ARTICLE

doi:10.1038/nature14963

Endosymbiotic origin and differential loss of eukaryotic genes

Chuan Ku[1], Shijulal Nelson-Sathi[1], Mayo Roettger[1], Filipa L. Sousa[1], Peter J. Lockhart[2], David Bryant[3], Einat Hazkani-Covo[4], James O. McInerney[5,6], Giddy Landan[7] & William F. Martin[1,8]

Chloroplasts arose from cyanobacteria, mitochondria arose from proteobacteria. Both organelles have conserved their prokaryotic biochemistry, but their genomes are reduced, and most organelle proteins are encoded in the nucleus. Endosymbiotic theory posits that bacterial genes in eukaryotic genomes entered the eukaryotic lineage via organelle ancestors. It predicts episodic influx of prokaryotic genes into the eukaryotic lineage, with acquisition corresponding to endosymbiotic events. Eukaryotic genome sequences, however, increasingly implicate lateral gene transfer, both from prokaryotes to eukaryotes and among eukaryotes, as a source of gene content variation in eukaryotic genomes, which predicts continuous, lineage-specific acquisition of prokaryotic genes in divergent eukaryotic groups. Here we discriminate between these two alternatives by clustering and phylogenetic analysis of eukaryotic gene families having prokaryotic homologues. Our results indicate (1) that gene transfer from bacteria to eukaryotes is episodic, as revealed by gene distributions, and coincides with major evolutionary transitions at the origin of chloroplasts and mitochondria; (2) that gene inheritance in eukaryotes is vertical, as revealed by extensive topological comparison, sparse gene distributions stemming from differential loss; and (3) that continuous, lineage-specific lateral gene transfer, although it sometimes occurs, does not contribute to long-term gene content evolution in eukaryotic genomes.

In prokaryotes, inheritance involves recombination superimposed upon clonal growth[1] and the mechanisms of recombination are the mechanisms of lateral gene transfer (LGT): transformation, conjugation, transduction, and gene transfer agents[2–4]. These mechanisms operate unidirectionally from donor to recipient and generate pangenomes[5,6]. In eukaryotes, sexual recombination is reciprocal, prokaryotic LGT machineries are lacking, and genetics indicate inheritance to be vertical[7,8]. Well-known exceptions to the vertical pattern of eukaryote evolution occurred at the origin of chloroplasts and mitochondria, where many genes entered the eukaryotic lineage via gene transfer from endosymbionts[9–11]. More controversial, however, are mounting claims for abundant and continuous LGT from prokaryotes to eukaryotes[12–17]. Such claims, if true, predict that cumulative effects of LGT in eukaryote genome evolution should be detectable in genome-wide surveys spanning many lineages. By contrast, endosymbiotic theory predicts that gene acquisitions in eukaryotes should correspond to the origins of chloroplasts and mitochondria[9] and to secondary endosymbiotic events among algae[18,19].

The evidence behind claims for widespread LGT from prokaryotes to eukaryotes, or from eukaryotes to eukaryotes, comes from genome sequences and rests upon observations of unexpected branches in phylogenetic trees[13,16] and patchy gene distributions across lineages[20,21]. Yet the same observations can stem from vertical evolution, with factors that influence phylogenetic inference causing unexpected branching patterns[22–25] and differential loss causing patchy distributions[26,27]. Distinguishing between these alternatives is not simple. Some cases of lineage-specific LGTs to eukaryotes are incontrovertible, in particular bacterial endosymbiont genome insertions into insect chromosomes[28,29] or viral acquisitions in placental evolution[30]. Yet if LGT to eukaryotes is continuously ongoing in evolution, it has to generate cumulative effects. Even if the average LGT frequency per genome is low, perhaps ~0.5% of all genes per genome[20], LGTs will still accumulate over time, like interest on a bank account: acquired genes will be inherited to descendant lineages, which themselves will continue to acquire new genes. The cumulative effect of LGT generates lineages that have increasingly different and continuously diverging collections of genes. This is exactly what is observed in prokaryotes, where known LGT mechanisms operate and pangenomes accrue[5,6]. Here we test the predictions of the competing alternatives to account for prokaryotic genes in eukaryotes—gradual LGT accrual versus episodic gene transfer from organelles—using gene distributions and maximum likelihood trees to uncover cumulative LGT effects.

Gene distributions bear out endosymbiotic theory

We clustered 956,053 protein sequences from 55 eukaryotes from six supergroups[31] and 6,103,025 sequences from prokaryotes (5,793,897 from 1,847 bacteria and 309,128 from 134 archaea) in a two-stage procedure. We first clustered all sequences within each domain (Supplementary Tables 1–5), then merged domain-specific clusters by a reciprocal best-cluster approach, resulting in 2,585 disjunct clusters containing sequences from at least two eukaryotes and at least five prokaryotes. For multidomain proteins, the cluster was assigned according to the most similar domain in the prokaryote–eukaryote comparison, favouring the detection of recent LGTs from prokaryotes, if they are present. The distributions of taxa for the 2,585 eukaryote–prokaryote clusters (EPCs) and for the 26,117 eukaryotic-specific clusters (ESCs) are shown in Fig. 1 and Extended Data Fig. 1a, respectively. The functional categories distributed across EPCs and ESCs are significantly different (Table 1 and Supplementary Table 6), reflecting the prokaryotic origin of core eukaryotic informational and operational genes[32], and the origin of eukaryotic-specific traits that followed the origin of mitochondria[33].

[1]Institute of Molecular Evolution, Heinrich-Heine University, 40225 Düsseldorf, Germany. [2]Institute of Fundamental Sciences, Massey University, Palmerston North 4474, New Zealand. [3]Department of Mathematics and Statistics, University of Otago, Dunedin 9054, New Zealand. [4]Department of Natural and Life Sciences, The Open University of Israel, Ra'anana 43107, Israel. [5]Department of Biology, National University of Ireland, Maynooth, County Kildare, Ireland. [6]Michael Smith Building, The University of Manchester, Oxford Rd, Manchester M13 9PL, UK. [7]Genomic Microbiology Group, Institute of Microbiology, Christian-Albrechts-University of Kiel, 24118 Kiel, Germany. [8]Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, 2780-157 Oeiras, Portugal.

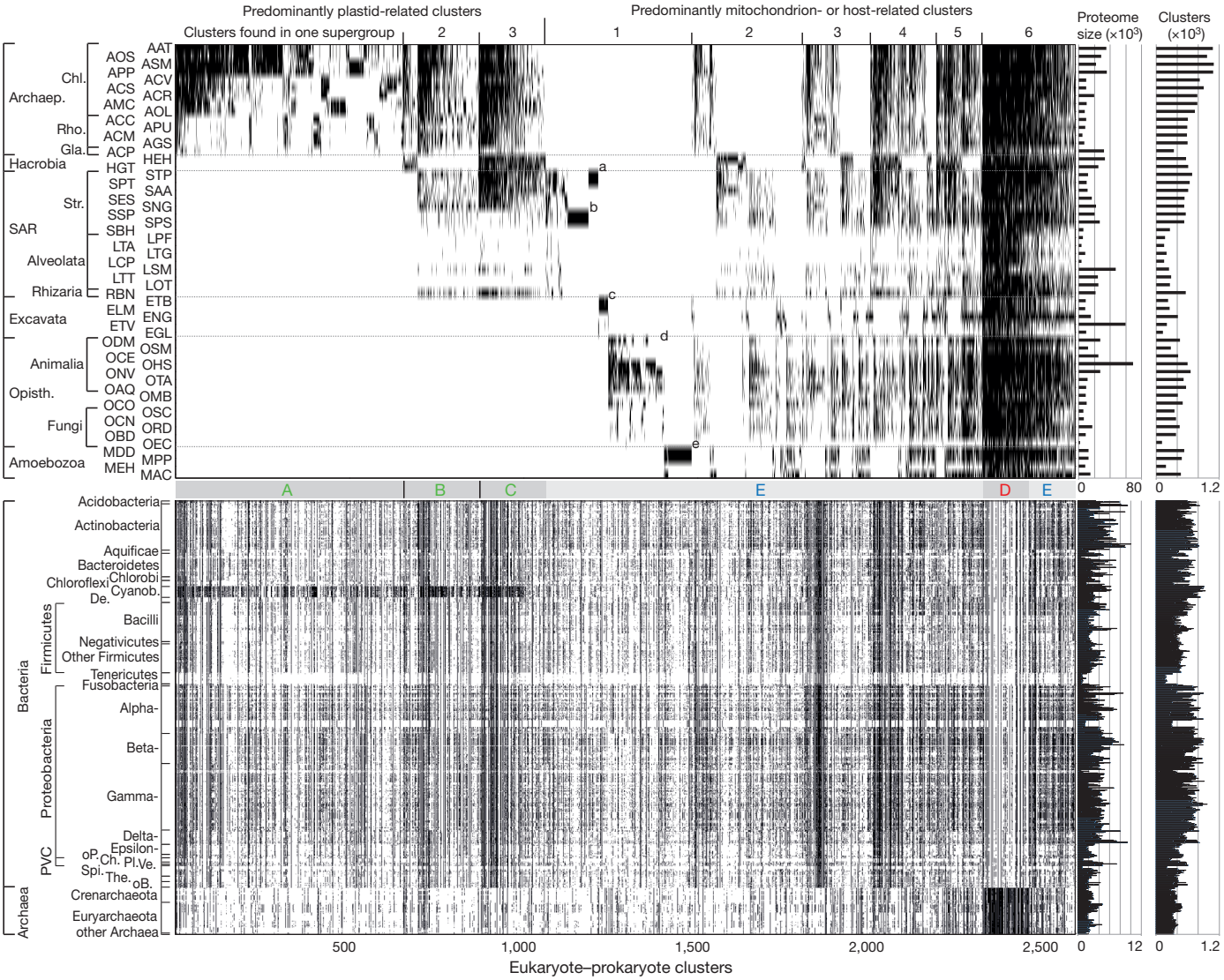


Figure 1 | Distribution of taxa in EPCs. Each black tick indicates gene presence in a taxon. The 2,585 EPCs (*x* axis) are ordered first according to their distribution across six eukaryotic supergroups with clusters specific to lineages with photosynthetic eukaryotes (blocks A–C) on the left, then according to the number of supergroups within which the clusters occur. Clusters most densely distributed in archaea among prokaryotes (block D) and others (block E) are indicated. Lower-case letters label clusters whose distribution is suggestive of recent lineage-specific acquisitions. The numbers of protein sequences and EPCs per genome are shown on the right. Taxon abbreviations are given in Supplementary Tables 1 and 3.

The phyletic distributions of the EPCs reveal blocks of genes with distinctly shared patterns that carry the unmistakable imprint of endosymbiosis in eukaryote evolution. The eukaryotic genes in blocks A–C are present in photosynthetic eukaryotes and related lineages only (Fig. 1), and are densely distributed among one particular group of prokaryotes—the cyanobacteria—as endosymbiotic theory[11] would predict. Block D encompasses genes that were present in the eukaryotic ancestor, that are very densely distributed in archaea, and that are also more refractory to loss than any other group of eukaryotic genes. These correspond to the informational genes[32] representing the archaeal host lineage that acquired the mitochondrion in endosymbiotic theory[34–36]. The archaeal genes in eukaryotes are rarely lost (Fig. 1), being more essential than operational genes[37] and involved in information processing: unlike genes in metabolic pathways, their function cannot be replaced by importing amino acids or vitamins from the environment[29,38]. Block E encompasses many genes that

Table 1 | Functional classification of eukaryotic protein clusters

Functional category	ESCs	EPCs	EPC blocks					
			A	B	C	ABC	D	E
Cellular processes and signalling*	6,685	191	42	14	21	77	14	100
Information storage and processing*	3,940	351	67	28	27	122	75	154
Metabolism*	4,882	1,130	217	95	79	391	35	704
Poorly characterized	10,610	913	328	81	61	470	4	439
Total	26,117	2,585	654	218	188	1,060	128	1,397

The full list of clusters and functional categories is given in Supplementary Table 6. See Extended Data Fig. 10 and Methods for distribution of ESCs and EPCs under different clustering criteria and the tests comparing them.

*χ^2 test of the distribution of clusters across the three general functional categories (null hypothesis was that the distribution is independent of the sets of clusters). The sets of clusters compared (*P* value) were as follows: ESCs/EPCs (0.00), ABC/D (0.00), ABC/E (0.01), D/E (0.00), A/B (0.71), A/C (0.56), B/C (0.29).

were present in the eukaryotic common ancestor, as well as many that are shared across supergroups but are more sparsely distributed than the host-derived genes in block D. These could correspond to the mitochondrion alone[39] or to the mitochondrion plus additional donors that exist in various formulations of endosymbiotic theory[11].

Eukaryote gene distributions and origins

Among the 2,585 trees (Supplementary Table 7) plotted in Fig. 1, 1,933 (74.8%) recovered the eukaryotes as monophyletic and another 329 trees (12%) did not reject eukaryote monophyly in the Kishino–Hasegawa approximately unbiased test (AUT) (Extended Data Fig. 1b). The remaining 323 trees (12%) reject eukaryote monophyly at $P = 0.05$ in the AUT. But these 323 cases are not all necessarily bona fide cases of LGT, because endosymbiosis introduces gene redundancy (for example organelle and cytosolic ribosomes) into the eukaryotic lineage, because many sequencing contaminations are evident in these 323 trees, and because molecular phylogenetics sometimes simply fails[22–25] (Extended Data Figs 2 and 3, Supplementary Table 6 and Methods). Yet even if we assume that these 323 trees represent outright LGTs, the eukaryotes harbouring these genes are not expanding their gene content repertoire via LGT, they are merely re-acquiring members of EPC families already present in the eukaryotic lineage. Rather than dwelling on non-monophyletic exceptions, we investigated the monophyletic majority.

For the 1,933 trees that recovered eukaryote monophyly, we asked which prokaryotic groups were present in the sister group to the eukaryotic clade. Blocks A–C (Fig. 1) encompass 1,060 clusters that clearly correspond to the introduction of photosynthesis into the eukaryotic lineage[18] and its spread via secondary symbiosis[19]. The 188 genes in block C include those acquired during the cyanobacterial origin of plastids and transferred to the nucleus, and then transferred again in at least two independent secondary symbiotic events[18,19] involving the origin of (1) red secondary plastids (*Guillardia*, *Emiliania*, stramenopiles, and alveolates) and (2) green secondary plastids in the *Bigelowiella* lineage. The 218 genes in block B encompass plastid-related functions shared by Archaeplastida and one of the supergroups with secondary plastids.

The distributions of genes depicted in Fig. 1 reflect the endosymbiotic heritage of plastids far more clearly than do the underlying phylogenetic trees (Extended Data Fig. 4). Among the 889 eukaryote monophyly trees in blocks A–C (1,060 clusters), only 283 (31.8%) identified a sister group that contained cyanobacterial sequences only, while 5.9% identified a mixed sister group containing sequences from cyanobacteria and other prokaryotic groups. For the 1,397 genes in block E, 940 trees recovered eukaryote monophyly but only 5.6% identified an alphaproteobacterial sister group to eukaryotes, while 17.2% identified a mixed sister group containing sequences from alphaproteobacteria and other prokaryotic lineages. Did Archaeplastida acquire ~68% of their lineage-specific EPCs from hundreds of independent non-cyanobacterial donors, with similar, more radical implications (~94%) for the more ancient origin of the mitochondrion? That is what the trees imply, while the gene distributions suggest two episodic acquisitions, one endosymbiont donation each at the origin of plastids and mitochondria, respectively. Are the trees to be believed, or are they positively misleading? Within the EPC trees, both the prokaryote subtrees and the eukaryote subtrees address that question.

Organelle ancestors, LGT, and pangenomes

Within the prokaryotic subtrees among 2,585 EPC trees, only five prokaryotic groups were monophyletic in at least 50% of their clusters; they had no more than 15 members each. Eight prokaryotic groups were monophyletic in no more than 20% of their clusters, including alphaproteobacteria (Extended Data Fig. 2c). The extent of prokaryote non-monophyly probably reflects prokaryotic pangenomes and LGT[1–6,40]. Were eukaryotes engaging in pangenomic LGT with prokaryotes, they would have a prokaryote-like pangenome. The 55 eukaryotic genomes sampled identify homologues in only 2,585 prokaryotic clusters. But using the same clustering criteria, 54 strains of *Escherichia coli* identify 5,074 homologous prokaryotic clusters, while samples of 55 genomes from Rhizobiales (alphaproteobacteria) recover on average 8,154 homologous prokaryotic clusters (Extended Data Fig. 2d). That is, a single bacterial species pangenome (conspecific strains of *E. coli*) has sampled prokaryote gene diversity twofold more broadly than 55 eukaryotes have in >1.5 billion years of evolution[41]. Except at organelle origins, eukaryotes are clearly isolated from the pangenome-generating LGT that prokaryotes undertake with each other, an insight that requires simultaneously investigating both phylogenies (Extended data Fig. 2c) and gene distributions (Extended data Fig. 2d).

Prokaryote pangenomes and LGT also affect the inference of gene donors to eukaryotes, because prokaryotic membership in the sister groups to eukaryotes is heterogeneous, often containing representatives from various prokaryotic phyla (Extended Data Fig. 5). Moreover, even in trees where eukaryotes branch with a sister group consisting purely of cyanobacterial, alphaproteobacterial or archaeal sequences, the eukaryotes do not branch with the same cyanobacterial, alphaproteobacterial, or archaeal sister genomes; rather, they branch with homologues from diverse members of these three prokaryotic groups (Extended Data Fig. 6). The prokaryotic homologues of genes that eukaryotes sequestered at organelle origins have been affected by pangenomes and LGT during prokaryotic genome evolution.

This effect is particularly evident in Fig. 2, where for each prokaryotic taxon the frequency of occurrence in the eukaryotic sister group is plotted against the proteome size. Only cyanobacteria, alphaproteobacteria, and, at lower significance levels, two groups of the archaea are implicated as gene donors more often than expected from random distributions of leaves in the individual trees (Supplementary Table 8). The cyanobacterial signal for plastids[11], the alphaproteobacterial

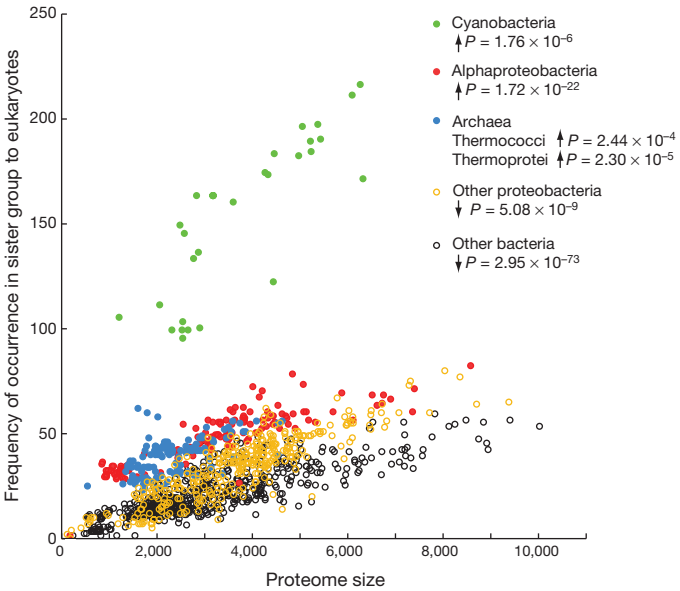


Figure 2 | Occurrence in the sister group versus proteome size. Prokaryotic taxa are plotted according to how frequently they are found in the sister group (defined as the nearest neighbour group) to a monophyletic group of eukaryotes in 1,933 trees against their proteome size. A two-sided Wilcoxon signed-rank test compares these frequencies with those generated by randomly selecting prokaryotic operational taxonomic units (OTUs) into the sister group (100 replicates). Upward and downward arrows indicate higher and lower frequencies in the real data set than in the randomized version, respectively. The test was adjusted for multiple comparisons. For complete statistics, see Supplementary Table 8.

signal for mitochondria[39], and the archaeal signal for the host[34–36] bear out the predictions of endosymbiotic theory. But beyond those three signals, no significant contributions are detected from other prokaryotes that are discussed in various formulations of endosymbiotic theory[14,42,43]. Moreover, individual trees contain information about the provenance of eukaryotic genes that is not better than random: if individual trees linking eukaryotes to prokaryotes are considered outside the context of the full set of trees to which they belong, they can—and do—deliver positively misleading results[44] about the prokaryotic subtree within which eukaryotes branch.

Eukaryote gene evolution is vertical

The eukaryote subtrees address the LGT versus endosymbiosis question even more decisively. There are only two biological mechanisms that could generate the 1,933 cases of eukaryote monophyly for the EPCs. Either the gene was present in the common ancestor of the eukaryotes possessing it and vertically inherited to descendant members[27], or it was acquired by one member of the group and then subsequently distributed via eukaryote-to-eukaryote LGT[21,45]. In the former case, the gene tree of the EPC will tend to be compatible with that observed for ESCs spanning the same taxa, whereas in the latter case the phylogenies will be very different and will differ again for each newly acquired EPC. We tested whether the ESC and EPC trees are drawn from the same distribution by comparing the similarity of trees spanning non-identical leaf sets.

Eukaryote gene evolution is resoundingly vertical (Fig. 3 and Extended Data Fig. 7), with all supergroups, and eukaryotes as a group, passing the test as not significantly different from vertical, while the eukaryote-to-eukaryote LGT alternative—a minimum topology perturbation of one random prune-and-graft LGT per tree—is strongly rejected in all cases. The crucial test case is Archaeplastida, which harbour the most supergroup-specific EPCs (Fig. 1). Although only a minority of Archaeplastida-specific EPCs phylogenetically trace directly to cyanobacteria sampled, they all trace to the Archaeplastida common ancestor (Fig. 3). The data thus indicate that (1) the Archaeplastida-specific EPCs were present in the Archaeplastida common ancestor, (2) their origin thus coincides with the origin of plastids, (3) many are directly involved in photosynthetic functions (Supplementary Table 6), but (4) the sister groups have heterogeneous membership (Extended Data Fig. 6).

This presents two alternatives. If we equate sister-group taxon labels in trees with biological donors, then plastid origin involved hundreds of independent gene donations by hundreds of different donors—the minority of them cyanobacteria—to construct, gene-by-gene, a photosynthetic eukaryote, without any of the individual donations being inactivated through mutation before the plastid was assembled to a functional unit. Alternatively, the gene trees are positively misleading, and these Archaeplastida-specific EPCs were acquired from the ancestor of plastids, which had a fully functional photosynthetic apparatus that merely needed to be integrated into the eukaryotic lineage via recurrent transfer of the necessary genes from the resident organelle to the nucleus[9], clearly the preferable alternative. The untenable proposition of gene-by-gene plastid assembly via hundreds of targeted LGTs arises from interpreting the trees, which can be positively misleading, at face value.

Episodic influx and differential loss

The Archaeplastida case is so important because exactly the same set of observations and the same reasoning applies to the mitochondrion. The host for the origin of plastids was a heterotroph; the transition to autotrophy was driven by endosymbiosis and gene transfer[9,11]. The gene distributions (Fig. 1) reflect that. Similarly, the host for the origin of mitochondria was an archaeon[34–36], the transition to chemiosmotic ATP synthesis in the mitochondrion also resulting from endosymbiosis and gene transfer from the organelle to the host[33]. As with plastids, mitochondria cannot have been constructed via one-by-one LGT, because hundreds of randomly acquired genes to assemble a respiratory organelle cannot be maintained by purifying selection until the mitochondrion is fully functional. Gene transfer from a respiring endosymbiont[9,46] is, by comparison, facile.

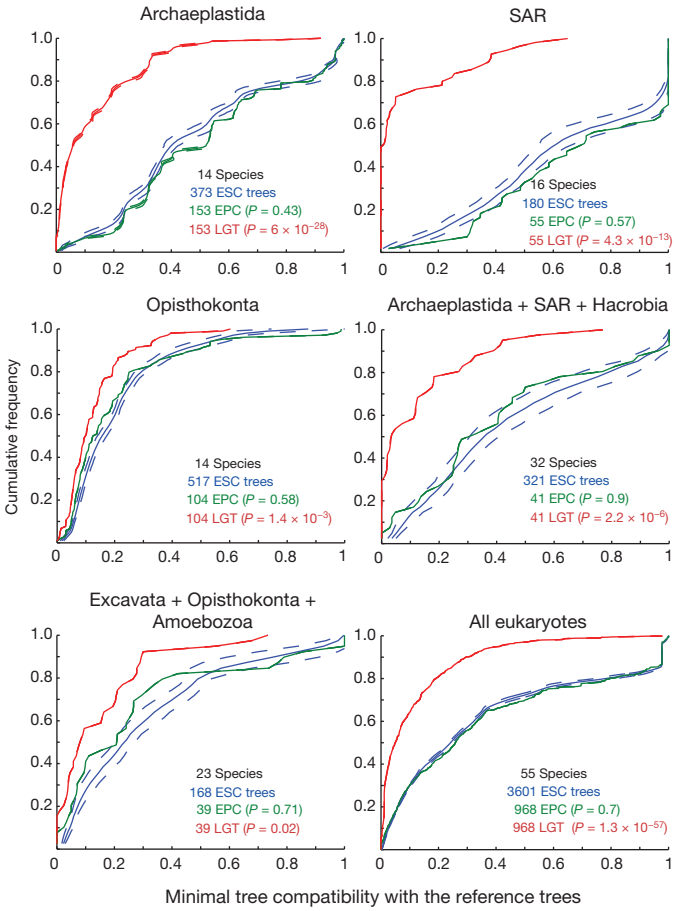


Figure 3 | Comparison of sets of trees for single-copy genes in eukaryotic groups. Cumulative distribution functions (*y* axis) for scores of minimal tree compatibility with the vertical reference data set (*x* axis). Values are number of species, sample sizes, and *P* values of the two-tailed Kolmogorov–Smirnov two-sample goodness-of-fit test in the comparison of the ESC (blue) data sets against the EPC (green) data set and a synthetic data set simulating one LGT (red). Dashed lines delineate the range of distributions in 100 replicates of random down-sampling. See also Extended Data Fig. 7.

Vertical gene inheritance in eukaryotes (Fig. 3 and Extended Data Fig. 7) has a further consequence: the patchy distributions of genes across eukaryotic lineages sampled here are not the result of eukaryote-to-eukaryote LGT, they are the result of differential loss. This is true not only for the EPCs shown in Fig. 1 but also for the ESCs (Extended Data Fig. 1a). Patchy gene distributions in prokaryotes generally indicate LGT, except in isolated species undergoing reductive evolution[38]. In eukaryotes, patchy distributions are often interpreted as evidence for LGT[13], yet the present findings show that patchy distributions in eukaryotes are better explained by differential loss. This leads to steadily declining genome size in terms of numbers of EPCs across eukaryote phylogeny (Extended Data Fig. 8a), with the notable exception of the origin of Archaeplastida, where EPCs double by the influx of ~1,000 clusters. Gene acquisitions in eukaryotes are episodic and correspond to symbioses (Extended Data Fig. 8b).

Finally, some gene distributions among EPCs are highly suggestive of lineage-specific acquisition, because many lineage-specific losses must be assumed. These include 67 dictyostelid-specific genes and 160 opisthokont-specific genes directly observable in Fig. 1, and 210 genes putatively acquired by the ancestor of land plants (Extended Data Fig. 9a). Were these genes recent LGTs, for example during land

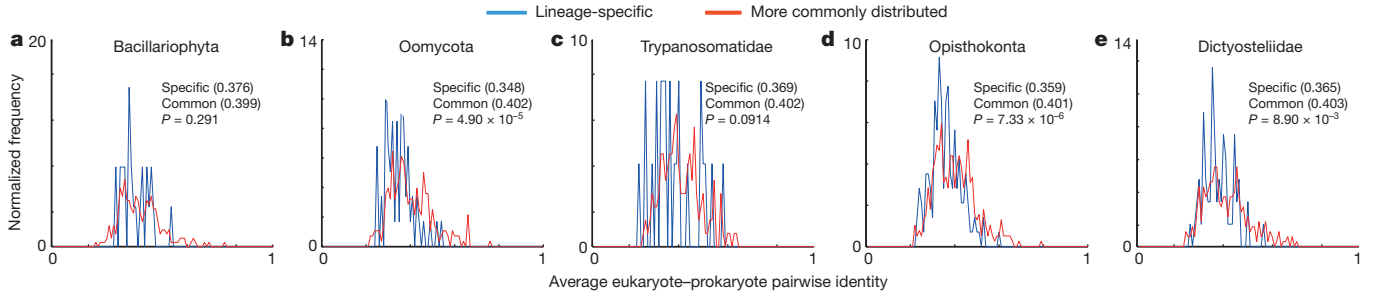


Figure 4 | Eukaryote–prokaryote sequence identities for genes with a tip distribution in eukaryotes versus those whose distributions trace their presence to a more ancient ancestor. **a**–**e**, Genes denoted by lower-case letters in Fig. 1 and those found in at least three of five major supergroups. The mean of the average pairwise identities is shown in parentheses. At $P = 0.05$, a two-sided Wilcoxon rank-sum test either did not reject the null hypotheses that the two sets of genes are not different (**a**, **c**) or suggested the tip-specific eukaryotic genes are less similar to their prokaryotic homologues (**b**, **d**, **e**). See also Extended Data Fig. 9.

plant origin ~450 million years ago[47], they should be more similar to their prokaryotic sisters than genes acquired at plastid and mitochondrial origin. The converse is observed (Fig. 4 and Extended Data Fig. 9). While we do detect genome-specific candidate LGTs (cLGTs), namely eukaryotic singletons that show high similarity to prokaryotic genes, their frequency is approximately four to ten times lower than that of nuclear insertions of mitochondrial and chloroplast DNA[46] (Supplementary Table 9). Thus, even on short timescales, the contribution of gene transfers from organelles is greater than that of cLGTs, whose numbers tend to decrease with updated genome annotations.

Conclusion

Eukaryote gene content evolution resembles the situation in archaea, where gene transfer also has an episodic tendency[48]. Despite many reports of LGT to and among eukaryotes, the combined analyses of all trees that would address the issue reveal no evidence for a detectable cumulative impact of continuous LGT on the evolution of eukaryote gene content. This indicates either (1) that lineage-specific LGTs rapidly undergo loss, having short residence times within their corresponding lineages, (2) that LGT-prone lineages do not give rise to evolutionarily stable descendants, with LGTs being concentrated in evolutionary dead-ends in a kind of terminal differentiation[49], (3) that many suspected LGTs are not really lineage-specific after all and with further eukaryote sampling they will eventually crop up in other distantly related eukaryotes as evidence for differential loss, or (4) any combination thereof. Eukaryotes obtain novel gene families via gene and genome duplication, prokaryotes undergo LGT[50]. Two episodes of gene influx—one from mitochondria and one from chloroplasts, followed by differential loss—account for the phylogeny and distribution of bacterial genes in eukaryotes, which sampled prokaryotic pangenomes at organelle origins.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 April 2015; accepted 20 July 2015.
Published online 19 August 2015.

1. Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* **55,** 709–742 (2001).
2. Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284,** 2124–2128 (1999).
3. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299–304 (2000).
4. Lang, A. S., Zhaxybayeva, O. & Beatty, J. T. Gene transfer agents: phage-like elements of genetic exchange. *Nature Rev. Microbiol.* **10,** 472–482 (2012).
5. Rasko, D. A. *et al.* The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.* **190,** 6881–6893 (2008).
6. Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol. Evol.* **5,** 233–242 (2013).
7. Szathmáry, E. & Maynard Smith, J. The major evolutionary transitions. *Nature* **374,** 227–232 (1995).
8. Nei, M. *Mutation-Driven Evolution* (Oxford Univ. Press, 2013).
9. Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Rev. Genet.* **5,** 123–135 (2004).
10. Lane, C. E. & Archibald, J. M. The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends Ecol. Evol.* **23,** 268–275 (2008).
11. Archibald, J. M. *One plus One Equals One: Symbiosis and the Evolution of Complex Life* (Oxford Univ. Press, 2014).
12. Andersson, J. O. Lateral gene transfer in eukaryotes. *Cell. Mol. Life Sci.* **62,** 1182–1197 (2005).
13. Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nature Rev. Genet.* **9,** 605–618 (2008).
14. Price, D. C. *et al. Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science* **335,** 843–847 (2012).
15. Boto, L. Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proc. R. Soc. B* **281,** 20132450 (2014).
16. Huang, J. L. Horizontal gene transfer in eukaryotes: the weak-link model. *Bioessays* **35,** 868–875 (2013).
17. Crisp, A., Boschetti, C., Perry, M., Tunnacliffe, A. & Micklem, G. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol.* **16,** 50 (2015).
18. Gould, S. B., Waller, R. R. & McFadden, G. I. Plastid evolution. *Annu. Rev. Plant Biol.* **59,** 491–517 (2008).
19. Curtis, B. A. *et al.* Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492,** 59–65 (2012).
20. Alsmark, C. *et al.* Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biol.* **14,** R19 (2013).
21. Keeling, P. J. & Inagaki, Y. A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1α. *Proc. Natl Acad. Sci. USA* **101,** 15380–15385 (2004).
22. Steel, M., Penny, D. & Lockhart, P. J. Confidence in evolutionary trees from biological sequence data. *Nature* **364,** 440–442 (1993).
23. Lockhart, P. J. *et al.* A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol. Biol. Evol.* **15,** 1183–1188 (1998).
24. Guo, Z. H. & Stiller, J. W. Comparative genomics and evolution of proteins associated with RNA polymerase II C-terminal domain. *Mol. Biol. Evol.* **22,** 2166–2178 (2005).
25. Semple, C. & Steel, M. *Phylogenetics* (Oxford Univ. Press, 2003).
26. Hughes, A. L. & Friedman, R. Loss of ancestral genes in the genomic evolution of *Ciona intestinalis*. *Evol. Dev.* **7,** 196–200 (2005).
27. Müller, M. *et al.* Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol. Mol. Biol. Rev.* **76,** 444–495 (2012).
28. Kondo, N., Nikoh, N., Ijichi, N., Shimada, M. & Fukatsu, T. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. *Proc. Natl Acad. Sci. USA* **99,** 14280–14285 (2002).
29. Husnik, F. *et al.* Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* **153,** 1567–1578 (2013).
30. Mi, S. *et al.* Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403,** 785–789 (2000).
31. Derelle, R. *et al.* Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl Acad. Sci. USA* **112,** E693–E699 (2015).
32. Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl Acad. Sci. USA* **95,** 6239–6244 (1998).
33. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467,** 929–934 (2010).
34. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504,** 231–236 (2013).
35. Guy, L., Saw, J. H. & Ettema, T. J. G. The archaeal legacy of eukaryotes: a phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* **6,** a016022 (2014).
36. Koonin, E. V. & Yutin, N. The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. *Cold Spring Harb. Perspect. Biol.* **6,** a016188 (2014).

37. Cotton, J. A. & McInerney, J. O. Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proc. Natl Acad. Sci. USA* **107,** 17252–17255 (2010).
38. Moran, N. A., McCutcheon, J. P. & Nakabachi, A. Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.* **42,** 165–190 (2008).
39. John, P. & Whatley, F. R. *Paracoccus denitrificans* and the evolutionary origin of the mitochondrion. *Nature* **254,** 495–498 (1975).
40. Koonin, E. V. & Wolf, Y. I. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* **36,** 6688–6719 (2008).
41. Parfrey, L. W., Lahr, D. J. G., Knoll, A. H. & Katz, L. A. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl Acad. Sci. USA* **108,** 13624–13629 (2011).
42. Margulis, L., Dolan, M. F. & Guerrero, R. The chimeric eukaryote: origin of the nucleus from the karyomastigont in amitochondriate protists. *Proc. Natl Acad. Sci. USA* **97,** 6954–6959 (2000).
43. Fuerst, J. A. & Sagulenko, E. Keys to eukaryality: Planctomycetes and ancestral evolution of cellular complexity. *Front. Microbiol.* **3,** 167 (2012).
44. Domman, D., Horn, M., Embley, T. M. & Williams, T. A. Plastid establishment did not require a chlamydial partner. *Nature Commun.* **6,** 6421 (2015).
45. Hug, L. A., Stechmann, A. & Roger, A. J. Phylogenetic distributions and histories of proteins involved in anaerobic pyruvate metabolism in eukaryotes. *Mol. Biol. Evol.* **27,** 311–324 (2010).
46. Kleine, T., Maier, U. G. & Leister, D. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annu. Rev. Plant Biol.* **60,** 115–138 (2009).
47. Yue, J. P., Hu, X. Y., Sun, H., Yang, Y. P. & Huang, J. L. Widespread impact of horizontal gene transfer on plant colonization of land. *Nature Commun.* **3,** 1152 (2012).
48. Wolf, Y. I. & Koonin, E. V. Genome reduction as the dominant mode of evolution. *Bioessays* **35,** 829–837 (2013).
49. Hao, W. L. & Golding, G. B. The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res.* **16,** 636–643 (2006).
50. Treangen, T. J. & Rocha, E. P. C. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **7,** e1001284 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the following funding agencies: the European Research Council grants 232975, 666053 (W.F.M.) and 281357 (G.L.; to T. Dagan); the Templeton Foundation grant 48177 (J.O.M.); the Open University of Israel Research Fund (E.H.-C.); the German-Israeli Foundation grant I-1321-203.13/2015 (E.H.-C., W.F.M.), the New Zealand BioProtection CoRE (P.J.L.); the German Academic Exchange Service PhD stipend 57076385 (C.K.); an Alexander von Humboldt Foundation fellowship (D.B.). Computational support of the Zentrum für Informations- und Medientechnologie at the Heinrich-Heine University is acknowledged.

Author Contributions C.K., G.L., S.N.-S., E.H.-C., D.B., M.R., P.J.L., J.O.M., and W.F.M. designed experiments. C.K., G.L., S.N.-S., M.R., F.L.S., and E.H.-C. performed analyses. C.K., S.N.S., F.L.S., P.J.L., D.B., E.H.-C., J.O.M., G.L., and W.F.M. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.F.M. (bill@hhu.de).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Sequence clustering. Protein sequences were downloaded from the NCBI database (version June 2012) for complete prokaryotic genomes and from respective genome sequencing websites for a phylogenetically diverse range of eukaryotes (Supplementary Table 1). Eukaryotic, bacterial, and archaeal protein sequences were clustered separately before homologous clusters from eukaryotes and prokaryotes were combined. The bacterial sequences (Supplementary Table 3) and the archaeal sequences (Supplementary Table 4) were clustered using the methods described[51] ('needle' global identity ≥25%). Eukaryotic sequences were clustered with the reciprocal best BLAST[52] (version 2.2.28; cut-off: expect (*E*) value $\leq 1 \times 10^{-10}$) hit (rBBH) procedure[53] followed by calculation of pairwise global identity (cut-off: global identity ≥40%) of each rBBH pair using the program 'needle' in the EMBOSS package[54] and MCL clustering[55] on the basis of the global identities. Because the prokaryotic genome sample is biased towards bacteria and because many bacterial species are represented by multiple strains (up to 54 for *E. coli*), before clustering, genome sequences from bacterial strains were combined into species pangenomes (Supplementary Table 3) and the rBBH procedure for bacteria (cut-off: *E* value $\leq 1 \times 10^{-10}$ and local identity ≥30%) was performed at the species level to take overrepresentation of bacteria and heavily sequenced bacterial species into account. To avoid combining clusters with different homologous protein domains due to gene fusion or recombination[56], a reciprocal best cluster procedure was used to compare and combine eukaryotic with prokaryotic clusters. Reciprocal all-against-all BLAST searches (cut-off: *E* value $\leq 1 \times 10^{-10}$ and local identity ≥30%) were conducted between 136,661 sequences in all 28,702 eukaryotic clusters containing sequences from at least two eukaryote genomes each, and 4,154,013 sequences in 102,089 bacterial clusters as well as 232,046 sequences in 11,992 archaeal clusters. Prokaryotic clusters containing sequences from not more than four taxa (Supplementary Table 1) were excluded. If ≥50% of the sequences of a eukaryotic cluster had their best hit in a bacterial or archaeal cluster, they were designated the best bacterial or archaeal cluster of the eukaryotic cluster, and vice versa. When a eukaryote cluster and a prokaryote cluster were reciprocally the best clusters for each other, the prokaryotic cluster was combined with the eukaryotic cluster, resulting in an EPC. In total, 2,585 EPCs containing one eukaryote cluster and one bacterial, one archaeal, or two prokaryotic clusters were obtained; the 26,117 remaining eukaryotic clusters were designated ESCs.

Different sets of EPCs and ESCs were generated with lowered thresholds for identifying the best cluster, including changing the BLAST local identity cut-off from 30% to 20% and the minimum proportion of sequences having the best hit in a cluster (best-hit correspondence) from 50% to 40%, 30%, 20% and 10%. Lowering the best-hit correspondence threshold to ≤50% can generate more than one 'best' cluster. To avoid combining two 'best' clusters corresponding to different domains of the sequences in the query cluster into one EPC, we adhered to the >50% threshold. Lowering the local identity or best-hit correspondence thresholds converts some ESCs to EPCs, but the distribution of clusters across eukaryotic taxa is not changed (Extended Data Fig. 10) and the distribution of the functional categories of the genes remains significantly different between ESCs and EPCs (Table 1; $P = 0.00$ for all thresholds in a χ^2 test). Different EPC sets generated with different thresholds are samples from the same pool of eukaryotic genes derived from prokaryotes; sampling lower thresholds for sequence conservation increases the proportion of poorly conserved genes in the alignment and phylogeny steps.

Functional annotation and test of independence. All eukaryotic protein sequences from the 28,702 clusters were BLASTed (cut-off: *E* value $\leq 1 \times 10^{-10}$ and local sequence identity ≥50%) against the eggNOG version 4.0 (ref. 57) database, and the eggNOG/cluster of orthologous groups (COG) identifier of the best hit was assigned to each sequence. A particular eggNOG/COG identifier was assigned to a cluster if it was assigned to more sequences in that cluster than any other identifier. Ties were broken by taking the first listed identifier. Each identifier was then mapped to the COG functional categories[58]. If an identifier was mapped to two or more categories, the category R (general function prediction only) was assigned. Functional annotations are in Supplementary Table 6.

If two sets of eukaryotic genes originated from different prokaryotic sources, the distribution of the functional categories should reflect that of the sources and could be significantly different. To test this, the COG functional categories were divided into four major categories: cellular processes and signalling, information storage and processing, metabolism, and poorly categorized proteins (including those clusters not assigned any eggNOG/COG identifier). A χ^2 test of independence (Table 1) was then used to compare the distribution of genes in the three former categories between ESCs and EPCs (on the basis of different thresholds for combining eukaryote and prokaryote clusters) and between the different blocks of EPCs (Fig. 1) that mainly corresponded to different sources (ABC, D, E) or the same one (A, B, and C).

Relationships between subgroupings within eukaryotes, archaea, and bacteria. A backbone tree of eukaryotes was constructed on the basis of recently published phylogenies[31,59–68]. The archaeal tree was based on the 70 single-copy genes present in the archaeal clusters and was generated in a previous study[51]. Since there was no single-copy orthologue present in every bacterial taxon, 32 nearly universal (present in at least 1,780 out of the 1,847 genomes) single-copy genes were used for inference of a bacterial reference tree (Supplementary Table 3). The OTU for the tree was species (see above). When a species pangenome had multiple sequences (in most cases, each from a different strain of the species) in a cluster, the first in the sorted list of the NCBI GI numbers was used as the representative sequence for this species. The sequences from each gene were aligned separately using MAFFT version 7.130 (ref. 69) with the option 'linsi' and concatenated into a single alignment. A maximum likelihood tree was reconstructed using RAxML version 7.8.6 (ref. 70) under the PROTCATWAG model. An initial tree revealed that some species had much longer branches. A second RAxML run was conducted without four long-branch taxa ('*Candidatus* Tremblaya princeps', '*Candidatus* Hodgkinia cicadicola', '*Candidatus* Zinderia insecticola', and '*Candidatus* Carsonella ruddi'). The reference tree generated was used to modify the taxonomic assignment of some taxa. For example, according to NCBI Taxonomy, *Erysipelothrix rhusiopathiae* strain Fujisawa is placed under Firmicutes in its own class, but the reference tree shows that it is nested within the clade formed by Tenericutes, so it should be placed under this phylum (as is also suggested by a recent study[71]). The curated taxonomic information for bacteria can be found in Supplementary Table 3.

Alignment, phylogenetic analyses, and test for eukaryote monophyly. Sequences in each of the 2,585 EPCs were aligned using MAFFT version 7.130 (ref. 69) with the option 'linsi'. The quality of alignment was compared between different sets of clusters using the HoT method[72,73] with the programs COS_v2.05.pl (in combination with MAFFT 7.130) and msa_set_score_v2.02. Maximum likelihood trees were reconstructed using RAxML version 7.8.6 (ref. 70) under the PROTCATWAG model, with special amino-acid characters U and J converted to X (unknown). The trees (Supplementary Table 7) were analysed using custom Perl scripts to determine whether the eukaryotic sequences formed a clade (Supplementary Table 6); if they did, the prokaryotic clade with the smaller average distance to the eukaryotic clade was identified as the sister group. This criterion is favoured over the use of the number of taxa in the neighbouring groups because the different prokaryotic higher-level taxonomic groups vary greatly in the number of species and genomes sampled (Supplementary Tables 3 and 4).

In cases where the eukaryotic sequences did not form a clade, we conducted the AUT implemented in the CONSEL package[74] to determine whether the apparent non-monophyly was statistically significant. From the maximum likelihood tree of each of the 652 EPCs where eukaryotes were recovered as non-monophyletic, we extracted a eukaryotic subtree by pruning the prokaryotic sequences and a prokaryotic subtree by pruning the eukaryotic sequences. We then generated the set of all trees formed by re-grafting the subtree with eukaryotic sequences into the subtree of prokaryotic sequences, keeping those closest to the original maximum likelihood tree in terms of Robinson and Foulds[75] distance (as computed by the program treedist of the PHYLIP package[76] version 3.695). For all these candidate trees, PhyML version 3.1 (ref. 77) was used to optimize parameters and calculate per-site likelihoods, using option –print_site_lnl, the WAG[78] evolution model, 25 evolutionary rate categories, estimation of gamma distribution shape parameter alpha, and by providing the alternative tree(s) as user tree. Note that only branch lengths and rate parameters, but not topology, were optimized using the -o lr option.

The program makermt in CONSEL version 1.16 was used with –phyml option and a file containing the site-likelihoods for the original tree together with those for the alternatives as input. The output file of makermt was provided to CONSEL version 1.20 and the program catpv was used to extract *P* values for the tree set.

If at least one of the alternative trees lay in the confidence interval of the original tree, namely in cases where the *P* value of the AUT from the multiple bootstrap (au) was not significant at the 5% level, the alternative tree with monophyletic eukaryotic sequences was considered to be equally likely (that is, not significantly worse than the original tree) and monophyly was not rejected (Extended Data Fig. 1b and Supplementary Table 6).

Classification of eukaryote non-monophyly trees. The 323 EPCs that failed the AUT for eukaryote monophyly were manually inspected and classified into categories according to the distribution of eukaryotic sequences in the respective phylogenetic trees. The categories were assigned as follows. Eukaryotes appear as one clade with the exception of sequences from at most one (1) or

two (2) eukaryotes as outlier(s). (3) Eukaryotes appear in two separate clades. Archaeplastida/SAR (stramenopiles + alveolates + Rhizaria)/Hacrobia (photosynthetic eukaryotes and their relatives) and the other eukaryotes form two separate clades (4) with the exception of sequences from at most one eukaryotic outlier (5). Cyanobacterial sequences branch within a single clade of Archaeplastida/SAR/Hacrobia (6) with the exception of one (7) or two (8) eukaryotic outlier(s). (9) Trees contain sequences from only two distinct eukaryotes that do not form a clade. (10) Trees where eukaryotic monophyly could be achieved by removing one sequence or one small clade of prokaryotes. (11) Remaining trees with more complex interleaving of prokaryotic and eukaryotic sequences. The frequency of outlier organisms in the trees was recorded (Supplementary Table 6). To investigate the relationship of gene-copy numbers with eukaryotic monophyly within EPCs, the number of EPCs containing more than one sequence per eukaryote was counted. A χ^2 goodness-of-fit test was used to compare different categories of EPCs with the eukaryote monophyletic EPCs; significance values at the 5% level are reported (Supplementary Table 6).

Prokaryotic gene sharing by eukaryotes and prokaryotes. To compare the number of genes shared by eukaryotes and prokaryotes and those by prokaryotic groups and other prokaryotes, we performed the same clustering procedure as used to generate EPCs for the prokaryotic groups shown in Fig. 1. Protein sequences from 55 prokaryote genomes randomly sampled from within a given group were clustered, as were sequences from the prokaryotes excluding the group, using the same criteria as those used to generate EPCs. The clusters from the sample were combined with the other clusters using the reciprocal best cluster procedure. The number of clusters shared between the 55-prokaryote sample and the remaining prokaryotes was counted (Extended Data Fig. 2d). The procedure was repeated for 100 random samples of 55 genomes (or a single sample of 54 *E. coli* genomes in our data set). Relative to eukaryotes, the extent of prokaryote gene sharing is slightly underestimated owing to smaller prokaryote gene pools as a result of removal of the given group.

Randomization test. All prokaryotic higher-level taxa and almost all prokaryotic species sampled occur in the sister group to eukaryotes in at least one tree (Supplementary Table 8); but instead of bona fide direct gene transfers to eukaryotes, this could result from phylogenetic errors and other factors such as LGT among prokaryotes and gene loss[79]. To evaluate whether the number of times a particular group identified as a putative donor lineage was statistically significant, we compared this number with the expected number of donor inferences in randomized versions of the phylogenetic trees. The frequency of occurrence was counted as the number of trees in which any sequence from a species was found in the sister group to eukaryotes (Fig. 2). The counting was performed for the 1,933 eukaryote monophyletic trees and for 1,933 trees with the same OTUs and the sister group of the same size where OTUs were randomly chosen to be in the sister group. The randomization procedure was repeated 100 times and the counts were averaged. A two-sided Wilcoxon signed rank test was performed in MATLAB R2013a (signrank) with the null hypothesis that the frequency of occurrence normalized by the proteome size for taxa from a taxonomic group was not different between the original 1,933 trees and the randomized data set. A procedure for controlling the false discovery rate[80,81] was used to correct for multiple comparisons involving different taxonomic groups.

Comparison of tree sets. Data sets. We considered six species groupings: (1) Archaeplastida; (2) SAR; (3) Opisthokonta; (4) Archaeplastida, SAR, and Hacrobia; (5) Excavata, Opisthokonta, and Amoebozoa; and (6) any eukaryotic group(s). The data set for each grouping consisted of three tree sets: (1) the verticality reference set consisting of the ESC trees, restricted to the species under consideration; (2) the imports set consisting of the EPC trees, restricted to the species under consideration; and (3) a synthetic data set, 'LGT', derived from the EPC set (2) by the introduction of one random LGT event, simulated by a random prune-and-graft topological operation. Only trees with more than three eukaryotic taxa were considered, which were further subject to two inclusion variants: (1) trees where the gene was present as a single-copy gene in each eukaryote, and where the eukaryotic taxa were monophyletic (Fig. 3); and (2) a more inclusive criterion, where intraspecific paralogues (inparalogues) in the EPC/ESC trees were reduced to one before the remaining eukaryote sequences were realigned and trees re-done, EPCs that passed the AUT for eukaryote monophyly (Supplementary Table 6) were included, and species with multiple copies of the gene were allowed (Extended Data Fig. 7). In the last case, multiple-gene-copy taxa were pruned from the tree to avoid paralogy obfuscation. ESC and EPC trees in Newick format for these two inclusion variants can be found in Supplementary Tables 1 and 7.

Congruence tests. The congruence of individual trees or sub-trees with the entire ESC tree set was measured using the minimal compatibility measure[51]. The trees in each set were layered according to the number of taxa, and pooled together using the random down-sampling procedure[51]. We performed 100 replicates of this procedure, and for each set derived the average cumulative distribution function. The fit between the ESC reference set and the EPC imports and LGT set was tested using a two-tailed Kolmogorov–Smirnov two-sample goodness-of-fit test[82], operating on the average cumulative distributions of the minimal compatibility scores.

Code availability. The MATLAB code used to compare tree sets (Fig. 3 and Extended Data Fig. 7) is available in the source data for Fig. 3.

Identities between eukaryote sequences and prokaryote sister-group sequences. Gene families that are specific to a eukaryotic group or where it forms a distinct clade from other eukaryotes in the tree raise the possibility of a recent lineage-specific transfer. If that were the case, such genes (recent set) are expected to exhibit higher similarities to their prokaryote homologues than more ancient acquisitions (ancient set). To test this, we performed two comparisons of eukaryote–prokaryote sequence identities between the two sets of genes. In the first comparison (Fig. 4), the recent set comprised genes specific to a eukaryote lineage. These are marked with lower-case letters in Fig. 1 and include 28 genes present in bacillariophytes in Fig. 4a, 59 genes present in oomycetes in Fig. 4b, 26 genes present in trypanosomatids in Fig. 4c, 160 genes present in opisthokonts Fig. 4d, and 67 genes present in dictyostelids in Fig. 4e. The ancient set consists of genes commonly present in eukaryotes (found in at least three supergroups, excluding Hacrobia, which are too narrowly sampled). Pairwise sequence identities were calculated as the fraction of amino-acid positions identical between two sequences in the EPC alignments using the program protdist of PHYLIP[76]. For the recent set, pairwise identities were calculated for any eukaryote sequence in the respective monophyletic clade of group-specific genes (lower-case letters in Fig. 1) and all prokaryote sequence in the respective sister group. For the ancient set, pairwise identity was calculated between any sequence from the target eukaryote lineage (for example all bacillariophytes in Fig. 4a) and any prokaryote sequence in the sister group to eukaryotes, in trees where all eukaryote sequences were monophyletic.

For the second comparison (Extended Data Fig. 9), we analysed all EPC trees to test the possibility that LGT from prokaryotes occurred continuously throughout eukaryote lineages. Genes were sorted into potentially recent and potentially ancient acquisition bins. Several criteria were applied to determine whether a gene was probably acquired in a eukaryote common ancestor (for example present in Chloroplastida + Rhodophyta) on the basis of gene distribution, as follows. (1) The gene needs to have a high density distribution: present in at least 33% of the species sampled for each descendent lineage. In the example of (Chloroplastida + Rhodophyta), at least three green lineage and two red lineage members should have the gene. (2) All sequences from this lineage form a clade in the tree. (3) The sister group to this clade consists only of prokaryotic sequences. The patterns suggestive of LGT within each supergroup were inferred under these criteria and mapped onto the eukaryote reference tree (Extended Data Fig. 9a). They were separated into two sets based on the age of the last common ancestor of the eukaryote lineage that apparently acquired the gene: if the last common ancestor was younger than 800 million years according to the reference time tree of eukaryotes[41], the apparent LGT belonged to the recent set; if not, it belonged to the ancient set. In total, the numbers of genes included in recent/ancient sets were 417/254 (Archaeplastida), 130/17 (SAR), 48/4 (Excavata), 41/70 (Opisthokonta), and 79/12 (Amoebozoa). If the age of a particular node (for example, the last common ancestor of *Dictyostelium* and *Polysphondylium*) could not be inferred from the reference time tree, its age was inferred on the basis of its position relative to other nodes in reference trees for the individual supergroups (for example, ref. 64). Pairwise identities were calculated between any sequence in the recipient eukaryote lineage and any prokaryote sequence in the sister group.

For both comparisons, all pairwise identities were averaged for each tree. In Fig. 4 and Extended Data Fig. 9b, the frequencies of the average pairwise identities were normalized so that the area under the curve equalled one. A two-sided Wilcoxon rank-sum test (MATLAB: ranksum) was used to compare identities between the two sets of genes.

Reductive genome evolution in eukaryotes. Our results suggest that the vast majority of EPCs originated from only three prokaryotic donors and have been vertically inherited, followed by differential loss. This is indicated by the gene distributions themselves (Fig. 1), the presence of only three significant prokaryotic donors (Fig. 2), verticality of eukaryotic genes (Fig. 3 and Extended Data Fig. 7), lack of evidence for recent acquisitions based on sequence identity (Fig. 4 and Extended Data Fig. 9), and a strong barrier against LGT between prokaryotes and eukaryotes (Extended Data Fig. 2d). Under this premise, eukaryote ancestral genome sizes were reconstructed using a loss-only model[83] by assuming that all genes in blocks D and E and in blocks A–C originated at the root of eukaryotes and the root of Archaeplastida, respectively, and that patchy distributions result from differential loss. Although it is widely accepted that secondary symbioses spread genes from green algae to two eukaryotic lineages via secondary symbiosis, the number and nature of secondary symbioses giving rise to plastids in the

Hacrobia and SAR lineages (blocks B and C in Fig. 1) is still a matter of debate[18,19,67]. Therefore, for Hacrobia and SAR, genes in blocks B and C were not counted as part of the ancestral genome size (Extended Data Fig. 8a).

Symbiosis and gene transfer in eukaryote genome evolution. Prokaryote reference trees were generated. The archaeal reference tree was condensed into a 13-OTU backbone tree, with each OTU representing a major group of archaea. RAxML trees were reconstructed using the same parameters for each individual gene of the 70 single-copy genes used for the backbone tree, with taxa from each archaeal group constrained to be monophyletic. Similarly, individual gene trees were reconstructed for the 32 bacterial genes, with taxa from each of the 23 major groups constrained to be monophyletic. The non-Bacilli and non-Negativicutes Firmicutes, which form a grade instead of a clade, were forced to be monophyletic and collectively denoted 'Clostridia'. To see how well the individual trees supported the reference tree and how their topologies conflicted with each other, each individual tree was compared with the reference tree and each branch on the latter was colour-coded by how often (white: 0%; black: 100%) the proximal node of this branch was recovered. The bacterial tree was arbitrarily rooted with Thermotogae and the archaeal root was put between Euryarchaeota and the other archaea, a position similar to a recently proposed one[84] except that Nanoarchaeota is not regarded as part of Euryarchaeota.

To indicate the distribution of the nearest prokaryotic neighbours of eukaryotic genes (Extended Data Fig. 8b), which according to the present data were mainly acquired in the eukaryote ancestor and the archaeplastidan ancestor, the prokaryote taxa in the sister group to eukaryotes were mapped with lateral edges linking prokaryotic groups to eukaryotic nodes corresponding to endosymbiotic events: the origin of mitochondria, the origin of plastids, and secondary symbioses. To avoid assigning genes to the wrong source, more conservative criteria were adopted. For the plastid origin, a gene needs to be present in at least two Archaeplastida species, the sequences from Archaeplastida need to be monophyletic or, given secondary endosymbiosis, form a clade where Hacrobia or SAR species are nested (that is, neither of the two descendent lineages of the root of this clade consists of purely Hacrobia or SAR), and the sister group to this clade needs to consist of prokaryotes instead of eukaryotes. Any prokaryotic group occurring in the sister group was counted once and a total frequency was calculated for each group across all trees. The lateral edges linking prokaryotic and eukaryotic trees were colour-coded according to the total frequencies. The reference trees used were the eukaryote reference tree and the prokaryotic backbone trees with shadings showing signal incongruence between individual genes used to construct each tree. For red secondary symbiosis, only one event is indicated for simplicity, but the single lateral red edge makes no statement about the number or timing of events that might have occurred in evolution. Similarly, two secondary symbioses involving green plastids have occurred, but plastid-bearing euglenids are not present among the current genome sample.

Recent organelle insertions in eukaryote genomes. Mitochondrial, plastid, and nuclear genomes were downloaded (Supplementary Table 1). Out of 55 genomes, given the available organelle data, we were able to analyse 39 nuclear genomes for the existence of nuclear mitochondrial DNA copies (*numts*) and 24 nuclear genomes for the existence of nuclear plastid DNA copies (*nupts*). Each organelle genome was BLASTed against the corresponding nuclear genome using Blast+[85] with the blastn task, E value $\leq 1 \times 10^{-4}$, and with the dust flag on for masking low-complexity regions. With a combination of in-house Perl scripts and MySQL queries, the BLAST hits were further filtered and counted as described below. To avoid including contaminating organelle DNA sequences in the count, only BLAST hits with a subject (contig) coverage of $<70\%$ were retained. Two different sets of criteria were then applied to produce two sets of BLAST hits: hit identity $\geq 80\%$ and length ≥ 100 base pairs, or hit identity $\geq 95\%$ and length ≥ 50 base pairs. Hits by identical sequences in different positions of the organelle were counted only once. To estimate the minimal number of independent insertion events in each nuclear genome, the following approach was applied. First, when several organelle fragments had hits to the exact same nuclear fragment, one was randomly chosen. Next, if several organelle fragments had hits to overlapping nuclear fragments, the longer one was chosen for further analysis. Finally, closely spaced organelle hits were concatenated if the nuclear distance between them was smaller than 2 kilobases. This is a permissive version of the method described in ref. 86. To get a minimum estimate, we chose here to concatenate any tandem organelle hits and hits on both nuclear strands, irrespective of the positions or order of the query sequences in the organelle genome (Supplementary Table 9).

Candidate LGTs in eukaryote genomes. The number of cLGTs specific to each eukaryote genome was estimated by BLAST[52] version 2.2.26 searches using all prokaryotic protein sequences and the eukaryotic proteins that were not clustered with any protein from another eukaryote (that is, those found neither in ESCs nor in EPCs). The number of protein sequences with at least one prokaryote hit (E value $\leq 1 \times 10^{-5}$, identity $\geq 95\%$) was reported for each eukaryotic genome (Supplementary Table 9).

Eukaryote non-monophyly in phylogenetic trees. In this study we detected 1,933 EPCs that recovered eukaryotic monophyly in maximum likelihood trees in addition to 329 EPCs that did not reject eukaryote monophyly in AUTs (Extended Data Fig. 1b). The remaining 323 EPCs produced maximum likelihood trees in which the eukaryotic sequences neither formed a monophyletic group nor passed the AUT (Extended Data Fig. 1b). It is possible that these 323 trees represent LGTs, but it is also possible that factors pertaining to the inference of phylogenetic trees are responsible for the failure of the eukaryotic sequences to form a monophyletic group. At least three well-known classes of factor can cause a proportion of eukaryote genes to branch in a non-monophyletic manner in molecular phylogenies: biological causes (for example, host and endosymbiont copies of a given gene persist), contamination in genome sequences, and limitations of phylogenetic methods.

First, among the 323 non-monophyly cases, biological causes constitute a significant class. It is uncontested that, during eukaryotic evolution, endosymbiosis brought together at least three different prokaryotic partners, which served as sources of nuclear genes: cyanobacteria, alphaproteobacteria, and archaea (Fig. 2). For essential cellular functions that were common to both endosymbiont and host such as ribosome biogenesis, amino-acid biosynthesis, nucleotide biosynthesis, cofactor biosynthesis, or carbohydrate metabolism, endosymbiosis brings together divergent but often homologous gene copies within the same cell. This occurs both at the origin of mitochondria and at the origin of plastids (including secondary symbiosis). The phenomenon, called functional redundancy through endosymbiosis[87], is reasonably well known. It often happens that both a host copy and an endosymbiont copy persist in a given eukaryotic lineage, ribosomal proteins being one example[88], chloroplast–cytosol isoenzymes being another[87]. Such homologous gene copies, sequence conservation permitting, can come to reside within the same EPC. Within the 323 non-monophyly cases (Supplementary Table 6), 218 genes (67%) are involved in such essential function: 38 genes (trees) are involved in ribosome biogenesis (including 19 ribosomal proteins), 55 in amino-acid metabolism, 27 in carbohydrate metabolism, 23 in nucleotide metabolism, 16 in cofactor metabolism, 33 in energy conservation, 11 in lipid metabolism, and 13 in post-translational modification. In cases of symbiotic redundancy, if copies from more than one symbiotic partner persist in any eukaryotic lineage sampled, eukaryotic sequences will form two or three distinct clades in the trees, if, that is, that phylogeny is reconstructed accurately in that regard. Before it was known how widespread LGT among prokaryotes is, there was an expectation that genes affected by symbiotic redundancy should branch with cyanobacterial and alphaproteobacterial homologues[87], but that expectation turned out to be too optimistic (Fig. 2) and has been revised[79]. Many of the 323 non-monophyly cases will ultimately be attributable to symbiotic redundancy, but it is not our aim to present that interpretation here. In addition to patterns suggesting LGT to eukaryotes, eukaryote non-monophyly patterns suggesting LGT from eukaryotes to prokaryotes were also observed. Many prokaryotes can take up foreign DNA present in the environment[1,3,89]. Among the 323 cases of non-monophyly, 21 trees show prokaryotic sequences nested within a eukaryote clade (Supplementary Table 6).

Second, bacterial contaminations during genome sequencing will generate non-monophyletic trees for eukaryotes (prokaryotic sequences with eukaryotic taxon labels). We took the data from the genomes as it was, without cleaning or purging for possible contaminations, which would have biased our results towards eukaryote monophyly in trees. Probable cases of contaminating DNA could be found in the eukaryote genome sequence data used in this study. In 78 trees, eukaryotes were non-monophyletic owing to the presence of only one or two eukaryotic outlier organisms. A notable source of outliers is the genome sequence of the sea anemone *Nematostella*[90], which was shown to contain sequences from Proteobacteria and Bacteroidetes[91]. In eukaryote non-monophyly EPC trees, putative contaminations in *Nematostella* were often found as the single outlier (7 out of 52, 13%; Supplementary Table 6) or together with an additional outlier (6 out of 28, 21%; Supplementary Table 6), frequently with either Proteobacteria (for example, E6978_B51) or Bacteroidetes (for example, E3129_B78) taxa in its sister group. Further evidence for contaminating DNA in the *Nematostella* genome comes from the observation that over half of the cLGTs in the 55 genomes stem from the *Nematostella* sequences (Supplementary Table 9). Another source of putative prokaryotic contaminations is the sponge *Amphimedon*[92], an organism known to have dense communities of symbiotic prokaryotes, which could be sources of bacterial contaminants as a result of sequence misassembly[93]. In 9 out of 52 (17%) eukaryote non-monophyly EPC trees with a single eukaryotic outlier organism, and in 9 out of 28 (32%) trees with two eukaryotic outlier organisms, *Amphimedon* (Supplementary Table 6) was an outlier. Single *Amphimedon* outliers in the eukaryote non-monophyly EPC trees

tend to be nested within a clade of gammaproteobacterial sequences as a long-branch (for example, E841_B491, E869_B486, E3655_B52). This is suggestive of the fast-evolving characteristic of symbiotic bacteria[94] and explains why, in contrast to *Nematostella*, the cLGT detection approach (BLAST local identity ≥95%) revealed no cLGT in *Amphimedon* (Supplementary Table 9), despite these putative contaminating bacterial sequences revealed by the trees. In addition, 32 eukaryote non-monophyly trees contain only two eukaryotic organisms, with *Amphimedon* and/or *Nematostella* accounting for 50% of those occurrences (Supplementary Table 6). Although putative contaminations are especially abundant in aquatic organisms or organisms with symbiotic prokaryotes, such as the known case of *Hydra* endosymbiotic bacterial contaminants[95], they can also be found in multicellular land organisms, such as mammals[96] or plants[97]. Contaminations need not stem from the DNA sample sequenced, but can also be introduced from vectors during the sequencing process[97]. The same putative contamination can even be present in genome sequences of different eukaryotes through the use of similar sequencing procedures. An example might be the EPC E14272_B12261, where a transposase gene only present in *Oryza* and *Trypanosoma* (both sequenced using the bacterial artificial chromosome) is 100% identical to the *E. coli* homologue. We used the genome data without purging for possible contaminations, which are, however, present in the data.

Third, factors affecting phylogeny can generate eukaryote non-monophyly in trees. Phylogenetic algorithms strive to find the best tree under a given evolutionary model[22,23,34]. If the model is misspecified, the best tree by a likelihood criterion need not be the true tree[25]. In eukaryote evolution, the duplication of genes and whole genomes is a very frequent phenomenon[98]. In duplicated families, functional constraints can change across sequence positions and across subfamilies, leading to covarion/covariotide phenomena (heterogeneity of the substitution process across sites and across the tree), which can generate phylogenetic artefacts, especially when gene duplicates are present[34,99,100]. We counted the number of EPCs in which any eukaryote was represented with more than one sequence. Among the 323 eukaryote non-monophyletic clusters that failed the AUT, such EPCs are overrepresented in comparison with monophyletic clusters (χ^2 goodness-of-fit test, $P = 6.06 \times 10^{-11}$; Supplementary Table 6). A significant, although much higher, P value was obtained for non-monophyletic clusters that passed the AUT ($P = 3.47 \times 10^{-4}$; Supplementary Table 6). Sampling is also an issue for phylogenetic analyses. We found 23 cases where cyanobacterial sequences were nested within the photosynthetic eukaryotes and their relatives (7 additional cases in which an outlier, possible sequencing contamination, appeared in the tree; Supplementary Table 6). Tree E1689_B206_A295 for example, contains 1,746 sequences and fails the AUT for eukaryote monophyly; however, adding merely ten new top BLAST[52] prokaryote hits from the most recent NR database[101] using the *Arabidopsis* sequence as the query (as of 17 April 2015), produces a highest likelihood tree with Archaeplastida monophyly (Extended Data Fig. 3). That taxon sampling affects phylogeny is well-known[102]; it affects all analyses, not just the present one. Another factor is clustering. Clustering and alignment can introduce phylogenetic biases; larger clusters produce eukaryote non-monophyly significantly more often than smaller clusters ($P = 1.45 \times 10^{-61}$) as do trees generated from the least reliable alignments ($P = 2.04 \times 10^{-10}$; Extended Data Fig. 2). The two-step clustering procedure used in this study avoids combining sequences into families that are too large and complex in terms of shared protein domains: the joining of a cluster for protein A to a cluster for protein B via a single AB fusion protein generates extremely large families, sometimes called giant connected components[103]. However, the universal identity threshold across all clusters could result in over-clustering in some cases: grouping of distinct prokaryotic families, each with eukaryotic homologues, into a single cluster with two eukaryotic branches, each monophyletic, but generating eukaryote non-monophyly for the cluster.

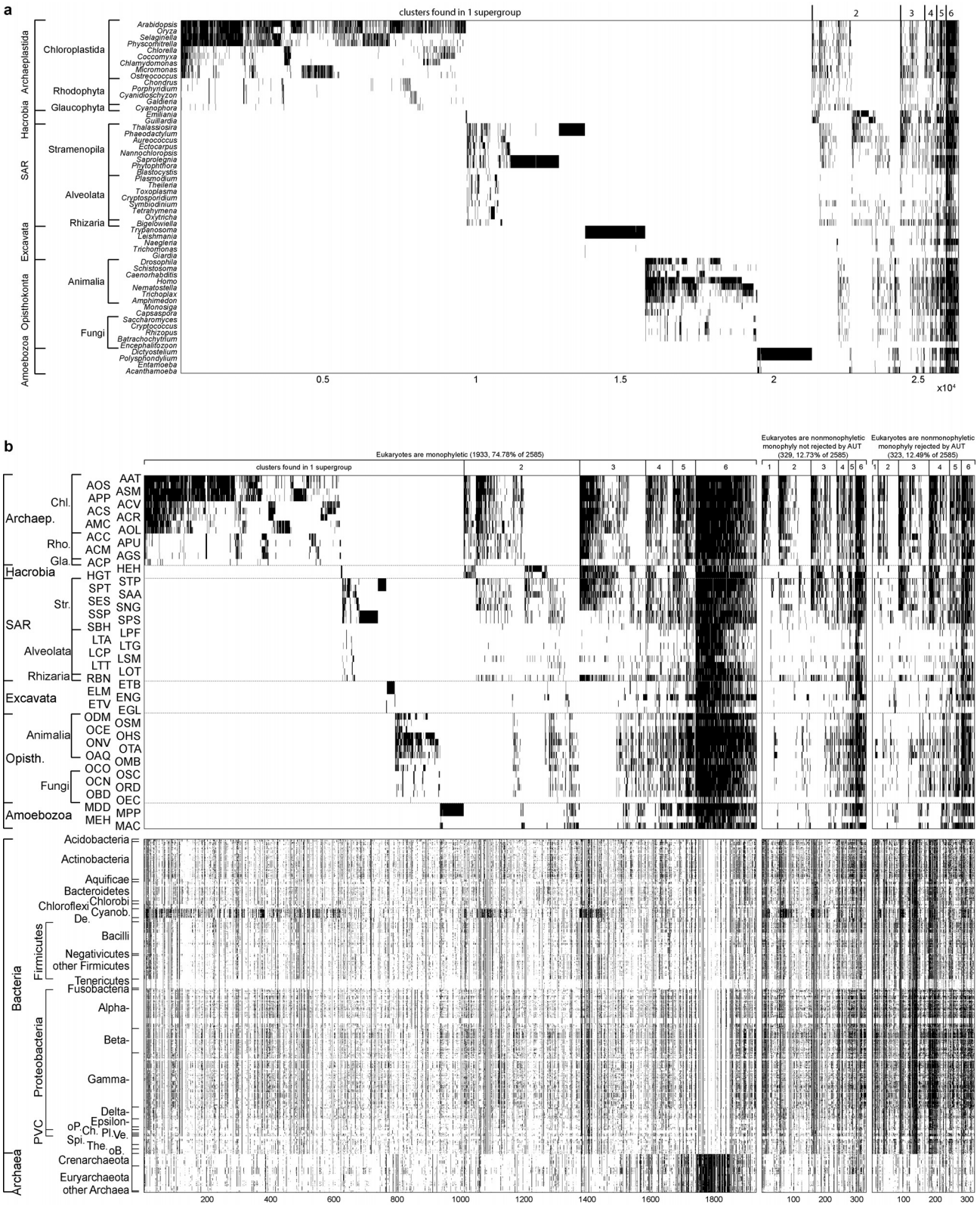
For 134 trees, there was no obvious contamination problem or case of cyanobacteria and plants interleaving. These 134 cases were therefore classified as putative LGT (Supplementary Table 6). But when the 134 cases were compared with the eukaryote monophyletic EPCs, we found significantly more trees than expected with any eukaryote having more than one gene copy (duplicates) ($P = 1.72 \times 10^{-13}$; Supplementary Table 6); in the remaining 189 cases the P value increased to 4×10^{-3}. The presence of an additional, divergently branching copy can result from functional redundancy through endosymbiosis[87] and differential loss, through heterogeneity of the substitution process across sites and across the tree[34,99,100], or through lineage-specific LGT. Of course, many of the trees in question might be affected by more than one of these factors. If LGT is the cause of these 323 cases, which for this paper we conservatively assume, then the eukaryotes in question are still not expanding their gene repertoire, they are merely reacquiring fresh copies of genes already present in the eukaryotic lineage. The details of these 323 trees are in Supplementary Table 6; the trees themselves are in Supplementary Table 7.

Estimating the relative contributions of the host, mitochondria, and plastids to the gene repertoire of present-day eukaryotes. The proportion of genes contributed by the archaeal host is calculated as the proportion of eukaryote monophyly EPC trees where archaea are found in the sister group, including 314 with pure archaeal sister groups and 33 with both archaea and bacteria in the sister group (Extended Data Fig. 5): 347/2,585 = 13.42%. The contribution from the plastid ancestor is calculated by regarding all clusters in the ABC block (Fig. 1) as genes of plastid origin other than those (83) where eukaryotes are monophyletic with archaea in the sister group: (1,060 − 83)/2,585 = 37.79%. The mitochondrion-derived genes are all the other genes: 100% − 13.42% − 37.79% = 48.79%.

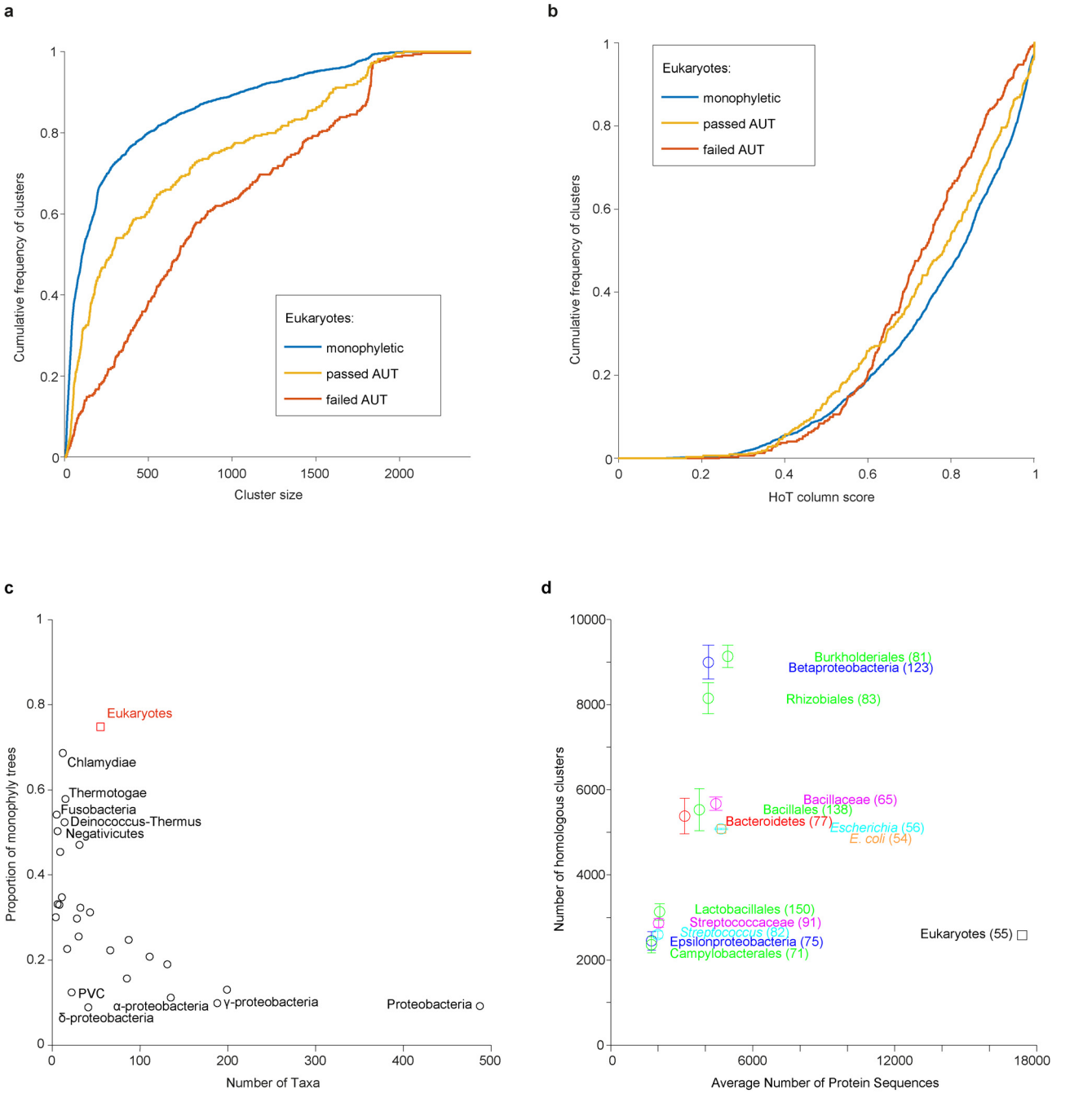
Note that the number for the host contribution is probably an underestimate, as only EPCs with a monophyletic eukaryotic clade in the maximum likelihood tree were counted. For genes of plastid origin, it might be a slight overestimate, since there would also be genes of plastid–host origin that are now specific to Archaeplastida/SAR/Hacrobia and found in the ABC block as the result of differential loss. Another complication is that there can be clusters with genes from more than one source (see above), so there can be, for example, E block clusters of partial plastid and partial mitochondrial origin.

51. Nelson-Sathi, S. *et al.* Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517,** 77–80 (2015).
52. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
53. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
54. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).
55. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
56. Apic, G., Gough, J. & Teichmann, S. A. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310,** 311–325 (2001).
57. Powell, S. *et al.* eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42,** D231–D239 (2014).
58. Tatusov, R. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinform.* **4,** 41 (2003).
59. Yoon, H. S., Muller, K. M., Sheath, R. G., Ott, F. D. & Bhattacharya, D. Defining the major lineages of red algae (Rhodophyta). *J. Phycol.* **42,** 482–492 (2006).
60. James, T. Y. *et al.* Reconstructing the early evolution of fungi using a six-gene phylogeny. *Nature* **443,** 818–822 (2006).
61. Okamoto, N., Chantangsi, C., Horak, A., Leander, B. S. & Keeling, P. J. Molecular phylogeny and description of the novel katablepharid *Roombia truncata* gen. et sp. nov., and establishment of the Hacrobia taxon nov. *PLoS ONE* **4,** e7080 (2009).
62. Hampl, V. *et al.* Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc. Natl Acad. Sci. USA* **106,** 3859–3864 (2009).
63. Janouškovec, J., Horák, A., Oborník, M., Lukeš, J. & Keeling, P. J. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc. Natl Acad. Sci. USA* **107,** 10949–10954 (2010).
64. Lahr, D. J. G., Grant, J., Nguyen, T., Lin, J. H. & Katz, L. A. Comprehensive phylogenetic reconstruction of Amoebozoa based on concatenated analyses of SSU-rDNA and actin genes. *PLoS ONE* **6,** e22780 (2011).
65. Adl, S. M. *et al.* The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* **59,** 429–493 (2012).
66. Leliaert, F. *et al.* Phylogeny and molecular evolution of the green algae. *Crit. Rev. Plant Sci.* **31,** 1–46 (2012).
67. Keeling, P. J. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annu. Rev. Plant Biol.* **64,** 583–607 (2013).
68. Jackson, C. J. & Reyes-Prieto, A. The mitochondrial genomes of the glaucophytes *Gloeochaete wittrockiana* and *Cyanoptyche gloeocystis*: multilocus phylogenetics suggests a monophyletic Archaeplastida. *Genome Biol. Evol.* **6,** 2774–2785 (2014).
69. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
70. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
71. Yutin, N. & Galperin, M. Y. A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. *Environ. Microbiol.* **15,** 2631–2641 (2013).
72. Landan, G. & Graur, D. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.* **24,** 1380–1383 (2007).
73. Landan, G. & Graur, D. Local reliability measures from sets of co-optimal multiple sequence alignments. *Pacif. Symp. Biocomput.* **13,** 15–24 (2008).
74. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17,** 1246–1247 (2001).
75. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53,** 131–147 (1981).
76. Felsenstein, J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266,** 418–427 (1996).

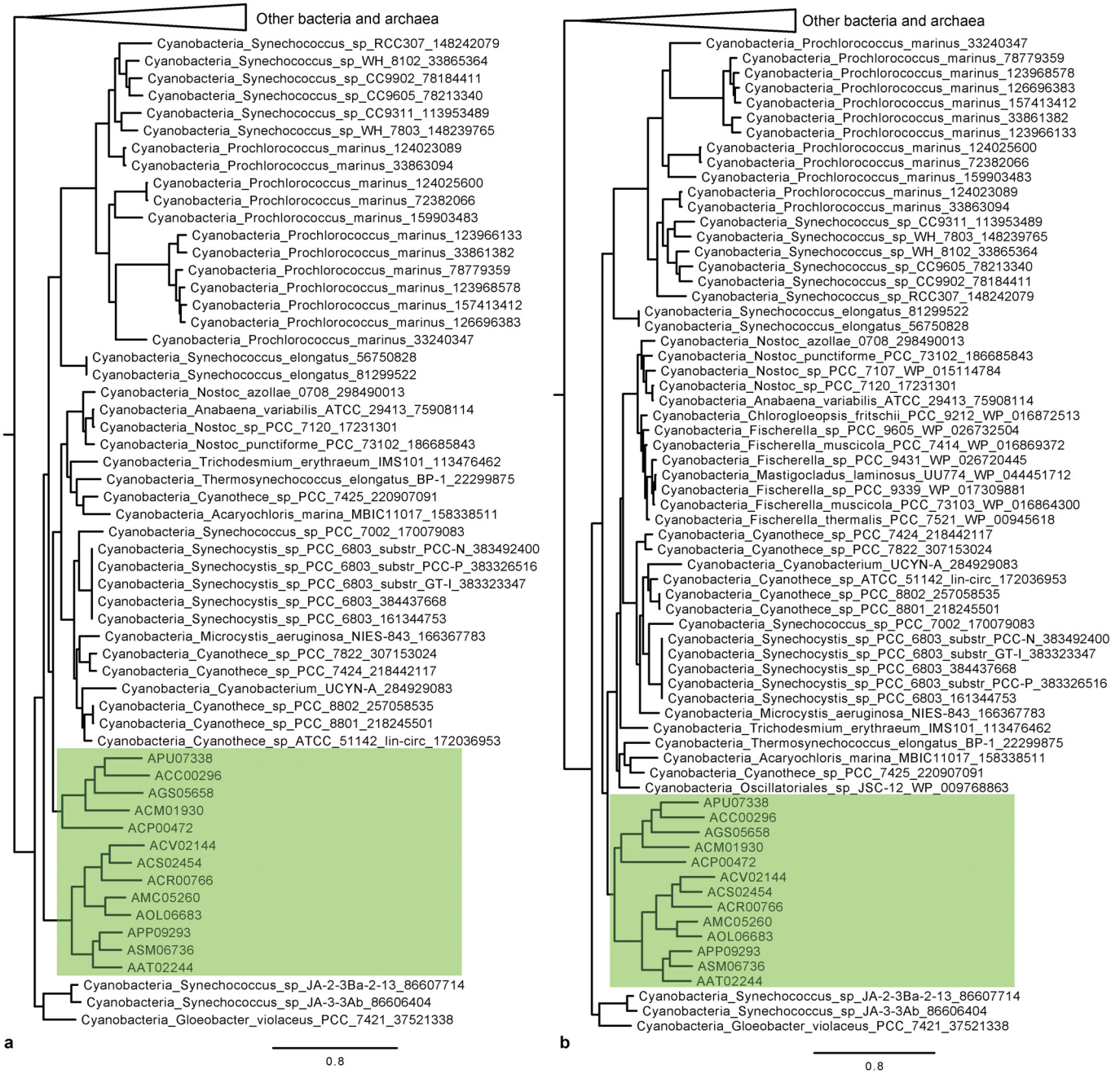
77. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
78. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18,** 691–699 (2001).
79. Ku, C. *et al.* Endosymbiotic gene transfer from prokaryotic pangenomes: inherited chimerism in eukaryotes. *Proc. Natl. Acad. Sci. USA* (2015).
80. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57,** 289–300 (1995).
81. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29,** 1165–1188 (2001).
82. Zar, J. H. *Biostatistical Analysis* Ch. 22 (Pearson, 2014).
83. Dagan, T. & Martin, W. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl Acad. Sci. USA* **104,** 870–875 (2007).
84. Petitjean, C., Deschamps, P., Lopez-Garcia, P. & Moreira, D. Rooting the domain Archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biol. Evol.* **7,** 191–204 (2015).
85. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinform.* **10,** 421 (2009).
86. Hazkani-Covo, E. & Graur, D. A comparative analysis of numt evolution in human and chimpanzee. *Mol. Biol. Evol.* **24,** 13–18 (2007).
87. Martin, W. & Schnarrenberger, C. 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,** 1–18 (1997).
88. Maier, U. G. *et al.* Massively convergent evolution for ribosomal protein gene content in plastid and mitochondrial genomes. *Genome Biol. Evol.* **5,** 2318–2329 (2013).
89. de Vries, J. & Wackernagel, W. Integration of foreign DNA during natural transformation of *Acinetobacter* sp. by homology-facilitated illegitimate recombination. *Proc. Natl Acad. Sci. USA* **99,** 2094–2099 (2002).
90. Putnam, N. H. *et al.* Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317,** 86–94 (2007).
91. Artamonova, I. I. & Mushegian, A. R. Genome sequence analysis indicates that the model eukaryote *Nematostella vectensis* harbors bacterial consorts. *Appl. Environ. Microbiol.* **79,** 6868–6873 (2013).
92. Srivastava, M. *et al.* The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* **466,** 720–726 (2010).
93. Hentschel, U., Piel, J., Degnan, S. M. & Taylor, M. W. Genomic insights into the marine sponge microbiome. *Nature Rev. Microbiol.* **10,** 641–654 (2012).
94. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nature Rev. Microbiol.* **10,** 13–26 (2012).
95. Wenger, Y. & Galliot, B. RNAseq versus genome-predicted transcriptomes: a large population of novel transcripts identified in an Illumina-454 *Hydra* transcriptome. *BMC Genom.* **14,** 204 (2013).
96. Langdon, W. B. Mycoplasma contamination in the 1000 Genomes Project. *BioData Min.* **7,** 3 (2014).
97. Lang, D., Zimmer, A. D., Rensing, S. A. & Reski, R. Exploring plant biodiversity: the *Physcomitrella* genome and beyond. *Trends Plant Sci.* **13,** 542–549 (2008).
98. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA* **102,** 5454–5459 (2005).
99. Lockhart, P. J., Larkum, A. W. D., Steel, M. A., Waddell, P. J. & Penny, D. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc. Natl Acad. Sci. USA* **93,** 1930–1934 (1996).
100. Lockhart, P. J. *et al.* How molecules evolve in eubacteria. *Mol. Biol. Evol.* **17,** 835–838 (2000).
101. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33,** D501–D504 (2005).
102. Zwickl, D. J. & Hillis, D. M. Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* **51,** 588–598 (2002).
103. Alvarez-Ponce, D., Lopez, P., Bapteste, E. & McInerney, J. O. Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc. Natl Acad. Sci. USA* **110,** E1594–E1603 (2013).



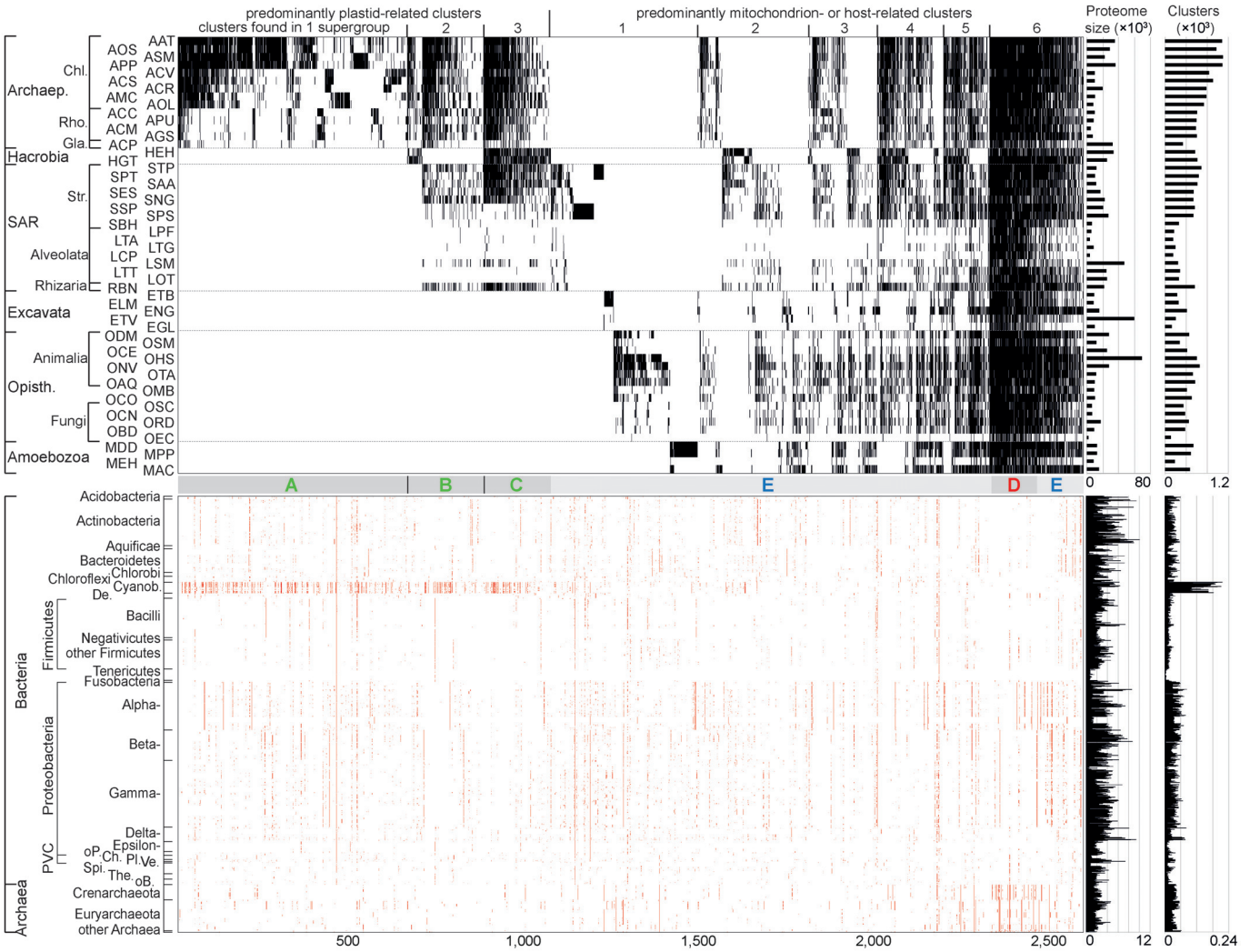
Extended Data Figure 1 | Additional gene distribution patterns.
a, Distribution of ESCs. Each black tick indicates the presence of a cluster in a taxon. The 26,117 ESCs (*x* axis) from 55 eukaryotic genomes (Supplementary Table 1) are sorted according to their distribution across the six eukaryotic supergroups. **b**, Distribution of taxa in EPCs and monophyly of eukaryotes. Each black tick indicates the presence of a cluster in a taxon. The 2,585 EPCs (*x* axis) are separated into three sets according to the monophyly of eukaryotes and the results of the AUT and, within each set, are ordered according to their distribution across the six eukaryotic supergroups. Clusters where eukaryotes were resolved as non-monophyletic in the maximum likelihood tree tend to occur more frequently in bacterial taxa. Archaep., Archaeplastida; Opisth., Opisthokonta; Chl., Chloroplastida; Rho., Rhodophyta; Gla., Glaucophyta; Str., Stramenopila; De., Deinococcus-Thermus; oP., other Proteobacteria; Ch., Chlamydiae; Pl., Planctomycetes; Ve., Verrucomicrobia; Spi., Spirochaetae; The., Thermotogae; oB., other Bacteria. For abbreviations of eukaryotes, see Supplementary Table 1.



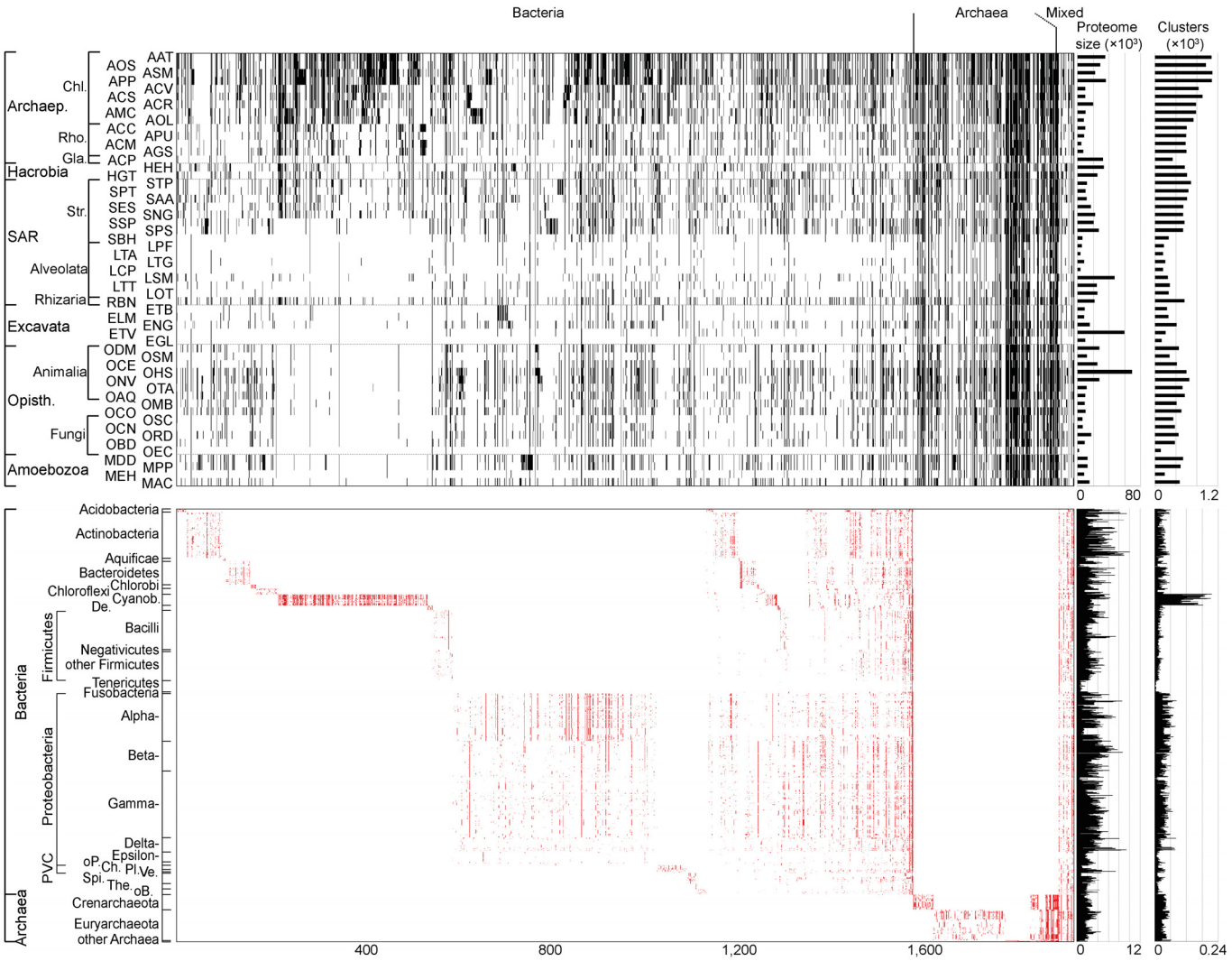
Extended Data Figure 2 | Clustering, monophyly, and gene sharing. **a**, **b**, Monophyly of eukaryotes in maximum likelihood trees, cluster size, and alignment quality. Cumulative frequency of clusters with different cluster size (**a**) or different HoT[72] column scores (**b**) is plotted for three sets of EPCs that differ in terms of the monophyly of eukaryotes in the maximum likelihood trees (monophyletic: resolved as monophyletic in the original tree; passed AUT: resolved as non-monophyletic in the original tree, but at least one alternative tree with eukaryote monophyly (see Methods) was as likely at $P = 0.05$ in an AUT; failed AUT: alternative trees were not as likely as the original tree where eukaryotes were resolved as non-monophyletic). One-sided Kolmogorov–Smirnov two-sample goodness-of-fit test (cluster size/HoT column scores): monophyletic versus passed AUT, $1.04 \times 10^{-13}/7.9 \times 10^{-3}$; monophyletic versus failed AUT, $1.45 \times 10^{-61}/2.04 \times 10^{-10}$; passed AUT versus failed AUT, $3.40 \times 10^{-13}/4.00 \times 10^{-3}$. **c**, **d**, Prokaryotic monophyly and gene sharing. **c**, Proportion of trees showing monophyly for taxonomic group. Prokaryotic phyla and classes (Supplementary Tables 3 and 4) that are monophyletic in the reference trees and that have at least five taxa (genomes in archaea or species in bacteria) are plotted according to the number of taxa and the proportion of EPC trees with at least two sequences from a prokaryotic group where it forms a monophyletic group. The proportion of eukaryote monophyly trees is higher than that of any prokaryotic group, including those with many fewer taxa. **d**, Gene sharing between a prokaryotic group and other prokaryotes. Using the same procedure for the generation of EPCs, 55 genomes were randomly sampled from a group of bacteria and the number of clusters (EPCs) they shared with prokaryotes not from this group was counted. The average number of shared clusters was mapped for each taxonomic group with 55–150 genomes (error bar, s.d.; number of genomes in parentheses). For *E. coli* and the eukaryotes (shown for comparison), there was only one sample. Colour coding for taxonomic levels: red, phylum; blue, class; green, order; magenta, family; cyan, genus; orange, species.



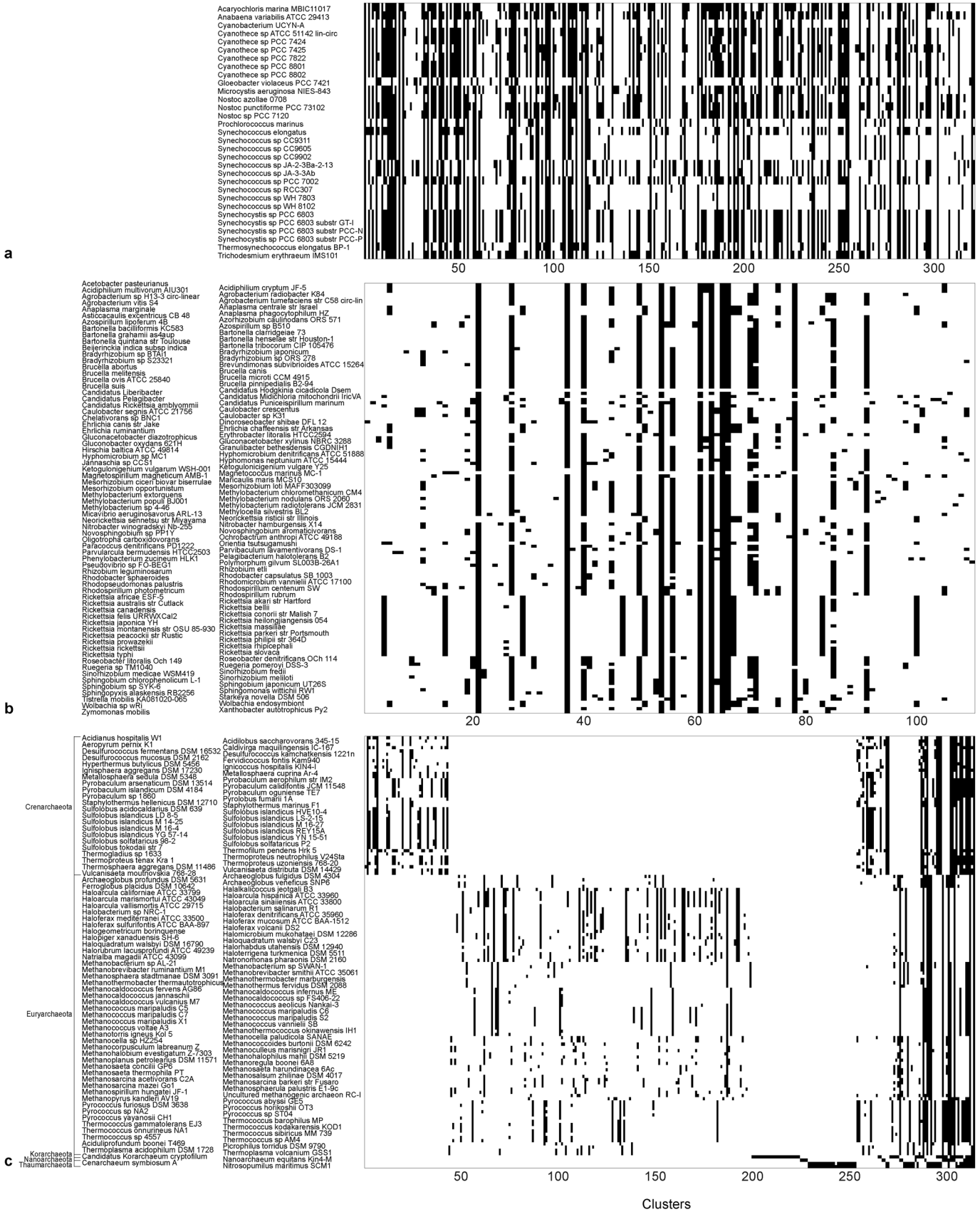
Extended Data Figure 3 | Effect of taxon sampling on eukaryote monophyly in phylogenetic trees. After ten sequences (bold) were added to the original data set (EPC E1689_B206_A295), the relationships among Archaeplastida taxa (highlighted in green) changed from non-monophyly (**a**) to monophyly (**b**). Abbreviations are shown for eukaryotic sequences (Supplementary Table 2) and NCBI GI numbers for cyanobacterial sequences (Supplementary Table 3; RefSeq accessions are shown for the added sequences).



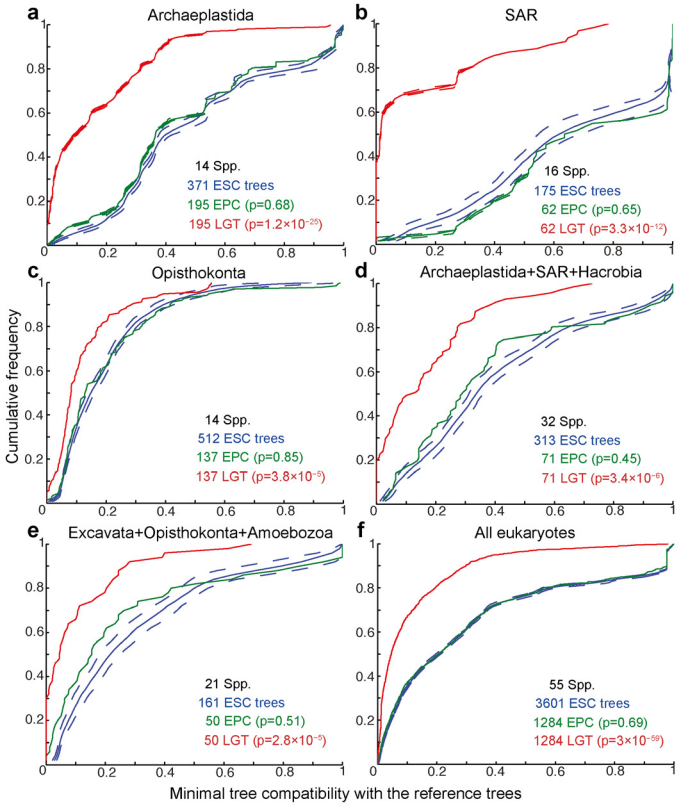
Extended Data Figure 4 | Distribution of prokaryotic taxa in the sister group to eukaryotes, with EPCs sorted by eukaryotic supergroups. Top: each black tick indicates the presence of a eukaryote taxon in one of the 2,585 EPCs. Bottom: each red tick indicates the presence of a prokaryote taxon in the sister group to eukaryotes in one of the 1,933 EPC maximum likelihood trees where eukaryotes were resolved to be monophyletic. The 2,585 EPCs, proteome size, and cluster size are as in Fig. 1. The number of EPCs present and the frequency of occurrence in the sister group to eukaryotes ('clusters') are shown for eukaryotes and prokaryotes, respectively. Archaep., Archaeplastida; Opisth., Opisthokonta; Chl., Chloroplastida; Rho., Rhodophyta; Gla., Glaucophyta; Str., Stramenopila; De., Deinococcus-Thermus; oP., other Proteobacteria; Ch., Chlamydiae; Pl., Planctomycetes; Ve., Verrucomicrobia; Spi., Spirochaetae; The., Thermotogae; oB., other Bacteria. For abbreviations of eukaryotes, see Supplementary Table 1.



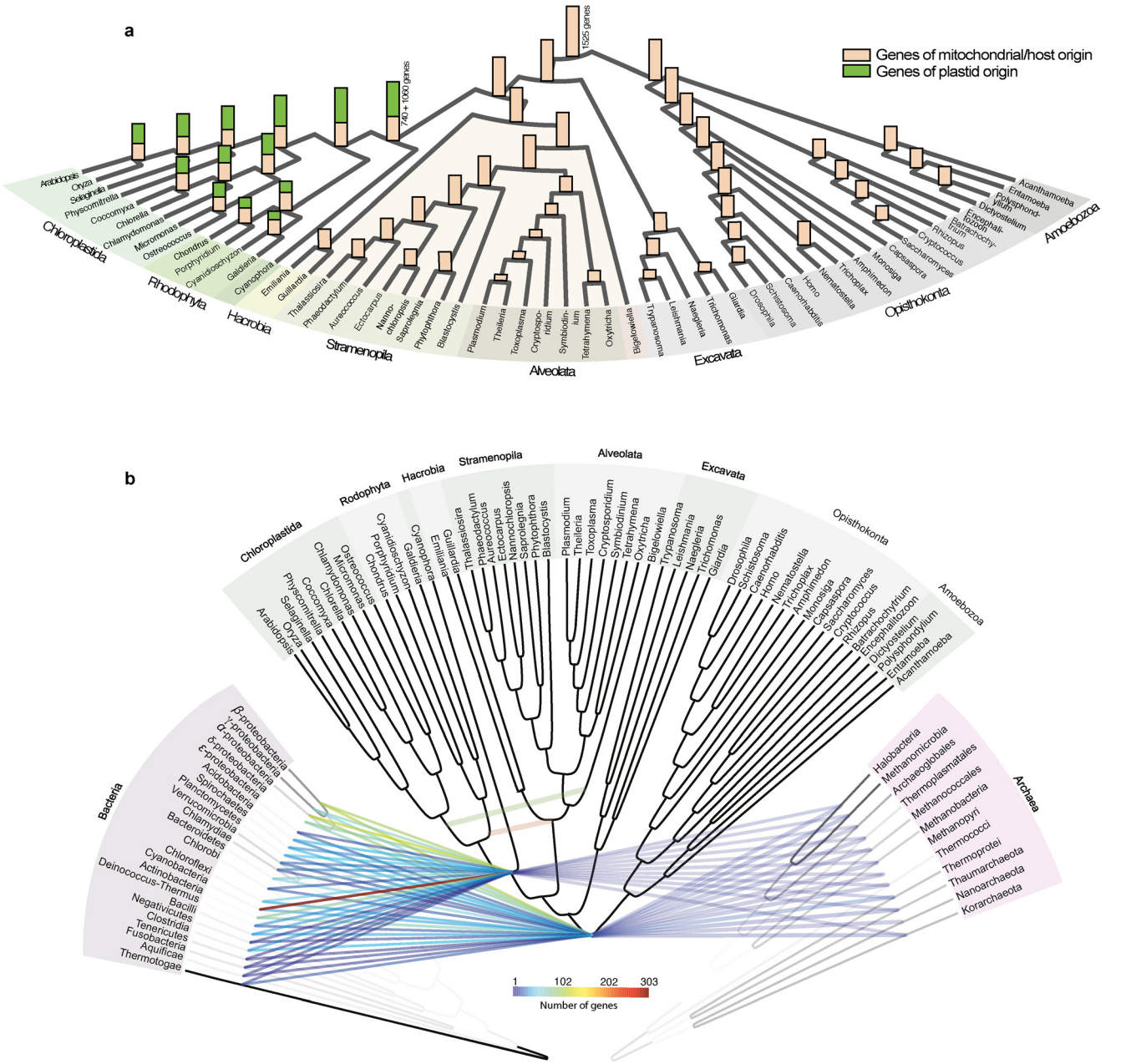
Extended Data Figure 5 | Distribution of prokaryotic taxa in the sister group to eukaryotes, with EPCs sorted by prokaryotic groups. Top: each black tick indicates the presence of a eukaryote taxon in one of the 1,933 EPC maximum likelihood trees where eukaryotes were resolved to be monophyletic. Bottom: each red tick indicates the presence of a prokaryote taxon in the sister group to eukaryotes in one of those 1,933 EPC trees. The EPCs (*x* axis) are ordered according to the taxonomic groups to which the prokaryotes in the sister group to eukaryotes belong (separated into three blocks where only bacteria (1,586 EPCs), only archaea (314 EPCs), or both bacteria and archaea (33 EPCs) are found in the sister group). There are 16 bacterial groups (including 'other Bacteria'; Firmicutes, Proteobacteria, and the PVC superphylum (Planctomycetes, Verrucomicrobia, and Chlamydiae) are regarded as single groups) and five archaeal groups (the five phyla). The number of EPCs present and the frequency of occurrence in the sister group to eukaryotes are shown for eukaryotes and prokaryotes, respectively. Archaep., Archaeplastida; Opisth., Opisthokonta; Chl., Chloroplastida; Rho., Rhodophyta; Gla., Glaucophyta; Str., Stramenopila; De., Deinococcus-Thermus; oP., other Proteobacteria; Ch., Chlamydiae; Pl., Planctomycetes; Ve., Verrucomicrobia; Spi., Spirochaetae; The., Thermotogae; oB., other Bacteria. For abbreviations of eukaryotes, see Supplementary Table 1.



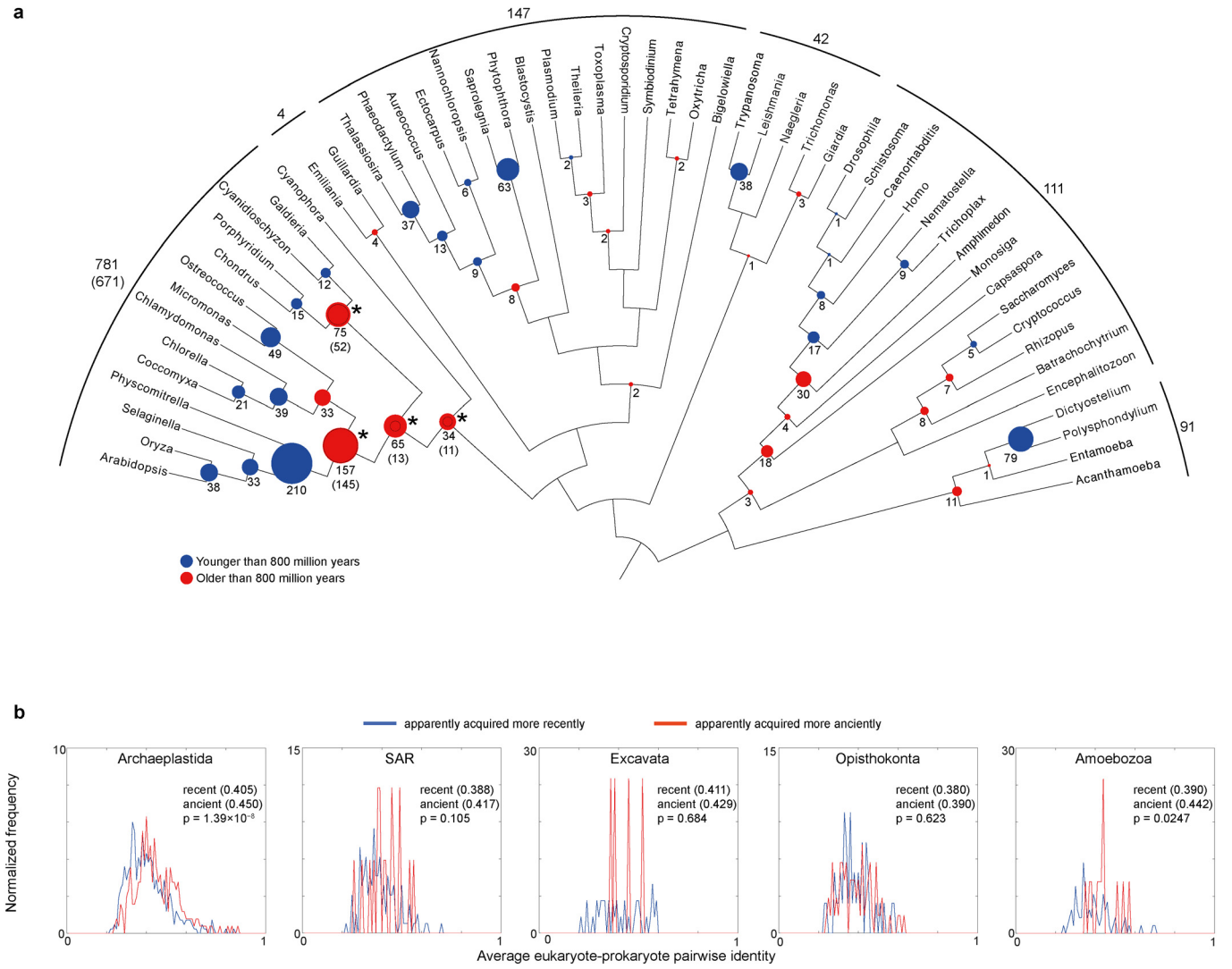
Extended Data Figure 6 | Distribution of taxa in the sister groups consisting purely of cyanobacteria, alphaproteobacteria, or archaea. Each black tick indicates the presence of a prokaryotic taxon in the sister group to eukaryotes in an EPC tree. **a–c**, Distributions of taxa in all pure-cyanobacterial (**a**), pure-alphaproteobacterial (**b**), and pure-archaeal (**c**) sister groups. The clusters are ordered alphanumerically according to the eukaryotic cluster numbers (Supplementary Table 5), whereas for archaea (**c**) the taxa are further sorted by the five archaeal phyla.



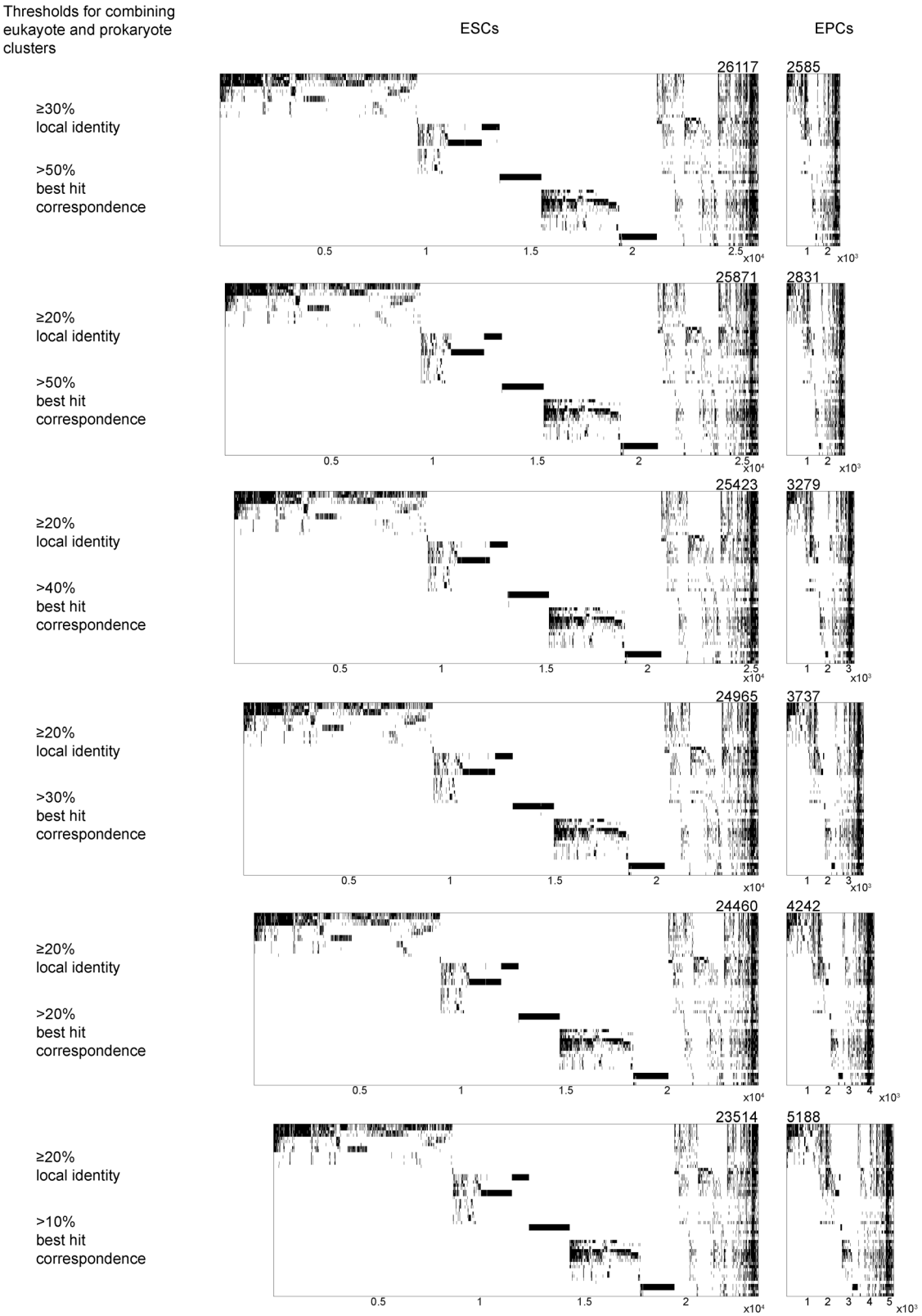
Extended Data Figure 7 | Comparison of sets of trees for single-copy genes in eukaryotic groups, with more inclusive criteria. a–f, Cumulative distribution functions (*y* axis) for scores of minimal tree compatibility with the vertical reference data set (*x* axis). Values are number of species, sample sizes, and *P* values of the two-tailed Kolmogorov–Smirnov two-sample goodness-of-fit test in the comparison of the ESC (blue) data sets against the EPC (green) data set and a synthetic data set simulating one LGT (red). Dashed lines delineate the range of distributions in 100 replicates of random down-sampling. The criteria for tree inclusion were less stringent than those for Fig. 3 (see Methods).



Extended Data Figure 8 | Overview of eukaryote gene content evolution. **a**, Eukaryotic evolution by gene loss. Genome sizes (number of EPCs present) were mapped onto the eukaryotic reference tree. Ancestral genome size in each eukaryotic ancestral node was calculated using a loss-only model, with all EPCs in blocks A–C and those in blocks D and E (Fig. 1) entering the eukaryotic lineage via the plastid ancestor (green) or the eukaryote ancestor (wheat colour). Plastid-derived genes are not shown for the ancestral nodes within SAR and Hacrobia, because of current debates about the number and nature of secondary symbioses, but are indicated by the greenish shading. **b**, Endosymbiotic gene transfer network. The network connecting apparent gene donors to the common ancestor of eukaryotes and Archaeplastida is mapped onto the reference phylogeny (vertical edges) of bacteria (left), eukaryotes (middle), and archaea (right). Grey shading (white to black) in the prokaryote reference trees (70 for archaea and 32 for bacteria) indicates how often a branch associated with a particular node was recovered within the trees of individual genes that were concatenated for inferring the reference topology. Lateral edges indicate gene influx at the origin of eukaryotes and at the origin of plastids. Edge colour corresponds to the frequencies with which a prokaryotic group appears in the sister group to eukaryotes. The archaeal reference tree was rooted between euryarchaeotes and other taxa, and the bacterial tree with Thermotogae. Secondary endosymbiotic transfers are indicated in light green and red. That members of both the Crenarchaeota and the Euryarcheaota are implicated as host relatives is probably because of the small archaeon sample[34–36].



Extended Data Figure 9 | Apparent gene transfers and eukaryote–prokaryote sequence identities. **a**, Patterns suggestive of LGT from prokaryotes inferred from EPC trees. All EPC trees were searched for phylogenetic patterns suggestive of gene acquisitions by the common ancestor of each eukaryote lineage within the six supergroups (see Methods). The size of each circle is proportional to the number of such putative acquisitions, with the total number of putative acquisitions shown for each supergroup. The colour shows the age of nodes according to a eukaryotic time tree (blue, younger than 800 million years; red, older than 800 million years). For the four lineages with an asterisk, phylogenetic patterns where SAR/Hacrobia are nested within a clade formed by Archaeplastida were also counted as putative acquisitions to take into account secondary plastid endosymbioses. The numbers of acquisitions without such patterns are indicated in parentheses (and shown as inner circles). **b**, Eukaryote–prokaryote sequence identities for genes apparently acquired more recently and more anciently in eukaryotes (**a**). The mean of the average pairwise identities is shown in parentheses. At $P = 0.05$, a two-sided Wilcoxon rank-sum test either did not reject the null hypotheses that the two sets of genes are not different or suggested the tip-specific eukaryotic genes are less similar to their prokaryotic homologues.



Extended Data Figure 10 | Distribution of ESCs and EPCs across eukaryotes under different criteria. Different thresholds were applied to find eukaryote clusters with prokaryote homologues, including BLAST local identity for each eukaryote–prokaryote hit (30% or 20%) and levels of best-hit correspondence (10–50%) for identifying reciprocal pairs of eukaryote and prokaryote clusters. Distributions of ESCs and EPCs are drawn as in Extended Data Fig. 1a and Fig. 1, respectively.