimal tree(s). If eukaryotes are chimeras, genome-wide data would be expected to convey at least two independent phylogenetic signals (each associated with one of the symbiotic partners). Tree-based methods, including supertrees, visualize only the principal signal in data (Pisani and Wilkinson 2002), so we used the phylogenetic signal-stripping approach to identify subsignals that could provide clues about the nature of the symbiotic partners, if any existed. We first identified the most strongly supported sister group of Eukaryota for the whole data set. We then removed all gene trees displaying this relationship and performed new supertree analyses to identify a second possible sister group of the Eukaryota. This procedure was repeated until all positions of the Eukaryota were identified. Note that if the Eukaryota are not chimeras, supertree-based signal stripping should identify only one well-supported position for the Eukaryota in the tree of life. Trees to be removed for the phylogenetic signal-stripping analyses were identified screening every ML tree derived from a phylogenetically informative data set that included at least one eukaryote, using CLANN.

If Eukaryota are chimeras (see also above), our data will contain multiple phylogenetic signals and bootstrap analyses should provide low support for (1) the Archaebacteria-Eubacteria split, and (2) the split separating the Eukaryota from their multiple sister groups (identified using the phylogenetic signal stripping method detailed above). However, low bootstrap values could also result from the absence of a clear phylogenetic signal in the data. To identify the causes of the low bootstrap values obtained for the Archaebacteria-Eubacteria split and for the split separating Eukaryota from the three sister groups pinpointed by the primary, secondary, and tertiary signals identified by our phylogenetic signal stripping method in the eukaryotic genomes (see *Results and Discussion*), a further set of bootstrap analyses were performed. For these analyses, two new data sets were generated. The first included all the trees that did not include eukaryotes and all and only the eukaryotic trees consistent with the primary signal. The second data set scored all the trees that did not include the eukaryotes, and all and only the eukaryotic trees consistent with the secondary signal. If lack of signal in the data caused the low support observed for the sister group relationships of the eukaryotes associated with the primary and secondary signals in our data, then these bootstrap analyses should also provide low bootstrap support for the same relationships. If the low support was caused by conflicting signals in the data (i.e., the presence of the secondary and tertiary signal in the data set that was used to identify the primary signal and the presence of the tertiary signal in the data set used to identify the secondary signal) as we postulate, then these bootstrap analyses should result in a marked increase in the support observed for the split separating the Eukaryotes from their sister groups, and the Eubacteria from the Archaebacteria. These analyses were performed only for the 126 species data set using the NJ-AC supertree method.

### Yet Another Permutation Test Analysis

To assess the hypothesis that Lateral Gene Transfer (LGT), completely erased the phylogenetic signal in our data sets, a NJ-AC (Lapointe and Cucumel 1997) based implementation of the "Yet Another Permutation Tail Probability" test (YAPTP; Creevey et al. 2004) was used to assess whether agreement across input trees was better than expected by random chance.

For each of our original data sets (the 126, 140, and 168 species data sets), 100 random data sets were generated by permuting the leaves on each input tree. This effectively removed the phylogenetic information that our original data sets could potentially convey, while maintaining other aspects of our sets of ML trees (e.g., the sample of species on each tree, and their imbalance). Neighbor-Joining–Average Consensus supertrees were generated for each original data set and for each of the randomly permuted ones. For each data set, the least-squares score of the optimal NJ-AC supertree was compared with those obtained from the 100 corresponding permuted data sets. The YAPTP *P* value is the probability that the least-squares score of the optimal NJ-AC supertree is lower than that of the NJ-AC supertrees obtained from the randomized data. Rejecting the null hypothesis implies that our sets of ML-derived gene trees are more congruent then equivalent sets of random trees. The NJ-AC supertrees were generated using CLANN, and the least-squares scores of the trees on an AC distance matrix corresponding to the original set of trees have been estimated using PAUP4b10.

### Identification of the Sister Group of the Eukaryota in Single Gene Phylogenies

For each single gene family tree derived from a phylogenetically informative data set that included at least one eukaryote, we identified the sister group of the eukaryotes implied by that tree. This was done screening every tree using CLANN. Trees where Eukaryota were not forming a clan (sensu Wilkinson et al. 2007b) were counted as conveying an "unclear signal" (see Table 2). If only one possible sister group of Eukaryota was implied by a tree, i.e., if orthologs of a given eukaryotic gene were found only in one prokaryotic group, the tree was counted as supporting a sister group relation between Eukaryota and that prokaryotic group. If orthologs of a eukaryotic gene were found in more than one prokaryotic group (say Archaebacteria and Cyanobacteria), and the tree could support only two alternative sister groups of Eukaryota, that tree was taken to support both, and in Table 2 it was counted as providing a support of 0.5 to each group. If a tree supported more than two sister groups, it was counted as providing "unclear support" in Table 2. The only exception was for trees that supported multiple proteobacterial sister groups of Eukaryota. These were counted as supporting a sister group relation with an "undetermined proteobacterium." Finally, trees where Eukaryota was nested among other eubacterial groups (those for which little signal was found in the eukaryotic genomes) were grouped together as providing support to "other Eubacteria." These groups include, for example, Thermatogales, Deinococcus/Thermus, Planctomycetales, Mollicutes, and Bacilli.

Several disambiguation rules were used to root the gene trees wherever possible, allowing us to identify the sister group of Eukaryota supported by a given gene. The first rule used was that if both Eubacteria and Archaebacteria were present in a tree, the branch separating these clans was assumed as the rooting point of the tree. This provided directionality and the possibility to identify a single sister group of Eukaryota in trees with multiple possible prokaryotic sister groups. The second rule used was applied only when Eukaryota were nested within Proteobacteria. As the support for the monophyly of the Proteobacteria is high, if Proteobacteria plus Eukaryota could be defined as a clan, to the exclusion of other prokaryotes, we assumed the root of the tree to be outside this clan. Again, rooting provided the directionality necessary to identify a single sister group even if multiple possible prokaryotic sister groups of the eukaryotes were present in a tree. Finally, if the eukaryotes were bracketed between different members of a single prokaryotic group, the eukaryotic version of the considered gene was assumed to have originated within that clan, even if other prokaryotes were present in the tree. We thus assume the monophyly of derived prokaryotic taxa (Creevey et al. 2004).

Results of this analysis are reported in Table 2 in raw form (i.e., total number of genes that originated from every considered prokaryotic group), in normalized form (i.e., total number of eukaryotic genes that originated from every given prokaryotic group divided by the total number of genes from that group considered in our analyses), and as a proportion of the total number of genes. For the normalized values, we also estimated the median and the third interquartile, identifying groups of genes that are significantly abundant in eukaryote genomes as those that are above the third interquartile. Numbers of single gene families supporting alternative sister groups of eukaryotes were calculated using optimal trees, 70% majority rule consensus trees, and excluding trees derived from alignments including sequences that were found to be compositionally heterogeneous using the $\chi^2$ test (see above).

## Results and Discussion

The supertree-based phylogeny of the prokaryotes in figure 1 (see also Figure S1 in the Supplementary Material online) is based on 5,741 single-copy genes and shows that complete genomes support most traditionally recognized prokaryotic groups. Lateral gene transfer causes different

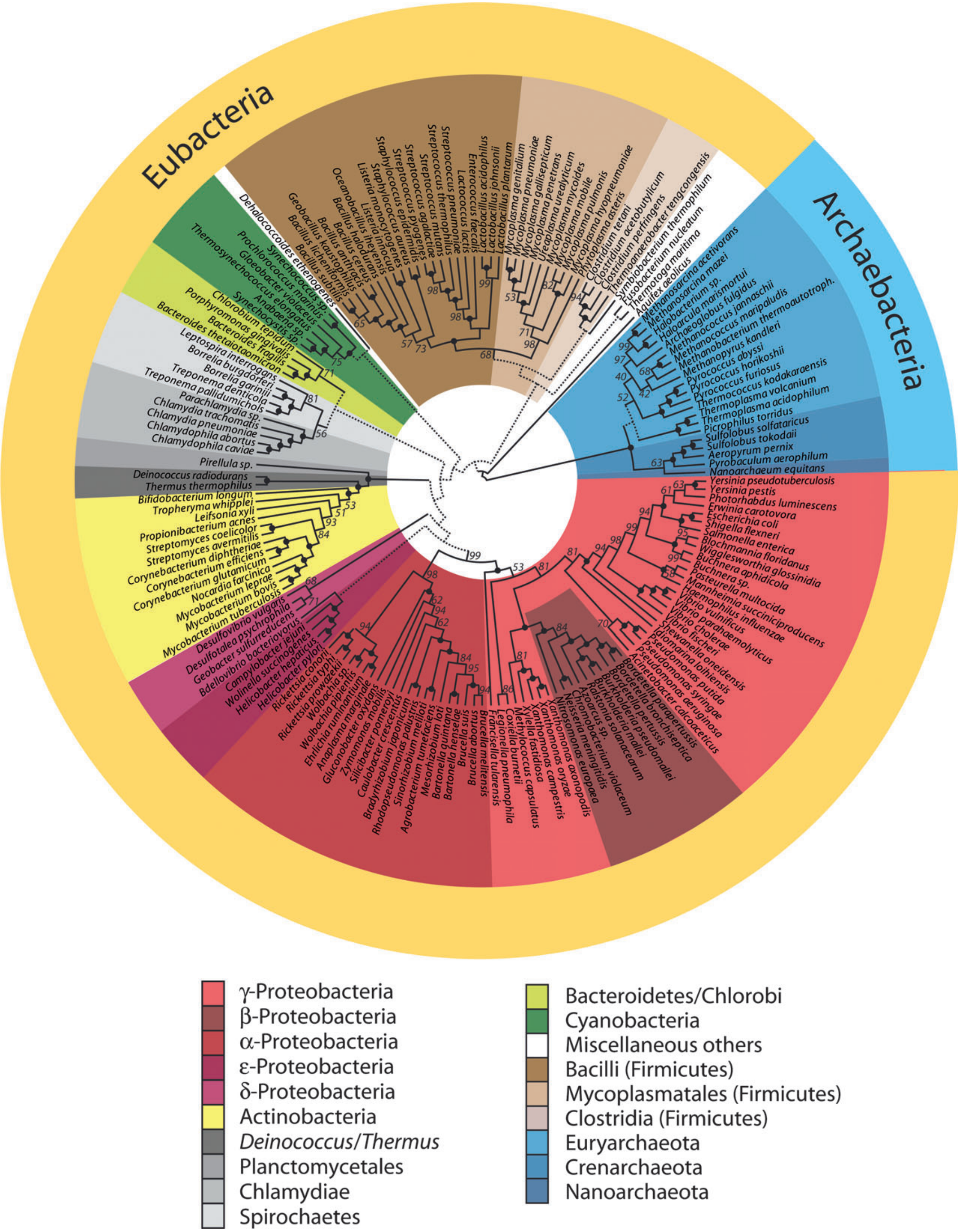

FIG. 1.—A phylogenetic supertree of the prokaryotes based on 168 species and 5,741 genes. Numbers at the nodes represent bootstrap proportions. Full circles indicate nodes with 100% bootstrap support.

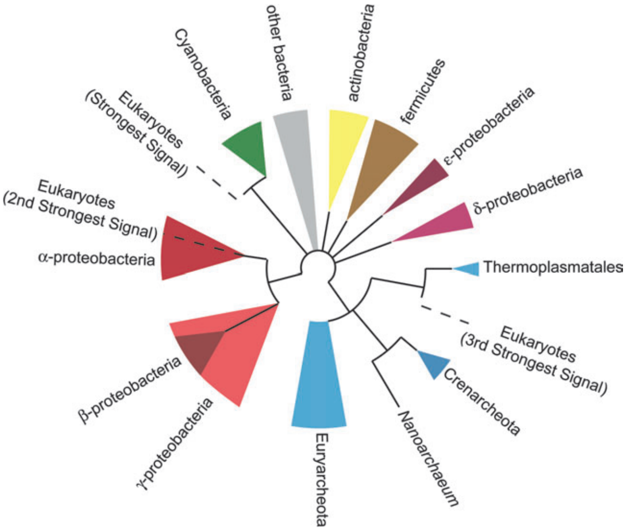

FIG. 2.—Summary of the sister group relationships inferred for the Eukaryota using complete genomes, and the matrix representation with parsimony (MRP) supertree method as the basis of our new "phylogenetic signal-stripping"method (see Supplementary Material online). The strongest signal places Eukaryota as the sister group of Cyanobacteria. The second-strongest signal places Eukaryota within the α-Proteobacteria. The third-strongest signal places them within Archaebacteria (Euryarchaeota), as the sister group of the Thermoplasmatales.

markers to have incompatible evolutionary histories, so assuming that LGT has occurred more-or-less randomly across the tree (but see Gogarten, Doolittle, and Lawrence 2002), the relatively high resolution of the tree in figure 1 suggests that LGT has only partially erased the genealogical signal within prokaryotes. This is confirmed by the YAPTP test $P < 0.01$. The best-supported deep branch in figure 1 separates the Archaebacteria from the Eubacteria, bootstrap proportions on both the MRP and AC supertrees (BP-MRP and BP-AC; see also figure S1 in the Supplementary Materials online) are 100%. Other deep branches are less well supported, consistent with previous reports suggesting that recovering deep phylogenetic events will be difficult (Creevey et al. 2004).

The relationships of the Eukaryota were investigated using two overlapping data sets including 126 species and 140 species (8 and 17 eukaryotes, respectively). These gave very similar results, suggesting that taxon sampling is not influencing our analyses, and trees from the 140 species analyses are not reported here. When all single gene families with less than 50% highly variable sites were used for supertree reconstruction, the eukaryotes clustered within Eubacteria, making the latter paraphyletic (see fig. 2 and figures S2 and S3 in the Supplementary Material online). If the tree in figure S2 of the Supplementary Material online is rooted on the branch separating Archaebacteria and Eubacteria (Gogarten et al. 1989; Iwabe et al. 1989), then the Eukaryota forms the sister group of the Cyanobacteria, and no rooting of this tree can support a partition of life into three domains (Woese and Fox 1977; Woese, Kandler, and Wheelis 1990; Pace 2006). An overwhelming majority of eukaryotic genes with a prokaryotic homolog are eubacterial: approximately 83% (if one exclude the genes providing unclear support—see Table 2), and approximately 25% of these have a cyanobacterial origin. However, genes of cyanobacterial origin are mostly limited to photosynthetic Eukaryota, suggesting that they moved to the nucleus from the chloroplast via endosymbiotic LGT (Timmis et al. 2004). Their high number could reflect either selection or the relative recency of the symbiosis through which the chloroplast entered the eukaryotic cell. Because the monophyly of the Eukaryota is well supported, these genes, despite their limited taxonomic distribution, caused all eukaryotes to cluster as shown in figure 2 (and in figures S2 and S3 of the Supplementary Material online). Support for the branch joining the Eukaryota and the Cyanobacteria is low (BP-MRP = 17%; BP-AC = 11%), consistent with multiple phylogenetic signals in the data and with the concept that only chloroplast-derived genes support this result. Support for the branch separating Eubacteria plus Eukaryota from Archaebacteria (BP-MRP and BP-AC = 66%) is lower than that separating Eubacteria and Archaebacteria in analyses that do not include Eukaryota (fig. 1). This is expected if Eukaryota originated from the symbiosis of an archaebacterium and a eubacterium, as eukaryotic genes will then have different origins and will cluster with either Eubacteria or Archaebacteria.

To visualize the strongest subsignal(s) in our data, we removed those trees supporting a Cyanobacteria plus Eukaryota relationship. This analysis places Eukaryota as the sister group of the α-Proteobacteria (fig. 2, Table 2; figs. S4 and S5 of the Supplementary Material online), consistent with a mitochondrial origin for these genes (Timmis et al. 2004). Support for the Eubacteria plus Eukaryota versus Archaebacteria branch (BP-MRP = 19%; BP-AC = 23%), and for the Eukaryota plus α-Proteobacteria branch (BP-MRP = 28%; BP-AC = 8%) are low, consistent with the presence of further signals in the data. We then also removed trees supporting a sister-group relationship between α-Proteobacteria and Eukaryota. This resulted in a tree showing a basal trichotomy between the Archaebacteria, Eubacteria, and Eukaryota (not shown). Several genes of Eubacterial origin could still be found, but they do not convey a consistent signal (see also Table 2), and they may represent erroneous homology assignments, phylogenetic inaccuracy, or independent LGTs into eukaryotic genomes. To assess the relationship between Archaebacteria and Eukaryota, we removed all these bacterial trees (results reported in the Supplementary Material, figs. S6 and S7), and when trees with up to 90% highly variable sites were included, we could observe strong support for the Eukaryota to be nested within Archaebacteria, as either the sister group of the Thermoplasmatales (BP-MRP = 79%) or within a Thermoplasmatales plus Nanoarchaeota group (BP-AC = 76%; fig. 2; but see also Supplementary Material online, and figures S6 and S7). Including trees with up to 90% variable sites in the analysis did not change our main conclusion that most eukaryotic genes that are not eukaryotic-specific are eubacterial (Table 2).

Removing trees cannot increase the overall amount of phylogenetic signal in a data set, but if there are multiple signals in the data, the nonrandom removal of the trees supporting one of the signals should improve the signal-to-noise ratio in favor of the remaining signals. We therefore used pruned data sets to test whether the low bootstrap support observed for eukaryotic relationships was caused by conflicting signals in the data or by the absence of phylogenetic signal. When all gene trees supporting an archaebacterial or α-proteobacterial sister group of Eukaryota were removed, we observed (with reference to the values reported in fig. S3 of the Supplementary Material online) a 64% increase in the support for the Cyanobacteria-Eukaryota group, and a 15% increase in the support for the Eubacteria-Archaebacteria split, up to 81%. When the gene-trees supporting either an archaebacterial or cyanobacterial sister group of Eukaryota were excluded, we could observe (with reference to the values reported in fig. S5 of the Supplementary Material online) a 48% increase in the support for the α-Proteobacteria-Eukaryota group and a 40% increase in the support for the Eubacteria-Archaebacteria split. This confirms that the low levels of support reported in figs. S3–S5 of the Supplementary Material online are not caused by the lack of phylogenetic signals, they are caused by the presence of three main conflicting signals in the data.

## Conclusions

Previous studies of small and potentially biased gene samples have suggested that eukaryotes and Archaebacteria are sister groups (Ciccarelli et al. 2006), but the largest of these only examined about 1% of an average prokaryotic genome (Dagan and Martin 2006). Other studies have indicated that eukaryotes possess genes of both archaebacterial and eubacterial origin (Rivera and Lake 2004c), but the specific affinities of eukaryotic proteins at the whole genome level have not been addressed for large species samples. Our results show an archaebacterial origin for only ~17% of single copy genes with prokaryotic homologs in photosynthetic eukaryotes, and ~22% in nonphotosynthetic eukaryotes, so the majority of these genes have their origins within the Eubacteria. If a "democratic," genome-wide view is taken, all tree-based explanations for the origin of eukaryotes (Woese and Fox 1977; Rivera and Lake 1992; Cavalier-Smith 2002; Ciccarelli et al. 2006; Kurland, Collins, and Penny 2006; Pace 2006) that have been proposed to date (see Table 1) are thus incorrect.

The simplest scenario to explain our results is that eukaryotes originated through a symbiosis, but not all symbiotic theories (see Table 1) are supported by our data. If one consider the genes in the eukaryotic genomes that originated from a given prokaryotic group (see Table 2) it is evident that only the Cyanobacteria, the α-Proteobacteria, and the Archaebacteria contributed significantly to the composition of the eukaryotic genomes. There is little evidence of eukaryotic genes derived from other bacterial groups such as spirochaetes (Margulis et al. 2006) or δ-Proteobacteria (Lopez-Garcia and Moreira 2006), and these genes are probably better explained as LGTs that entered the eukaryotic nucleus independently, as genes that have previously been transferred to one of the symbiotic partners (Esser, Martin, and Dagan 2006), as phylogenetic artifacts, or as noise (see also Table 2).

The results of our systematic analysis of the phylogenetic signals of eukaryote genes confirm the mosaic nature of the eukaryotic nuclear genomes. There are thus only two primary lineages of life: Archaebacteria and Eubacteria [assuming the "traditional" root (Gogarten et al. 1989; Iwabe et al. 1989)], and these lineages are paraphyletic. Given that the plastid entered the photosynthetic eukaryotes relatively recently, our results identify the partners in an ancient symbiosis from which the first eukaryote arose as an α-proteobacterium and a Thermoplasmatales-like archaebacterium. The exact metabolic relationship that drove this symbiosis remains uncertain, as it depends on the metabolism of extinct species. However, given our results, the absence of evidence for the existence of amitochondriate eukaryotes (Embley and Martin 2006; de Duve 2007), and the metabolisms of modern euryarchaeotes, a sulfur-driven (Searcy and Hixon 1991) or hydrogen-driven (Martin and Muller 1998) syntropy is the most likely event leading to the origin of the eukaryotes.

## Supplementary Material

Supplementary materials are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

This work was supported by a Marie Curie Intra European Fellowship and a Michigan Center for Theoretical Physics grant to D.P., an IRCSET EMBARK postdoctoral fellowship to J.A.C., by a Science Foundation Ireland Research Frontiers Programme grant to J.O.McI, and by the Irish Centre for High-End Computing (ICHEC). The authors thank two anonymous reviewers for their challenging but constructive comments.

## Literature Cited

Archie JW. 1989. A randomization test for phylogenetic information in systematic data. Syst Zool. 38:219–225.

Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon. 41:3–10.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci USA. 102: 14332–14337.

Brinkmann H, Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol. 16:817–825.

Burleigh JG, Driskell AC, Sanderson MJ. 2006. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. Syst Biol. 55: 426–440.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Cavalier-Smith T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. Int J Syst Evol Microbiol. 52:297–354.

Charlebois RL, Doolittle WF. 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res. 14:2469–2477.

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

Creevey CJ, Fitzpatrick DA, Philip GK, Kinsella RJ, O'Connell MJ, Pentony MM, Travers SA, Wilkinson M, McInerney JO. 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proc Biol Sci. 271:2551–2558.

Creevey CJ, McInerney JO. 2005. Clann: investigating phylogenetic information through supertree analyses. Bioinformatics. 21:390–392.

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

de Duve C. 2007. The origin of eukaryotes: a reappraisal. Nat Rev Genet. 8:395–403.

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

Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science. 284:2124–2129.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci USA. 104:2043–2049.

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

Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D, Bryant D, Steel MA, Lockhart PJ, Penny D, Martin W. 2004. A genome phylogeny for mitochondria among alpha-Proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol. 21:1643–1660.

Esser C, Martin W, Dagan T. 2006. The origin of mitochondria in light of a fluid prokaryotic chromosome model. Biol Lett. 3:180–184.

Fitzpatrick DA, Logue ME, Stajich JE, Butler G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol. 6:99.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol. 19:2226–2238.

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, et al. 1989. Evolution of the vacuolar $H^+$-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci USA. 86:6661–6665.

Goldenfeld N, Woese C. 2007. Biology's next revolution. Nature. 445:369.

Gribaldo S, Philippe H. 2002. Ancient phylogenetic relationships. Theor Popul Biol. 61:391–408.

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA. 86:9355–9359.

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McInerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol. 6:29.

Kurland CG, Collins LG, Penny D. 2006. Genomics and the irreducible nature of eukaryote cells. Science. 312:1011–1014.

Lake JA. 2007. Disappearing act. Nature. 446:983.

Lapointe FJ, Cucumel G. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. Syst Biol. 46:306–312.

Lemmon AR, Moriarty EC. 2004. The importance of proper model assumption in bayesian phylogenetics. Syst Biol. 53:265–277.

Lopez-Garcia P, Moreira D. 2006. Selective forces for the origin of the eukaryotic nucleus. Bioessays. 28:525–533.

Margulis L, Chapman M, Guerrero R, Hall J. 2006. The last eukaryotic common ancestor (LECA): acquisition of cytoskeletal motility from aerotolerant spirochetes in the Proterozoic Eon. Proc Natl Acad Sci USA. 103:13080–13085.

Martin W, Hoffmeister M, Rotte C, Henze K. 2001. An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. Biol Chem. 382:1521–1539.

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

McInerney JO. 2006. On the desirability of models for inferring genome phylogenies. Trends Microbiol. 14:1–2.

McInerney JO, Wilkinson M. 2005. New methods ring changes for the tree of life. Trends Ecol Evol. 20:105–107.

Pace NR. 2006. Time for a change. Nature. 441:289.

Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Syst Biol. 53:978–989.

Pisani D, Wilkinson M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. Syst Biol. 51:151–155.

Ragan MA. 1992. Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol. 1:53–58.

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

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

———. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science. 257:74–76.

Searcy DG, Hixon WG. 1991. Cytoskeletal origins in sulfur-metabolizing archaebacteria. Biosystems. 25:1–11.

Swofford DL. 1998. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Massachusetts

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

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

Wilkinson M, Cotton JA, Creevey C, Eulenstein O, Harris SR, Lapointe FJ, Levasseur C, McInerney JO, Pisani D, Thorley JL. 2005a. The shape of supertrees to come: tree shape related properties of fourteen supertree methods. Syst Biol. 54: 419–431.

Wilkinson M, Cotton JA, Lapointe FJ, Pisani D. 2007a. Properties of supertree methods in the consensus setting. Syst Biol. 56:330–337.

Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM. 2007b. Of clades and clans: terms for phylogenetic relationships in unrooted trees. Trends Ecol Evol. 22:114–115.

Wilkinson M, Pisani D, Cotton JA, Corfe I. 2005b. Measuring support and finding unsupported relationships in supertrees. Syst Biol. 54:823–831.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA. 74:5088–5090.

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

William Martin, Associate Editor

Accepted May 9, 2007