On the Prokaryotic Origins of Eukaryotic Genes

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Chuan Ku
aus Taipeh, Taiwan

Düsseldorf, Juni 2016

Aus dem Institut für Molekulare Evolution
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. William F. Martin
Korreferent: Prof. Dr. Martin Lercher

Tag der mündlichen Prüfung: 28. September 2016

Im Laufe dieser Arbeit wurden mit Zustimmung des Betreuers folgende Beiträge veröffentlicht oder zur Veröffentlichung eingereicht:

Verena Zimorski, **Chuan Ku**, William F. Martin, Sven B. Gould: Endosymbiotic theory for organelle origins. Current Opinion in Microbiology 22:38-48.

Chuan Ku, Mayo Roettger, Verena Zimorski, Shijulal Nelson-Sathi, Filipa L. Sousa, William F. Martin: Plastid origin: Who, when and why? Acta Societatis Botanicorum Poloniae 83:281-289.

Chuan Ku, Shijulal Nelson-Sathi, Mayo Roettger, Sriram Garg, Einat Hazkani-Covo, William F. Martin: Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. Proceedings of the National Academy of Sciences of the United States of America 112:10139-10146.

Chuan Ku, Shijulal Nelson-Sathi, Mayo Roettger, Filipa L. Sousa, Peter J. Lockhart, David Bryant, Einat Hazkani-Covo, James O. McInerney, Giddy Landan, William F. Martin: Endosymbiotic origin and differential loss of eukaryotic genes. Nature 524:427-432.

William F. Martin, Mayo Roettger, **Chuan Ku**, Sriram G. Garg, Shijulal Nelson-Sathi, Giddy Landan: Late mitochondrial origin is an artefact. Genome Biology and Evolution (eingereicht)

Chuan Ku, William F. Martin: A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: The 70% rule. BMC Biology (eingereicht)

Summary

All cells can be divided into two forms: the simple prokaryotic cells and complex eukaryotic cells comprising organisms visible to the naked eye. Genetic and biochemical evidences point to a single origin of life on Earth, with eukaryotes arising much later than prokaryotes. Hypotheses abound on the origin of eukaryotes and the origins of eukaryotic genes from prokaryotes. The most widely accepted one, the endosymbiotic theory, suggests that the eukaryotic organelles – mitochondria and plastids – originated as endosymbiotic bacteria that have transferred much of their genomes to the nucleus, which also has abundant archaeal components. Other hypotheses propose that, in addition to those derived from organelle endosymbioses, eukaryotes have received genes from other prokaryotes, intracellular or free-living, through lateral gene transfer (LGT) during specific periods or throughout the history of eukaryotes, and that the majority of eukaryotic genes could have originated thus. After testing these hypotheses, this dissertation provides a clearer picture of eukaryotic genome evolution. The vast majority of eukaryotic genes correspond to two massive acquisitions at the origin of mitochondria from alphaproteobacteria and the origin of plastids from cyanobacteria. Rampant lateral transfer among prokaryotes and gene losses have blurred the phylogenetic information contained in gene sequences, as do imperfect methods for phylogenetic inference and incomplete data sampling. By taking phylogenetic noise into account, it is shown that there is no significant signal from bacteria other than the organelle ancestors such as Chlamydiae. Individual LGTs from prokaryotes to eukaryotes are observed, but there is no long-term cumulative effect. Eukaryote genomes are distinct from those of prokaryotes, as are their cellular structures, and there is a natural barrier to LGT across the eukaryote-prokaryote divide.

Zusammenfassung

Alle Zellen fallen in zwei Kategorien: die einfachen prokaryotischen Zellen und die komplexen eukaryotischen Zellen, aus denen fast alle mit bloßem Auge zu erkennenden Lebewesen bestehen. Genetische und biochemische Befunde deuten auf einen gemeinsamen Ursprung aller Lebewesen auf der Erde, wobei Eukaryoten viel später als Prokaryoten entstanden sind. Es gibt zahlreiche Hypothesen über den prokaryotischen Ursprung eukaryotischer Gene. Die anerkannteste davon, die Endosymbiontentheorie, besagt, dass die eukaryotischen Organellen, Mitochondrien und Plastiden, aus endosymbiotischen Bakterien entstanden, von denen die Mehrzahl der Gene bereits auf den zugleich mit vielen archaeellen Genen ausgestatteten Zellkern transferiert worden sind. Anderen Hypothesen zufolge haben Eukaryoten während bestimmter Phasen oder auch über ihre gesamte Evolutionsgeschichte hinweg zusätzlich zu den obengenannten Quellen auch Gene von anderen, intrazellulären oder freilebenden Prokaryoten durch lateralen Gentransfer erhalten, aus dem die meisten eukaryotischen Gene ihren Ursprung haben könnten. Diese Dissertation vermittelt ein besseres Bild der evolutionären Entwicklung der eukaryotischen Genome durch die Überprüfung der unterschiedlichen Hypothesen. Die Ergebnisse zeigen, dass die überwiegende Anzahl eukaryotischer Gene während zwei massiver Transferereignisse aufgenommen wurde, der Entstehung der Mitochondrien aus Alphaproteobakterien, sowie der Entstehung der Plastiden aus Cyanobakterien. Weitreichender lateraler Transfer zwischen Prokaryoten und der Verlust von Genen, sowie Fehler in der phylogenetischen Rekonstruktion und unvollständige Datensätze, haben jedoch das phylogenetische Signal einzelner Gensequenzen undeutlich gemacht. Wenn diese Störungen berücksichtigt werden, verbleibt kein bedeutsames Signal von Bakterien, die nicht die Vorläufer der Organellen sind, wie z.B. Chlamydien. Einzelne Gentransfers von Prokaryoten zu Eukaryoten mögen vorkommen, aber es gibt keine langfristigen kumulativen Auswirkungen. Eukaryotische Genome sind so verschieden von den prokaryotischen, wie sich auch ihre zellulären Strukturen unterscheiden, und es existiert ein natürliches Hindernis für den lateralen Transfer über die eukaryotisch-prokaryotische Grenze hinweg.

Table of Contents

1 Introduction

1.1
Origin, evolution and loss of genes

Every organism stores information for making proteins or RNAs in units of nucleic acid called genes. The differences in gene content and sequences contribute to the biological diversity we observe and is a major topic in evolutionary genomics. To study how organisms diversify, we need to understand the dynamics of gene births, changes, and deaths.

A gene can arise *de novo* or by copying from a pre-existing source. The former involves origination from non-genic regions (Carvunis et al. 2012), while the latter includes both gene duplication within a genome (Ohno 1970) and transfer of genes from other genomes vertically or laterally (horizontally). Whereas duplications create a family of genes similar in sequences, only *de novo* origination and lateral transfer would bring in genuinely new genes that are absent in the parental organism(s). There is no doubt that genes in the earliest life must have arisen *de novo*; however, the role of *de novo* origination in subsequent evolution still remains unclear. On the contrary, lateral gene transfer is a well-known factor in contributing to gene content diversity (cf. 317 results for the topic "de novo gene" and 11,324 results for "lateral gene transfer" or "horizontal gene transfer" on Web of Science as of May 10, 2016). The biological significance of lateral gene transfer will be discussed in detail in section 1.2.

Genes within a genome are subject to several types of mutations. On the one hand, there are point mutations (nucleotide substitutions, insertions, and deletions) driving the gradual sequence divergence between the same gene in different organisms and between different genes within the same gene family (Graur 2016). On the other hand, larger-scale mutational processes can cause regional duplications and rearrangements within a gene or between a gene and another genomic region. When different genes are recombined, fusion genes may occur, whose expressed products contain the properties of the individual products (Espinosa et al. 2001; Marsh and Teichmann 2010).

Mutations can eventually lead to the loss of genes (Albalat and Cañestro 2016). In the simplest case, genes can be entirely lost as part of chromosomal deletion. Small-scale mutations can also lead to gene loss. Nucleotide insertions or deletions can create frameshift mutations resulting in a premature stop codon that significantly shortens a gene. Accumulation of multiple point mutations can change the sequence of a gene to be significantly different from that of the same gene in other genomes. Although point mutations only incrementally change the gene sequence, a gene can appear to be lost when the extent of change is reached that it can no longer be recognized (dependent on detection methods) as the same gene.

While the major processes (Figure 1) affecting the dynamics of gene content and sequences are known, it is still challenging to reconstruct the history of genome dynamics accurately. Often it is difficult or even impossible to determine which processes have resulted in an observed pattern. For example, one explanation for discordance between gene trees is lateral gene transfer (Ravenhall et al. 2015). Another possibility is gene duplication followed by differential gene loss (Gogarten and Townsend 2005). In sexually reproducing diploid or polyploid organisms, incomplete lineage sorting and gene flow among incipient lineages can also result in discrepancy between trees (Rogers and Gibbs 2014). Another example is illustrated by the fact that genes changes through time. A gene that is not detected might never have been there, it might have been lost, or it might have gone undetected because it was fused with another gene or because it has undergone many mutations.

Therefore, all potential factors should be taken into account for reconstructing the ancient, complex, and dynamic history of genomes, which poses a major challenge to modern evolutionary biology. In this dissertation, the case of eukaryotes will be highlighted to show how we can better understand genome evolution with new methods, analyses, and perspectives.

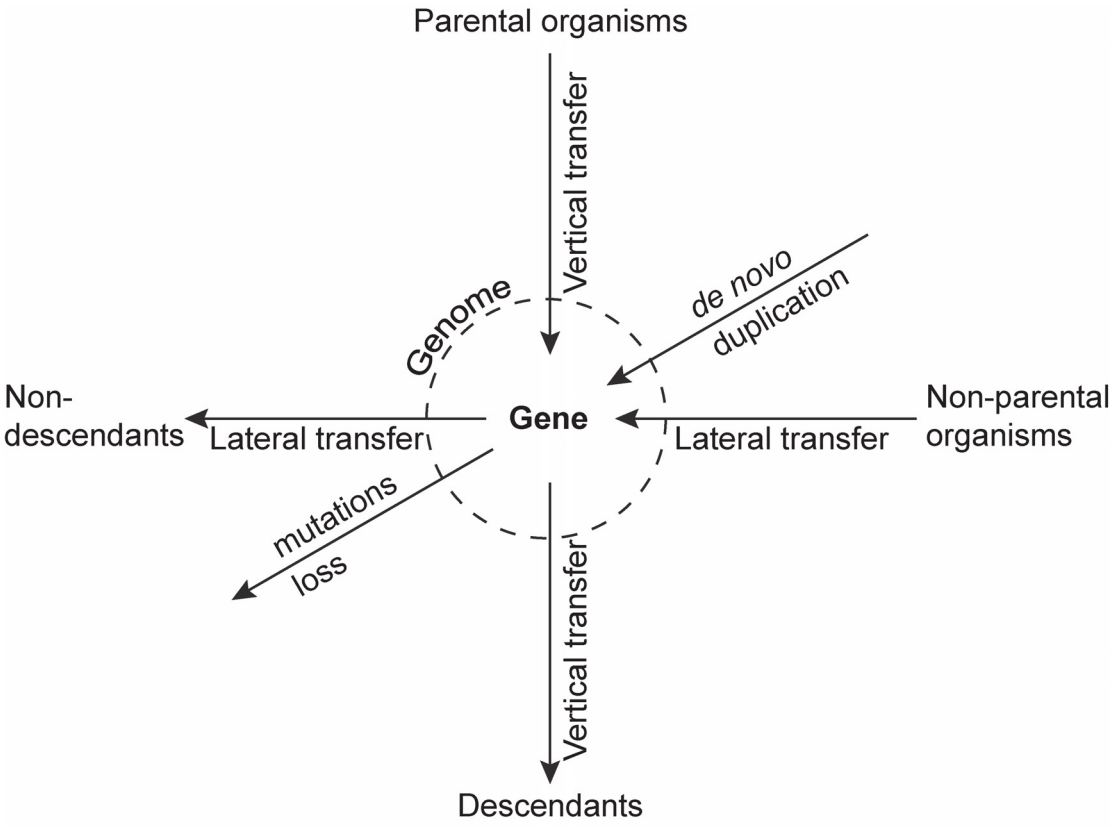


Figure 1. A three-dimensional view of genes within a genome. The arrows represent the progress in time. Two axes indicate the vertical and lateral gene transfer from one organism to another, while a third axis shows intra-genomic changes. They summarize the possible sources (vertical transfer, lateral transfer, *de novo* origination, duplication), changes (mutations), and death (loss) of a gene. For a comparison of the definitions of lateral gene transfer used here and in other studies, see Figure 2.

1.2
Lateral gene transfer

1.2.1 Definitions

The term, lateral gene transfer (LGT) or horizontal gene transfer (HGT), is used to contrast with vertical transfer and could be easily understood as 'non-vertical' transfer of genes. However, if one looks into the literature, one would find that there is no uniform definition for LGT and that the differences between the ones adopted by different authors can be non-trivial. Some examples of the definitions of LGT/HGT are:

(a) "nonsexual transfer of genetic information between genomes" (Kidwell 1993)

(b) "Transfer of a gene from one genome to another at some point in the evolutionary process: an outcome, not a specific genetic mechanism." (Doolittle 1999a)

(c) "the transfer of genes between different species" (Koonin et al. 2001)

(d) "the non-genealogical transfer of genetic material from one organism to another" (Goldenfeld and Woese 2007)

(e) "the sharing of genetic material between organisms that are not in a parent-offspring relationship" (Soucy et al. 2015)

The criteria used for defining LGT includes sex (Kidwell 1993), species boundary (Koonin, et al. 2001), genealogy (Goldenfeld and Woese 2007), gene sharing (Soucy, et al. 2015), or simply "transfer" of genes (Doolittle 1999a). While some definitions, such as "non-genealogical" or "transfer of a gene from one genome to another" ("transfer" can be either lateral or vertical (Kidwell 1993; Schönknecht et al. 2014)), seem too ambiguous to be interpreted in the same way by different researchers, others are obviously at odds with each other or with classical studies demonstrating LGT. It is clear that not all forms of nonsexual transfer of genetic information should be considered LGT, since many of them are just normal, vertical, asexual reproduction occurring in prokaryotes (binary fission) or in eukaryotes (e.g., plant vegetative reproduction, yeast budding (de Meeus et al. 2007)). When it comes to LGT, the transfer of genes across the species boundary is often seen as the defining characteristic

(Andersson 2005; Koonin, et al. 2001; Schönknecht, et al. 2014; Syvanen 2012), which is popularized Syvanen's paper (1985) *Cross-species gene transfer*. This common definition, however, has two problems. First, there is no universal species definition, especially for prokaryotes (Doolittle 1999b; Konstantinidis et al. 2006). Second, intraspecific transfers should not be excluded, as the earliest works showing movement of DNA through LGT involved different strains of the same species such as *Streptococcus pneumoniae* (Griffith 1928) and *Escherichia coli* (Tatum and Lederberg 1947). In the classical review paper *Biochemical Evolution* by Allan Wilson and colleagues (1977), both transfers within a species and between species are considered to be LGT. A reasonable definition should therefore not be limited to cross-"species" transfers.

Perhaps the most inclusive definition of LGT to date is gene sharing between organisms not in a parent-offspring relationship (Soucy, et al. 2015). Based on this definition, the authors also include as a type of LGT introgression (Soucy, et al. 2015), the gene flow from one entity (population or species) to another through hybridization and backcrossing (Harrison and Larson 2014; Rhymer and Simberloff 1996), both of which are sexual reproduction that has been rarely, if ever, considered as mechanism of LGT. In fact, according to the overly inclusive definition, one could easily arrive at the conclusion that genes shared by *Arabidopsis* and *Drosophila*, which clearly are not in a parent-offspring relationship, are the results of LGT, as are those shared by sister species, cousins, or even sisters and brothers.

Clearly enough it is necessary to have a sound definition of LGT in studies where the distinction between vertical and lateral is important. In this dissertation, LGT is defined as: transfer of a gene to an organism from a non-parental organism. Here non-parental organisms are those not in a reproductive relationship with the recipient organism of LGT, whereas parental organisms include those in asexual reproduction of unicellular organisms (parent cells), in asexual reproduction of multicellular organism (parent plants, spore-producing organisms, etc.), and in sexual reproduction (gamete-producing organisms). Overall, this definition is more exclusive than the one based on sex, but more inclusive than that based on the species boundary (Figure 2).

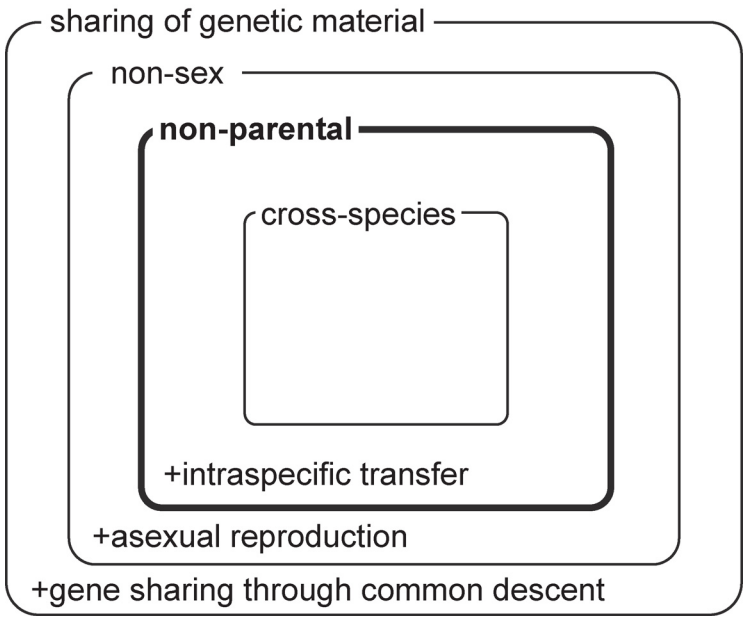


Figure 2. A comparison of definitions of LGT. The definition adopted in this dissertation is shown in bold.

1.2.2 Mechanisms

Ever since the first observation of lateral transfer, its mechanisms have been a central question in LGT studies. For LGT among prokaryotes, several mechanisms have been found that can be classified according to the media by which genetic information is transferred:

(a) Transformation: The first mechanism of lateral transfer was discovered in 1928 with Griffith's experiments showing a strain of *S. pneumoniae* can be transformed by taking up genetic material, later proved to be DNA (Avery et al. 2000), from another strain (Griffith 1928). Components in the type II and type IV secretion systems, as well as ABC transporters, are involved in the uptake of DNA from the environment through the cell membrane (Chen and Dubnau 2004).

(b) Conjugation: A mechanism requiring cell contact was found in *E. coli*, which led to gene recombination between different strains (Lederberg and Tatum 1946; Tatum and Lederberg 1947). Similar to transformation, conjugative DNA transfer involve type IV secretion system proteins that form pili or surface adhesins (Chen et al. 2005) to transfer diverse DNA molecules, such as plasmids and integrative and conjugative elements (Wozniak and Waldor 2010), from one cell to another.

(c) Transduction: A few years after Lederberg and his mentor Tatum reported the phenomenon of conjugation, he and his student Zinder discovered another mechanism with phages as the medium of LGT in *Salmonella typhimurium* (Zinder and Lederberg 1952). Phage particles transfer either host DNA adjacent to prophage DNA (specialized transduction) or any random segment of the host DNA (generalized transduction) to another host cell (Ikeda and Tomizawa 1965).

(d) Gene transfer agent (GTA): First reported in the 1970s (Marrs 1974), GTAs are phage-like entities that differ from transducing phages in that they cannot package the complete set of genes required for GTA production and that GTA-encoding DNA normally does not get replicated or excised from the host genome (Lang et al. 2012). GTAs are widely found in alphaproteobacteria (Lang, et al. 2012) and are important for LGT in oceanic environments (McDaniel et al. 2010).

(e) Others: For LGT in archaea, there exist not only the aforementioned mechanisms (Allers and Mevarech 2005; Lang, et al. 2012), but also more recently found cell fusion, where bidirectional gene exchange occurs (Naor and Gophna 2013; Rosenshine et al. 1989). A recently found mechanism in bacteria involves the formation of nanotubes for exchange of cellular molecules (Dubey and Ben-Yehuda 2011).

In contrast to the situation in prokaryotes, in eukaryotes relatively little is known about the *mechanisms* of LGT despite many *reports* of LGT. No proteins are known to mediate DNA uptake by or transfer to a eukaryotic cell and there have been few direct experimental observations of naturally occurring LGT to eukaryotic cells. Notable

exceptions include transfer of T-DNA from *Agrobacterium* to plants (Gelvin 2000) and a report of conjugative transfer from *E. coli* to *Saccharomyces cerevisiae* (Heinemann and Sprague 1989), but the former is limited to genes encoded by T-DNA while further studies on the mechanism and importance of the latter are lacking. Hypotheses have been proposed for how foreign DNA can enter the eukaryotic cell, including viral infection, feeding (phagocytosis), and endosymbiosis (Doolittle 1998; Schönknecht, et al. 2014; Soucy, et al. 2015). However, there have been few if any direct observations that genes are transferred to the eukaryotic genome through these processes.

1.2.3 Lateral gene transfer as an explanatory principle

For the first 60 years since its discovery in 1928, research on LGT focused on observation of LGT as a genetic phenomenon and experimental dissection of its processes and mechanisms. Since the first complete genome sequence of an organism was available (Fleischmann et al. 1995), sequence data have provided exciting insights into the important role of LGT in shaping prokaryotic genomes. At the same time, it also shifted the research on LGT from direct observations and experiments to sequence similarity- or tree-based approaches. In the genomic era, LGT figures in mainly as an explanatory principle. Whereas traditional studies on LGT are based on observations that some organisms acquired genes through LGT, more recent *reports of LGT* tend to be based on observations that some organisms have genes showing patterns that *can* be explained by LGT, although it is by no means the only possible explanation. While genomic data are being produced at an ever increasing rate and when inferring the affinity of genes by similarity search or phylogenetic analyses become a common practice, LGT is an easy-to-use explanatory principle for patterns that seem not to conform to vertical transfers. But such *reports of LGT* should always be taken with a grain of salt, as they do not provide evidence for LGT, but LGT as a hypothesis for origin of genes.

More often than not there exist other possibilities that are compatible with vertical transfer, when one takes into account the quality of the underlying data

(genomic sequences), methods for tree reconstruction, and how trees are interpreted. Genes reported to stem from an organism can be contaminating sequences from other organisms, such as in the genome reports of the tardigrade *Hypsibius dujardini* (Bemm et al. 2016; Koutsovoulos et al. 2016), the sea anemone *Nematostella vectensis* (Artamonova et al. 2015; Artamonova and Mushegian 2013), and the moss *Physcomitrella patens* (Lang et al. 2008). Artefacts can also be found in the phylogenetic trees. Phylogeneticists know very well that not every tree depicts the correct gene history, since many trees, each of which itself being an estimation of evolutionary processes, just contradict each other. All kinds of phylogenetic errors stemming from evolutionary models, taxonomic sampling, sequence alignment methods, tree reconstruction methods can result in an incorrect tree (Felsenstein 2004; Semple and Steel 2003) from which an incorrect hypothesis about gene origin may be formulated. Finally, even with the perfect sequence data and trees, there are several ways to interpret an individual tree. The origin of a gene in a lineage can be explained by acquisition through LGT, but it can also be explained by, for example, gene duplication followed by differential gene loss (Gogarten and Townsend 2005; Martin and Schnarrenberger 1997). Incomplete taxon or sequence sampling can also lead to identification of the wrong sister group of a clade, and thus, the wrong source of a gene. Even for LGT between two organisms, there can be more than one hypothesis, when transfers in both directions create the same phylogenetic pattern (for example, Wolf et al. (1999) vs. Greub and Raoult (2003) on the ATP/ADP translocase gene).

It is therefore crucial to distinguish between LGT as an observed phenomenon and LGT as an explanatory principle. When using the latter, other competing hypotheses should also be taken into account.

1.3

Endosymbiotic theory

The endosymbiotic theory developed in the days where the phenomenon of lateral gene transfer was unknown, even before the term "gene" was coined (Johannsen 1909). Based on similarities between plastids and free-living cyanobacteria, Mereschkowsky was the first to propose that plastids originated as cyanobacteria that became endosymbiotic within a plastid-free ancestral cell (Martin and Kowallik 1999; Mereschkowsky 1905). A similar endosymbiotic origin from bacteria was also proposed for mitochondria (Wallin 1927). Forty years later, the endosymbiotic theory was further synthesized and popularized among biologists by Lynn Margulis in her works on the origin of eukaryotic cell (Margulis 1970; Sagan 1967). A major support for the endosymbiotic theory came from the discovery that the two organelles contain stable extranuclear DNA (Nass and Nass 1963; Sager and Ishida 1963; Schatz et al. 1964) that appears to be the genome of their endosymbiotic bacterial ancestors. Additional sequencing and phylogenetic analyses showing that plastid and mitochondrial genomes are remnants of cyanobacterial and alphaproteobacterial genomes, respectively, provide further evidence for the endosymbiotic theory (Gray et al. 1999; Gray and Doolittle 1982; Martin et al. 1998), which is now widely accepted by biologists (Archibald 2014; Baum 2013; Degli Esposti et al. 2014; Deschamps and Moreira 2009; Falcón et al. 2010; Gould et al. 2008; Howe et al. 2008; Martin et al. 2012). It has also been suggested that plastids have been spread across eukaryotic lineages through secondary (involving Archaeplastida, eukaryotes with primary plastids) and even tertiary (involving eukarytoes with secondary plastids) endosymbioses (Gould, et al. 2008; Keeling 2013), further underscoring the role of endosymbiosis in organelle origination.

Sequences of the organellar genomes, however, also reveal one paradox, namely that the organellar DNA encodes only ~1% of genes found in the genomes of ordinary cyanobacteria or alphaproteobacteria, yet the organelles perform diverse biochemical functions found in free-living bacteria and contain over a thousand proteins (Mower and Bonen 2009; Richly and Leister 2004). This discrepancy is accounted for by the transport of polypeptides encoded in the nuclear genome to the organelles through translocases (Schleiff and Becker 2011). That vast majority of genes for bacterial

functions of the organelles are nucleus-encoded is explained by endosymbiotic gene transfer, where genes of the original endosymbiotic bacteria were transferred to the nuclear genome while the original copies were lost, resulting in the reduced bacterial genomes in present-day organelles (Embley and Martin 2006; Martin et al. 1993; Martin and Herrmann 1998; Martin et al. 2002; Timmis et al. 2004). Unlike reported LGTs from endosymbiotic bacteria, which is based on genome or transcriptome sequences of the nucleus and bacteria (Hotopp et al. 2007; Husnik et al. 2013), endosymbiotic gene transfer from organelles is an ongoing, observable process (Huang et al. 2005; Ju et al. 2015; Timmis, et al. 2004). Through organellar genomes have been much reduced, recently transferred nuclear copies of mitochondrial DNA (numts) and plastid DNA (nupts) are widespread among eukaryotes and have important impacts on the nuclear genome structure and variation (Hazkani-Covo and Covo 2008; Hazkani-Covo et al. 2010; Kleine et al. 2009).

1.4

Hypotheses on gene origins in eukaryotes

Since sequencing of genes became a common technique, scientists have always been intrigued by the relationships between the eukaryotic genomes – namely the nuclear, mitochondrial and plastid genomes – and the prokaryotic genomes (Bonen and Doolittle 1975; Woese and Fox 1977). Genomic information bears pivotal insights into the origin of the eukaryotic cell, whose separation from prokaryotes is regarded by some as the "greatest evolutionary discontinuity between living organisms" (Raff and Mahler 1972). Except for very few proposed cases of LGT to organelles, there is little dispute that plastid- or mitochondrion-encoded genes originated from the endosymbiotic cyanobacteria or alphaproteobacteria, respectively. There is, however, much dispute to the origins of genes in the nuclear genome. While most authors agree that it has a chimeric origin, there is little agreement on what sources constitute this chimerism other than the contributions from archaea and from the endosymbiotic mitochondrial and plastid ancestor (the three *cornerstone* partners). Different proposed origins are summarized in Table 1.

Table 1. Proposed prokaryotic sources of eukaryotic genes.

Sources of eukaryotic genes	Supporting references
I. Three cornerstone partners	
I.1 Archaea	Margulis (1996), Martin and Müller (1998), Rivera et al. (1998), Pisani et al. (2007), Williams et al. (2013)
I.2 Proteobacterial ancestor of mitochondria	Gray (1993), Martin and Müller (1998), Gray et al. (1999), Esser et al. (2004), Pisani et al. (2007)
I.3 Cyanobacterial ancestor of plastid	Gray (1993), Martin and Herrmann (1998), Martin et al. (2002), Pisani et al. (2007)
II. Other particular prokaryotes plays an important role	
II.1 Spirochetes	Margulis (1996)
II.2 Deltaproteobacteria	Moreira and López-García (1998), López-García and Moreira (1999)
II.3 Actinobacteria	Cavalier-Smith (2002)
II.4 Chlamydiae	Huang and Gogarten (2007) and various references in red listed in Figure 3
III. Continuous transfers from various other prokaryotes	Doolittle (1998), Larkum et al. (2007), Howe et al. (2008), Andersson (2009), Yue et al. (2012)
IV. Major contributions from various other prokaryotes	
IV.1 Prior to mitochondrial endosymbiosis	Gray et al. (2014), Pittis and Gabaldón (2016)
IV.2 Prior to or during plastid establishment or related to plastid functions	Suzuki and Miyagishima (2010), Reyes-Prieto and Moustafa (2012), Qiu et al. (2013b)
IV.3 Throughout eukaryote evolution/ No specific time points	Andersson (2005), Keeling and Palmer (2008) Syvanen (2012), Huang (2013), Qiu et al. (2013b), Boto (2014), Schönknecht et al. (2014)

A case that has received much attention involves genes that show special affinity between Archaeplastida, eukaryotes with primary plastids (Adl et al. 2005), and Chlamydiae, intracellular pathogenic or endosymbiotic bacteria found in animals and amoebae (Horn 2008). It illustrates how distinct scenarios (vertical transfer, LGT in different directions, and other hypotheses) can explain the same patterns, how different factors should be taken into account for studying genes across deep phylogenies, and how hypotheses on the origin of a group of genes could develop.

For the Archaeplastida-Chlamydiae genes (Figure 3), one can divide the development of the related hypotheses during the nearly 20 years since the first report (Stephens et al. 1998) of those genes into three phases. In the beginning, the topic was debated with diverse hypotheses: origin in Archaeplastida followed by transfer to Chlamydiae (Linka et al. 2003; Stephens, et al. 1998; Wolf, et al. 1999), bacterial genes transferred from Chlamydiae to Archaeplastida (Greub and Raoult 2003), origin in proteobacteria followed by mitochondrial endosymbiosis, differential gene loss in eukaryotes, and transfer to Chlamydiae (Amiri et al. 2003), or alternatives to Archaeplastida-Chlamydiae transfers such as "complex ancestral gene transfers" among Archaeplastida, cyanobacteira, and Chlamydiae (Horn et al. 2004), LGT among prokaryotes before mitochondrial (Royo et al. 2000) or plastid (Schmitz-Esser et al. 2004) endosymbiosis, or after the plastid endosymbiosis (Lange et al. 2000; Rujan and Martin 2001), phylogenetic artefacts or limited resolution (Lange, et al. 2000), and gene loss in or inadequate taxon sampling of cyanobacteria (Brinkman et al. 2002).

The second phase is marked by *ad hoc* attempts to find Archaeplastida-Chlamydiae genes using newly available genomic data. Instead of focusing on specific genes (e.g., ATP/ADP translocase) as in the first phase, these studies tried to list such genes for whole genomes or a specific compartment (mostly plastid) and, following the first such study (Huang and Gogarten 2007), attribute them to Chlamydiae-to-Archaeplastida transfers around the time of plastid establishment (Becker et al. 2008; Huang and Gogarten 2008; Moustafa et al. 2008; Park et al. 2009; Price et al. 2012; Reyes-Prieto and Moustafa 2012; Suzuki and Miyagishima 2010; Tyra et al. 2007), a view shared by related review or perspective articles (Ball et al. 2011; Fournier et al. 2009; Horn 2008).

In the first and second phases, the focus was on inferring scenarios of the organism-level history of Archaeplastida-Chlamydiae genes based on phylogenetic trees, each of which, as mentioned earlier, is a hypothesis on the relationships of the sequences it contains. In the third, it shifted to reconstruction of biochemical interactions in the scenario of cryptic (i.e., not found in extant Archaeplastida) chlamydial endosymbionts supposed to have helped established the plastid and to have transferred genes to Archaeplastida (Ball et al. 2016b; Ball et al. 2015; Ball and Greub 2015; Ball et al. 2013; Cencil et al. 2014; Facchinelli et al. 2013a; Facchinelli et al. 2013b; Karkar et al. 2015; Qiu et al. 2013a; Subtil et al. 2014). Such proposed biochemical interactions based on the premise of the existence of cryptic endosymbiotic chlamydiae, in the view of some authors (Ball et al. 2016a), can be used as "evidence" to support their existence, although biochemistry itself does not point to a particular lineage. Since the trees showing chlamydial sister groups, if they are correct and complete, are the only data that can be interpreted to imply a chlamydial participant at plastid origin, the argument based on chlamydial biochemistry becomes circular. What is much more problematic, however, is the circumstance that all prokaryotic groups appear as donors of genes of to the plant lineage, not just chlamydiae (Dagan et al. 2013). Thus, if one looks at all the data, there are two possibilities. Either i) there is no evidence for a specific participation of chlamydiae at plastid origin, or ii) the "chlamydioplast" (Facchinelli, et al. 2013a) hypothesis should be expanded to include a bacilloplast, a spirochetoplast, a clostridioplast, and so forth (Dagan, et al. 2013). Several articles, including those included in this cumulative dissertation, have pointed out pitfalls of inferring the cryptic "chlamydioplast" (Facchinelli, et al. 2013a) based on trees (Dagan et al. 2008; Dagan and Martin 2009; Dagan, et al. 2013; Deschamps 2014; Deusch et al. 2008; Domman et al. 2015; Moreira and Deschamps 2014; Thiergart et al. 2012).

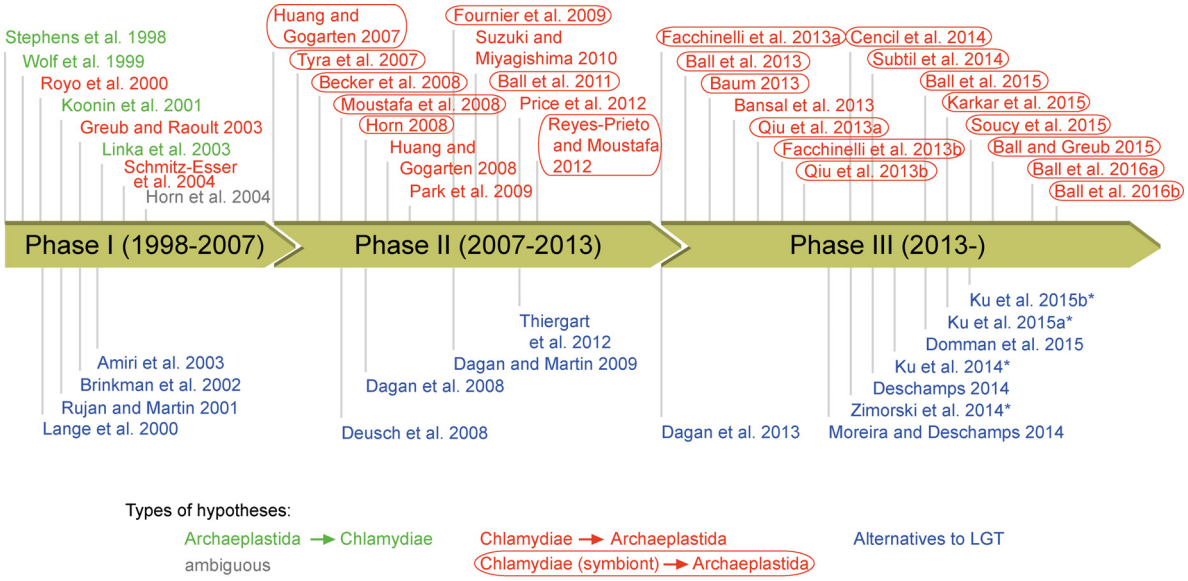


Figure 3. Hypotheses on the evolution of Archaeplastida-Chlamydiae genes. *Papers included in this dissertation.

2 Aims of the dissertation

There exist various hypotheses on the origins of eukaryotic genes, but few have been tested at the genome-wide level for many species. To better our understanding of eukaryotic genome evolution, this dissertation aims to elucidate the prokaryotic sources of eukaryotic genes, with special emphasis on genes of bacterial origins. There will be theoretical components, as well as analyses of empirical data that put the different hypotheses to the test. Specific goals of the individual publications are as follows:

(a) Origin of organelles and eukaryotic genes in light of endosymbiotic theory

(b) The setting of plastid endosymbiosis

(c) Impact of LGT on prokaryotic genomes and structure of prokaryotic pangenomes, and how they affect inference of prokaryotic origins of eukaryotic genes

(d) Testing the hypotheses on the prokaryotic origins by phylogenetic and statistical analyses of eukaryotic-prokaryotic sequence clusters

(e) Contrasting the impacts of LGT on prokaryotic and eukaryotic genomes.

3 Publications

3.1

Endosymbiotic theory for organelle origins

Verena Zimorski, Chuan Ku, William F Martin, Sven B Gould
Institute of Molecular Evolution, Heinrich-Heine-University of Düsseldorf, Germany
Corresponding author: bill@hhu.de

The presented manuscript was published in the journal *Current Opinion in Microbiology* in 2015.

Contribution of Chuan Ku (second author)
Manuscript writing: 15%

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

Endosymbiotic theory for organelle origins

Verena Zimorski, Chuan Ku, William F Martin and Sven B Gould

Endosymbiotic theory goes back over 100 years. It explains the similarity of chloroplasts and mitochondria to free-living prokaryotes by suggesting that the organelles arose from prokaryotes through (endo)symbiosis. Gene trees provide important evidence in favour of symbiotic theory at a coarse-grained level, but the finer we get into the details of branches in trees containing dozens or hundreds of taxa, the more equivocal evidence for endosymbiotic events sometimes becomes. It seems that either the interpretation of some endosymbiotic events are wrong, or something is wrong with the interpretations of some gene trees having many leaves. There is a need for evidence that is independent of gene trees and that can help outline the course of symbiosis in eukaryote evolution. Protein import is the strongest evidence we have for the single origin of chloroplasts and mitochondria. It is probably also the strongest evidence we have to sort out the number and nature of secondary endosymbiotic events that have occurred in evolution involving the red plastid lineage. If we relax our interpretation of individual gene trees, endosymbiotic theory can tell us a lot.

Addresses
Institute of Molecular Evolution, Heinrich-Heine-University of Düsseldorf, 40225 Düsseldorf, Germany

Corresponding author: Martin, William F (bill@hhu.de, w.martin@hhu.de)
Dedicated to Klaus V Kowallik on the occasion of his 75th birthday.

Current Opinion in Microbiology 2014, **22**:38–48

This review comes from a themed issue on **Growth and development: eukaryotes**

Edited by **Michael Bölker**

http://dx.doi.org/10.1016/j.mib.2014.09.008

Introduction

Endosymbiotic theory posits that plastids and mitochondria were once free-living prokaryotes and became organelles of eukaryotic cells. The theory started with plastids [1] and was further developed for mitochondria [2]. It was rejected by cell biologists in the 1920s and revived in the 1960s [3]. The main strength of the theory is that it accounts for the physiological and biochemical similarity of organelles to prokaryotic cells [4,5]. Important evidence in support of endosymbiotic theory comes from organelle genomes. Organelles tend to retain a miniaturized prokaryotic chromosome encoding 200 proteins or less in the case of plastids [6] or 63 proteins or less in the case of mitochondria [7]. Despite that genome reduction, both organelles harbour on the order of 2000 proteins each [8,9], which are involved in a broad spectrum of pathways germane to their ancestrally prokaryotic biochemistry. The discrepancy between the number of proteins that organelles encode and the number of proteins that they harbour is generally explained by a corollary to endosymbiotic theory involving gene transfer to the nucleus, or endosymbiotic gene transfer (EGT). During the course of evolution, many genes were transferred from the organelles to the chromosomes of their host. In the early phases of organelle evolution, before the invention of the protein import apparatus that allowed plastids and mitochondria to import proteins from the cytosol, the transferred genes either became pseudogenes or became expressed as cytosolic proteins. With the advent of organelle protein import, the transferred genes could obtain the necessary expression and targeting signals to be targeted back to the organelle from which the nuclear gene was acquired [10]. For functions essential to the organelle, only the third case allowed the gene to be lost from organelle DNA [11]. This process of organelle genome reduction has resulted in an expansion of the eukaryotic nuclear gene repertoire and in reductive genome evolution in the organelle. While it has long been known that the genes retained most tenaciously by plastids and mitochondria encode for proteins involved in the electron transport chain of the bioenergetic organelle or for the ribosome required for their synthesis [12], only recently was it recognized that even within the ribosome, the same core of proteins has been retained independently by plastids and mitochondria, probably owing to constraints imposed by the process of ribosome assembly [13].

Endosymbiotic theory was also an important testing ground for molecular evolution. In the 1970s, there were competing theories to explain organelle origins. Those theories called for autogenous rather than symbiotic organelle origins and saw plastids and mitochondria as deriving from invaginations of the plasma membrane [14], from restructuring of thylakoids in a cyanobacterial ancestor of eukaryotes [15], or from budding of the nuclear membrane [16], as opposed to origins through symbiosis. They had it that the DNA in organelles stems from, and hence should be more similar in sequence to, genes encoded in nuclear DNA than to genes from free-living prokaryotes. That was a prediction that could be tested with DNA sequence comparisons. Bonen and Doolittle [17] found evidence for similarity between plastid and cyanobacterial nucleic acids, and Butow [18] found

evidence for mitochondrial genes that had been transferred to the nucleus in yeast. By about 1980, endogenous theories could be excluded and through 16S rRNA analyses, it was possible to confirm the origin of plastids from their suspected cyanobacterial ancestors [19] and to trace the origin of mitochondria to a metabolically versatile group of prokaryotes then called purple non-sulphur bacteria [20], later renamed to proteobacteria [21].

Protein import machineries as beacons for endosymbiotic events

Plastids and mitochondria each have a single origin. The strongest evidence for this comes from the protein import apparatus [22,23]. Had mitochondria become established in independent eukaryotic lineages, they would hardly have independently invented, via convergent evolution, the same core set of TIM and TOM components (translocon of the inner/outer mitochondrial membrane) that unite all mitochondria and organelles derived thereof [24•,25]. The same is true for the TIC and TOC systems (translocon of the inner/outer chloroplast membrane) of plastids [26,27]. The unity of these import machineries among mitochondria and plastids, respectively, is thus widely regarded as the best evidence we have for the single origin of these organelles, as opposed to multiple independent symbiotic origins in different lineages, even from endosymbionts so closely related as to be indistinguishable in phylogenies [28]. The establishment of a symbiotic cyanobacterium and its transition to the plastid ancestor is called primary symbiosis, it occurred perhaps some 1.2 billion years ago [29]. Subsequent to that, a number of secondary symbioses took place during evolution [30–32], in which eukaryotic algae became established as endosymbionts within eukaryotic cells, giving rise to what are called complex plastids, a term used to designate plastids surrounded by three or more membranes [33]. It is undisputed that secondary endosymbiosis occurred on at least three different occasions during eukaryote evolution: one in the lineage leading to the Euglenoids, a second independent event in the lineage leading to the Chlorarachniophytes and *at least* one more that led to the secondary plastids of red algal origin in diverse algal groups (Figure 1). For more than 20 years, the number and nature of secondary endosymbiotic events involving red algae has been heftily debated. Most of the debate has focussed on interpreting the differences between conflicting gene trees for the same groups [31,34,35,36•,37].

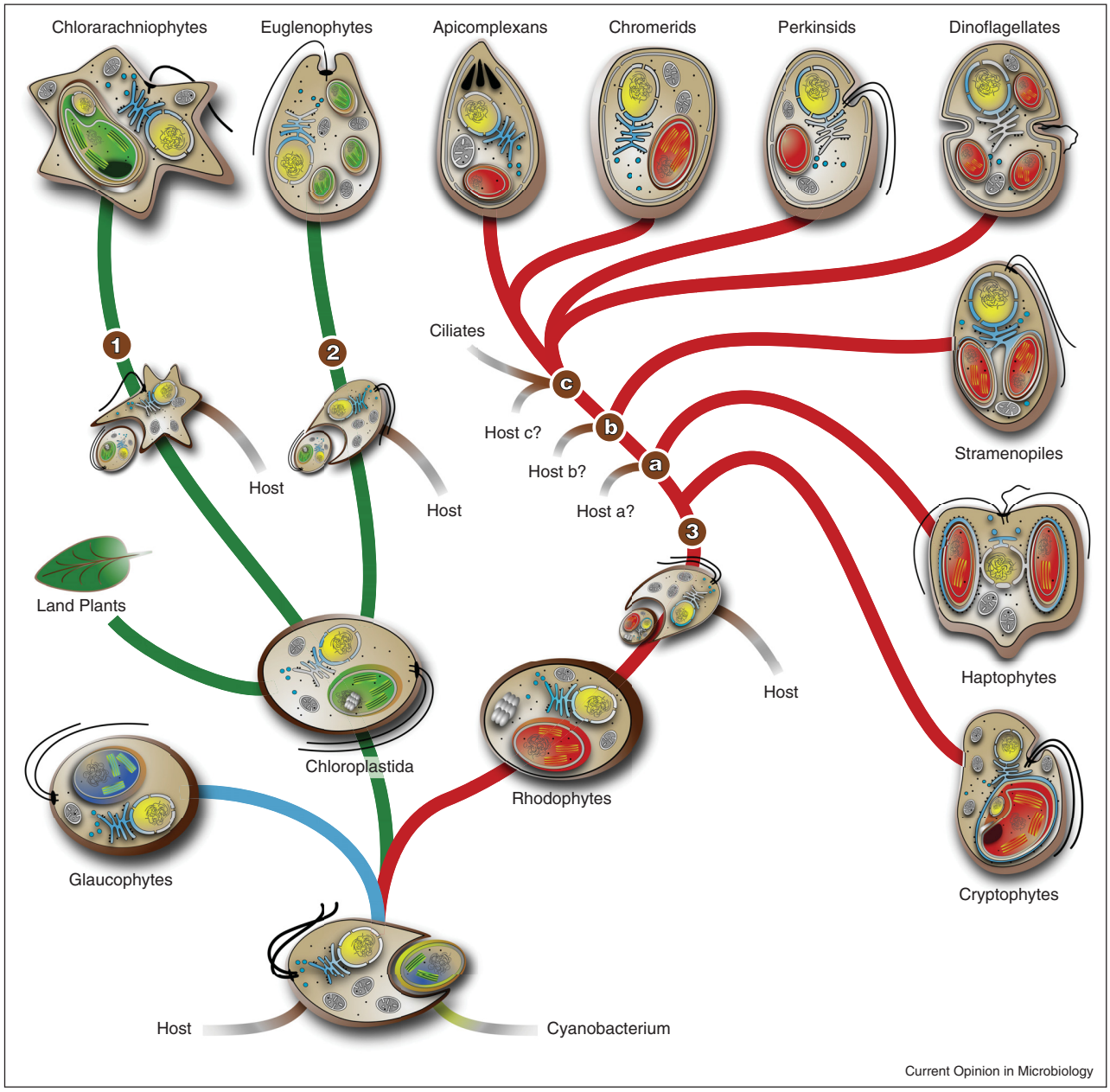
What if we step back from the trees and use the same reasoning and kind of data as the field uses to uncontentiously conclude that there was only one origin each of plastids and mitochondria? What if we look at the protein import machinery of red complex plastids of CASH lineages (Cryptophytes, Alveolates, Stramenopiles and Haptophytes)? Work in Uwe-G. Maier's group has shed light on the protein import machinery across the second outermost membrane of complex red plastids surrounded by four membranes [38,39]. That machinery is called SELMA (symbiont-specific ERAD-like machinery). SELMA is a multi-protein system that has been adopted from the symbiont's ERAD system (for endoplasmic reticulum (ER) associated degradation). In eukaryotic cells ERAD exports proteins from the ER for their degradation in the cytosol [40]. In *all* CASH plastids, a conserved N-terminal bipartite leader guides preproteins through the SELMA translocon across the second outermost membrane into the periplastidal compartment [39–42,43••].

For untangling red secondary symbioses, the crucial observation is that salient components of the SELMA are still encoded in the nucleomorph (the former nucleus of the engulfed red alga of cryptophytes; Figure 2) [38], and that protein import across the second outermost membrane of all CASH plastids involves a homologous SELMA machinery of monophyletic origin [42]. The SELMA machinery arose only once in evolution (like TIM/TOM and TIC/TOC), and it arose in the nucleus of the secondary endosymbiont that gave rise to the complex red plastid of cryptophytes (Figure 2). That tells us that all red secondary plastids are derived from the same algal endosymbiont that gave rise to cryptophyte plastids — and from that it follows that there was one single secondary endosymbiosis at the origin of the red secondary plastids (symbiosis 3 in Figure 1). So far so good, but in symbiosis it takes two to tango and a single origin of the red complex plastid still does not tell us how many hosts were involved. It could be that all CASH groups descend from the same endosymbiotic event as Cavalier-Smith suggested in the chromalveolate hypothesis [44]. Or they only share the same plastid, in which case one or more of the CASH lineages could have acquired plastids via tertiary symbiosis (like in the rhodoplex hypothesis [36•]) by engulfing a member of the ancient lineage that lead to cryptophytes (possible additional symbioses a–c in Figure 1). Should the plastid of cryptophytes also be of tertiary origin, then the secondary red alga that established SELMA has yet to be identified. Some might suggest that SELMA was passed around through lateral gene transfer (LGT), but considering its functional complexity (about a dozen or more proteins [36•]) that seems unlikely. Also note that chlorarachniophytes harbour a complex plastid still containing a nucleomorph, too, but it is of green algal origin and does not use a SELMA-like translocon [45]. Many conflicting gene trees addressing the issue of red secondary plastid origins have to be wrong, or misleading, or both.

How green are the reds, how red are the greens?

The origin of red secondary plastids highlights issues about trees and their interpretation. This can be illustrated with one recent study concerning diatoms, whose

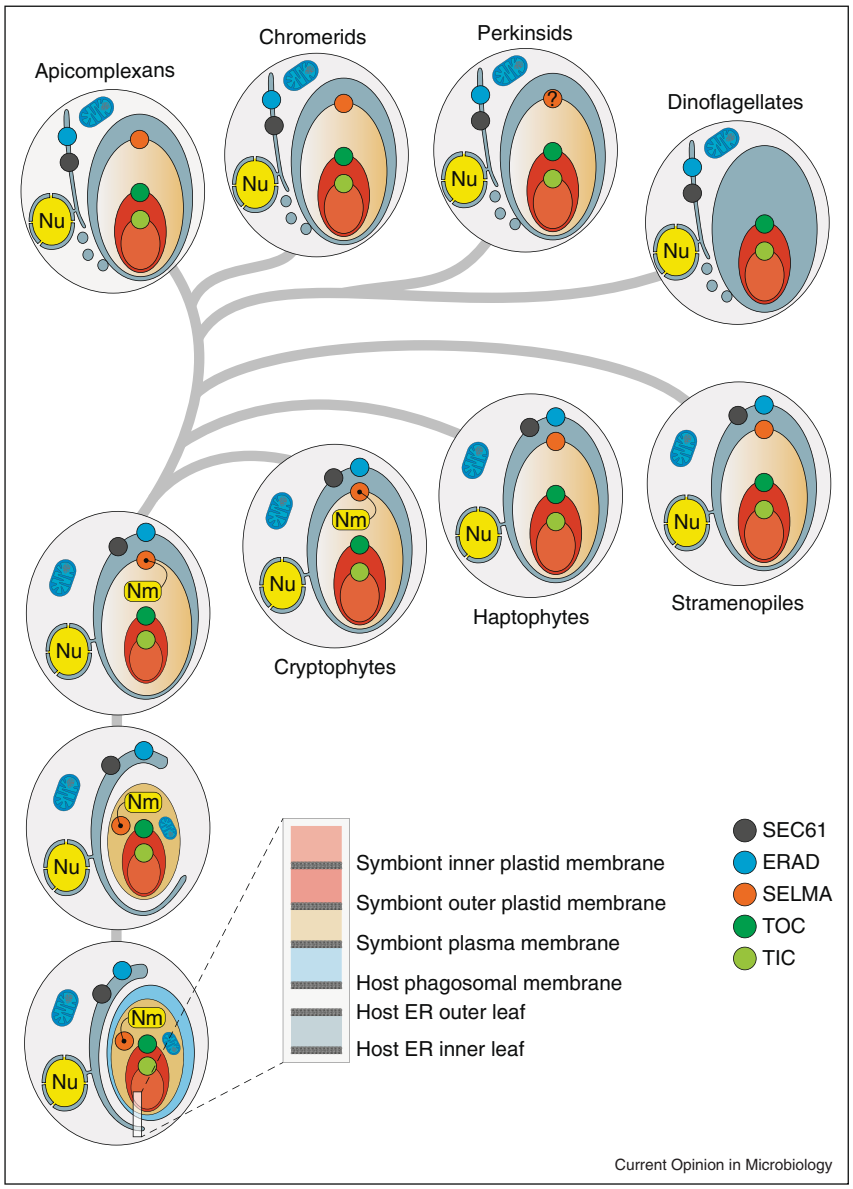
Figure 1



Plastid evolution. The initial uptake of a cyanobacterium by a heterotrophic host lead to three lineages: the Glaucophytes, Chloroplastida and Rhodophytes. Subsequently, two individual secondary endosymbiotic events involving algae of the Chloroplastida lineage and two heterotrophic hosts of unknown nature lead to the Chlorarachniophytes (symbiosis 1) and Euglenophytes (symbiosis 2). The radiation of secondary red plastids is not fully resolved, but the initial step was monophyletic, too (symbiosis 3) and connected to the origin of the SELMA translocon (see Figure 2 for details). While there is good evidence that the initial secondary plastid is of monophyletic origin, the amount of downstream-involved hosts remains uncertain (potential additional symbioses a–c). In some lineages red complex plastids could be of tertiary endosymbiotic origin. For details please refer to the text.
Modified from [30].

plastids unquestionably — based on plastid genome organization, not trees [46] — descend from red algae. Moustafa *et al.* [47] found that diatoms harbour many nuclear genes that branch with red algal homologues, as they should, *if* their plastids indeed are derived from the red lineage, which they are, and *if* many genes have been transferred from organelles to the nucleus during evolution, which has happened [48,49]. The problem is that

Figure 2



SELMA and the evolution of the CASH lineages. Schematic model for the evolution and radiation of SELMA among protists with complex red plastids. The red algal endosymbiont was initially encapsulated by a phagosomal membrane that separated it from the hosts' cytosol. That membrane was lost first, and after which a part of the hosts' endoplasmic reticulum wrapped around the endosymbiont (similar, but not identical to the 'autophagosome model' [102]). This step was accompanied by the loss of the endosymbionts' plasma membrane, mitochondrion and ER. The two eukaryotic cytosols fused and the nucleomorph (Nm)-encoded SELMA was now integrated into the inner face of the host ER membrane after the endosymbionts ER was lost. This process established the SELMA system, which is now found in all organisms with complex red plastids, but where it is now encoded in the nucleus (Nu), except for cryptophytes, where it remains Nm-encoded. Peridinin-containing dinoflagellates, whose plastids are surrounded by only three membranes, are the only exception: they appear to have lost the SELMA machinery altogether, when loosing an additional complex plastid membrane.

they found just as many diatom nuclear genes branching with green algae as with red. The same red versus green problem was observed in an independent study on *Chromera*, a photosynthetic relative of Apicomplexans [50]. And to complicate the matter, the same observation, but vice versa, was made in the genome of the chlorarachniophyte (Figure 1) *Bigelowiella natans* that houses an endosymbiont of green origin: of the 353 algal genes identified, 45 (22%) were found to branch with red algae [51••]. Hence, the results and the effects are reproducible. Some

will ask whether green plastids are frequently being replaced by red ones, and vice versa, during algal evolution, but maybe the first question we should ask is: Are trees simply fraught with systematic or random errors in such a way that diatoms end up on the green branch very often, when they really belong on the red branch [52]?

Is molecular phylogeny really that badly error prone? It well could be. In one study of a known phylogeny involving two grasses, a dicot, a gymnosperm, a liverwort and a red alga, only 40 out of 58 chloroplast encoded proteins (where there is no paralogy and no lateral gene transfer for the genes in question) recovered the true tree [53]. In a study of nine plastid genomes only 11 out of 42 genes recovered the consensus tree [54]. The simplest interpretation of such findings is that phylogeny is an imperfect art and that we should always expect some unexpected branches. The problem is that we do not know how many or which unexpected branches to expect. But the more ancient the phylogeny and the more species in the tree, the more we should expect to see spurious branches. In theory, for a tree with 38 leaves (taxa), there are roughly 10^{51} possible trees: the chances of getting the right one are the same as picking the same proton out of all the protons on Earth (6×10^{50}) twice in a row. So if we see a tree with three-dozen leaves, it is possible that many branches are wrong, we just don't know which ones are wrong or how wrong they are. Even the branches with strong bootstrap or other support values can be wrong, because support values just tell us how often the algorithm and the data produce the branch in the computer, not whether the model or the branch is correct [55]. And when the trees contain prokaryotic leaves, the problems get worse, because of LGT among prokaryotes [56].

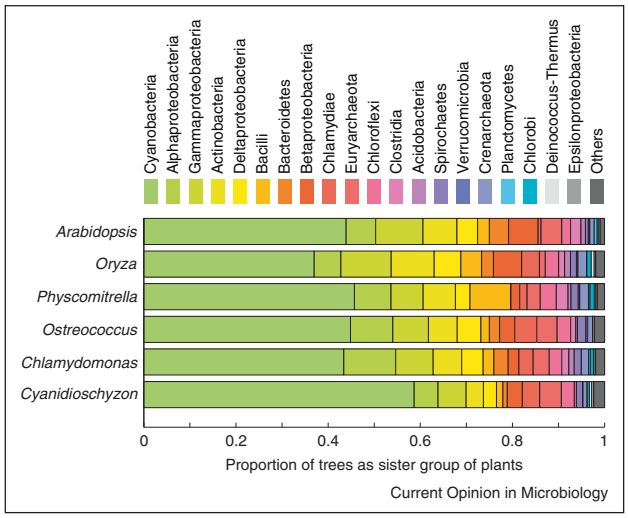
Of course, the alternative to assuming that phylogenetic trees are inherently imperfect is to assume that they are telling us the true course of history, just the way it was, every branch in every tree reflecting some past event, whose existence can be inferred because of some edge (the mathematical term for branch) that a computer produces. This is a good place to recall that plastids and mitochondria are biological entities in nature, things that we can observe and whose origins require an evolutionary explanation. By contrast, branches in phylogenetic trees are not observations of things in nature, they are things that computers generate when instructed by humans to produce them from input data — whether or not branches in phylogenetic trees require any explanation at all is debateable. Trees and branches are most effective when we use them as tools to test theories about evolution rather than as crayons to draw evolutionary history from scratch. The problem is that each tree tells a different story and if we can believe one tree all the others must be wrong, which can lead to exhausting debates of which gene tree is telling the true story, or if we look at the matter openly, whether *any* gene tree is telling the true story. Two recent developments concerning the use of gene trees in endosymbiotic theory, and the interpretation of those trees, underscore that point.

How much help did a cyanobacterium have becoming an endosymbiont?

In the genome sequence of *Chlamydia trachomatis* some genes were found that shared unexpectedly close phylogenetic relationships with plant homologs [57]. These unexpected branches were met with an array of explanations including direct LGT from eukaryotes to chlamydiae [57,58], or LGT in the other direction [59,60], indirect LGT to archaeplastids through the cyanobacterial endosymbiont [61], unrevealed relationship between archaeplastids and amoebas [54] or between cyanobacteria and chlamydiae [62], and gene transfer from mitochondria followed by differential loss [63]. Subsequent phylogenetic studies revealed a few more examples, and it was stated in its most recent formulation that '*Chlamydia-like pathogens are the second major source of foreign genes in Archaeplastida*' [64], and that the cyanobacterial origin of plastids was a symbiosis of three partners, with chlamydiae in an essential role of mediating metabolic integration of those partners [65–68].

The problem is not that modern cyanobacterial (endo)symbioses observable in nature (lichens, cycads, *Azolla*, *Gunnera* or *Rhopalodia*) get by with just the cyanobacterium alone, with no aid from chlamydias, spirochaetes, or any other helper bacteria. The problem is also not that the benefit afforded to the host in those cyanobacterial symbioses is fixed nitrogen, not carbohydrate [69,70]. The problem is that when we look at all the trees that include prokaryotic lineages, chlamydiae no longer stand out [71]. Not much attention is paid to the overall potential gene origin in studies focusing on chlamydiae and plants alone [66,67,72]. If we apply the rationale of the chlamydial-helper hypothesis to genes apparently stemming from other prokaryotes, the endosymbiont hypothesis for plastids would be one involving many more 'helper' prokaryotes. Moreover, the '*second major*' [64], and we stress, apparent '*source*' of prokaryotic genes in plants is not chlamydia, it is alphaproteobacteria, followed by gammaproteobacteria, then actinobacteria, deltaproteobacteria, bacilli, bacteroidetes, and betaproteobacteria, behind which chlamydiae range as another meagre apparent donor (Figure 3) [71,48]. Did all of these lineages, and the lesser apparent donors, such as euryarchaeotes, clostridias, spirochaetes, planctomycetes and chlorobia help the cyanobacterium to become established as an endosymbiont or plastid? That should be the conclusion, if one takes the trees at face value. Furthermore, the genes in the different apparent donor lineage trees do not even branch with the same chlamydia, or the same proteobacteria, or for that matter of fact the same cyanobacteria. In the end, the single gene

Figure 3



Apparent prokaryotic donors of genes to plant lineages. Genes of many major prokaryotic lineages appear as nearest neighbours to archaeplastid nuclear genes in phylogenetic trees. Note that the apparent contribution of chlamydias is smaller than that of lineages such as actinobacteria, bacilli or bacteroidetes.
The figure is reproduced with permission from [71].

trees in which plants branch with cyanobacteria tell us that plastids arose from 60 or more different cyanobacteria [71]. Could that be?

An alternative would be to consider factors that are too often overlooked in studies of eukaryote gene origins in the context of organelle origins: random phylogenetic errors, limited taxon sampling, individual gene losses and LGT among prokaryotes [71,73–77]. Even if in an analysis the phylogenetic inference is completely correct and homologs from all extant organisms are included, LGT and gene losses in prokaryotes alone could still have produced the observed patterns [71,74–76,78,79]. In fact, LGT among prokaryotes is even evident in the trees in studies suggesting direct LGT (e.g. [80]), where the prokaryotic sister group of the eukaryotic clade is formed by homologs from more than one prokaryotic lineage, an observation that would not be possible had the gene never been transferred among prokaryotes. Even if the true donor was a cyanobacterium and gene phylogeny was error-free, loss of this gene or its absence from our limited sample of cyanobacteria and its transfer among prokaryotes since the origin of plastids could easily produce the pattern of apparent LGT from non-cyanobacterial sources.

Because of the single origin of plastids, the cyanobacterial ancestor of plastids was a unique prokaryotic organism. But as such, it had a pan-genome [81••]. What was the composition of its specific genome of the symbiont within that cyanobacterial pan-genome at the time of symbiosis? The best estimate probably comes from analysis of a frozen accident: the genes that plants acquired at the origin of plastids and that have persisted to the present in plant genomes. An analysis of 51 modern cyanobacterial genomes reveals 18 000 cyanobacterial gene families and 47 000 singletons [71], or a cyanobacterial pan-genome encompassing some 65 000 genes, whereby only about 5000 are found in any one cyanobacterium. Similarly, 61 strains of *Escherichia coli* have a pan-genome of about 18 000 genes, whereby only about 4500 are packaged in any given cell and only about 1000 genes (about 20% of the genome) are common to all *E. coli* strains within the species [82]. Thus, were an *E. coli* strain to become an endosymbiont today with the fate of turning into an organelle in a billion years, only about 20% of its genome would be defining for *E. coli* at the time of symbiosis, and the remainder would be shared with free-living *E. coli* strains, which would be free to generate new combinations of genes within and among species for the next billion years. In a billion years, the collection of genes that we call *E. coli* will no longer exist as an *E. coli* species complex, but most of the genes will still be around as descendant copies somewhere, just distributed among various genomes that would not be called *E. coli*. We do not know what happened a billion years and more ago, but we should keep in mind that, firstly, the genomes of the symbionts were already chimaeras; secondly, the descendants of the free-living relatives continued to experience LGT with other prokaryotes; and thirdly, phylogenetic tools are far from perfect.

An autogenous, ATP-consuming origin of mitochondria?

Another development that has unfolded around endosymbiosis could be called an issue of lumping and splitting. It centres around the origin of mitochondria. A good bit of progress has been made in understanding the role of mitochondria in eukaryote evolution in recent years. First, all eukaryote lineages are now known either to have or to have had a mitochondrion in their past [83••]. Second, the host that acquired the mitochondrion stems from a lineage that branches within the archaebacteria (or archaea), not as their sister [84••,85,86•]. Third, the presence of internalized bioenergetic membranes was the key attribute provided by mitochondrial endosymbiosis, which afforded eukaryotes many orders of magnitude more energy per gene than is available to prokaryotes [87]. Thus, while it has now been evident for some time that the common ancestor of eukaryotes possessed a mitochondrion, it is now clear why that was so: the lack of true intermediates in the prokaryote-to-eukaryote transition has a bioenergetic cause [87].

But beyond that, the origin of mitochondria is debated. Different phylogenomic analyses come to different

results regarding the nature of the free-living bacteria that are the closest relatives of mitochondria. Recent studies focussing on genes located in mitochondrial DNA, which is very AT-rich and thus prone to associate mitochondria, phylogenetically, to AT-rich proteobacteria, disagree with respect to the relationship of mitochondria to clades of free-living prokaryotes [88,89]. Different genes in mitochondrial DNA appear to trace to different sources in phylogenetic studies [90–92], as do different eukaryotic nuclear genes associated with mitochondrial functions [76,93,94]. Like in the case of plastids discussed above, such differences have causes that involve phylogenetic reconstruction, pan-genomes, and gene transfer among prokaryotes themselves [95], the relative contributions of which have however yet to be resolved. Amidst those debates, a careful and detailed survey of bioenergetic pathways and the diversity among components of the membrane-associated electron transport chain in free-living proteobacteria points to methylotrophic ancestors for mitochondria [96••], which is particularly interesting as the methylotrophs are metabolically versatile prokaryotes and have invaginations of their plasma membrane that rival the ultrastructural complexity of mitochondrial cristae [97].

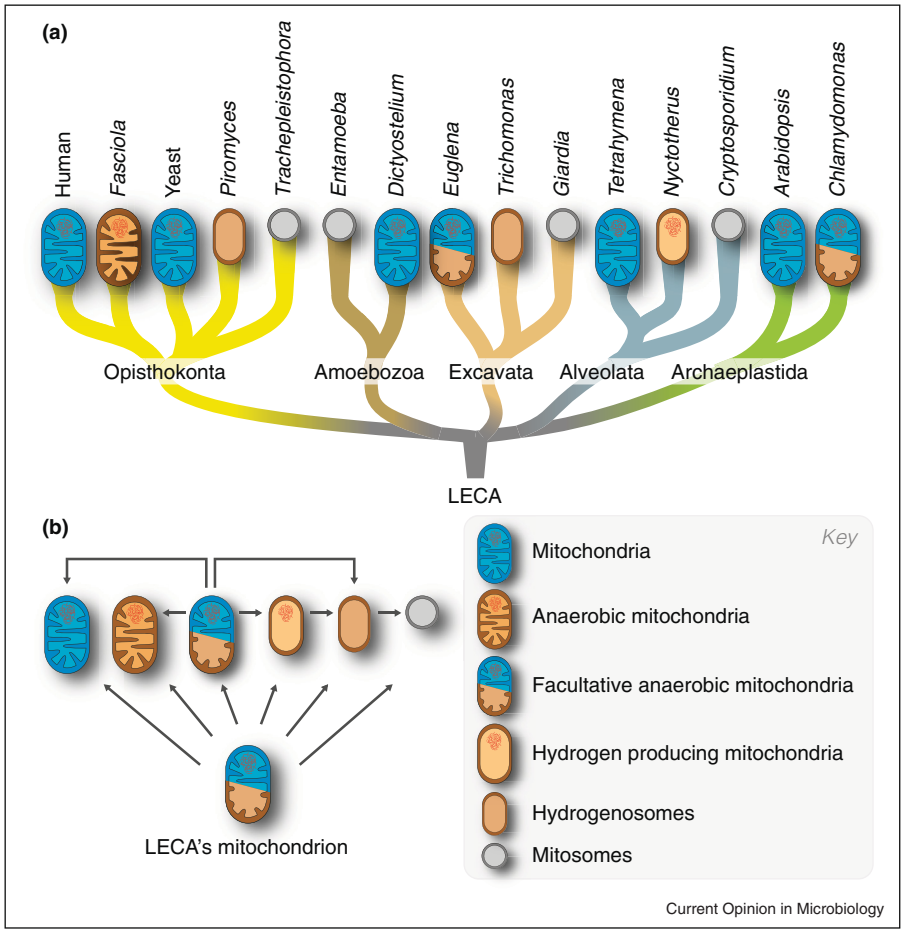
Some people still think that the main advantage of mitochondria and the key to eukaryote complexity was a roughly sixfold increase in energy yield from glucose. Indeed, with O_2-respiring mitochondria eukaryotes can harvest about 32 mol ATP per glucose, while with anaerobic mitochondria they can only glean about 5 mol ATP per glucose, and with hydrogenosomes they only harness about 4 mol ATP per glucose [98•]. But O_2 respiration cannot be the key to eukaryote complexity, for were that true, then *E. coli* and all other (facultative) aerobic prokaryotes should have become just as complex as eukaryotes, for the same reason of improved aerobic energy yield from glucose. The different manifestations of mitochondria in eukaryotes — aerobic, facultatively anaerobic, anaerobic, hydrogenosomes and mitosomes — have arisen independently as ecological specializations in different eukaryotic lineages (Figure 4), but essentially all of the genes involved in ATP production in organelles of mitochondrial origin were present in the eukaryote common ancestor [98•]. A competing alternative that the genes for anaerobic energy metabolism in eukaryotes were acquired late in eukaryotic evolution from donors that were distinct from the mitochondrion and then passed around from one eukaryote to another is favoured by some researchers [99,100], but the theory only accounts for sparse distributions of genes, which is just as simply accounted for by differential loss.

In search of one sentence on mitochondrial origin with which all prospective readers of this paper could agree, one could have recently risked: Mitochondria are organelles derived from a symbiosis between a bacterium that became the mitochondrion and a host. Yet, that formulation would not agree with the most recent view of mitochondrial evolution by Gray [101•], who was once a strong proponent of the endosymbiotic theory, but who now argues that the mitochondrial compartment was present before the organism that we call the mitochondrial endosymbiont entered the cell. His argument is that only comparatively few genes for mitochondrial proteins — 10–20% in his estimate — tend to reflect an alphaproteobacterial ancestry in single gene phylogenetic trees. The rest do not, they branch elsewhere among prokaryotic or eukaryotic homologues. From that he infers that only the genes that branch with alphaproteobacterial homologues come from endosymbiosis, while the remainder, the majority of genes whose products function in mitochondria today, were already present before the alphaproteobacterial symbiosis in an autogenously originated compartment: the pre-mitochondrion, which is envisaged as an ancestrally ATP consuming compartment. Its proteinaceous contents were specifically retargeted to the alphaproteobacterial invader, transforming it into a mitochondrion.

Gray's hypothesis, called the pre-endosymbiont hypothesis [101•], is not designed to explain the origin of mitochondria, it is designed to explain the origin of the many mitochondrial proteins that do not branch with alphaproteobacterial homologues. That is, it is designed to explain branching patterns in individual gene trees, which, as we saw in the case of chlamydiae, can be more complicated than it would seem at first glance. Like the chlamydial-helper hypothesis, the pre-endosymbiont hypothesis divides the world into, in this case, mitochondrial proteins whose trees branch with a particular group (chlamydiae, alphaproteobacteria) and those that do not. A disconcerting aspect of the theory is that it arbitrarily lumps and splits: it splits off into one bin all the mitochondrial proteins that branch with present-day alphaproteobacterial homologues and lumps together into a second bin all the ones that do not. While the former are assumed to come from the alphaproteobacterial symbiont, the origin of the latter is not addressed, they are just assumed to be present in the cell that acquired a few alphaproteobacterial genes.

The kind of transition between the pre-mitochondrion (not derived from proteobacteria) and the mitochondrion (derived from an alphaproteobacterium) that Gray envisages entails several *ad hoc* components, such as precise retargeting of all the proteins that a mitochondrion needs from the pre-mitochondrion to the mitochondrion. During the origin of plastids, the plant mitochondrion, which had its protein import apparatus in place, did not become transformed so as to become green and photosynthetic, the two compartments remained distinct, rather than showing a tendency to merge, and the plastid ended up having its own import machinery, which arose

Figure 4



Mitochondria and related organelles all have a single origin. **(a)** Different types of mitochondria-related organelles (e.g. mitosomes or hydrogenosomes) can be found in different taxa of all eukaryotic super groups, such as the Amoebozoa and the Alveolata. **(b)** The last eukaryotic common ancestor (LECA) contained a 'universal' facultative anaerobic mitochondrion of alphaproteobacterial origin and the different types of mitochondria-related organelles evolved subsequently from the common ancestor, and depending on the ecological niche the host colonized.

independently of that in mitochondria. Gray's theory is an excellent example of a thoughtful theory that is designed to explain unexpected branches in trees, but not to explain the similarity of mitochondria to bacteria. As Gray [101•] points out, it has quite a lot in common with autogenous theories for the origins of organelles, which were also not designed to explain the similarity of mitochondria to bacteria, rather they were designed to explain the presence of DNA in plastids and mitochondria [14–16].

Conclusion

Endosymbiotic theory for the origin of organelles is still by far the best tool we have to explain why chloroplasts and mitochondria are so similar to free living bacteria. Alternatives to endosymbiotic theory often share several important, but unstated assumptions: they start with the premise that endosymbiotic theory somewhere stated or predicted that *all* genes that the plant lineage acquired from cyanobacteria need to branch with present-day cyanobacterial homologues, and that *all* genes that eukaryotes acquired from mitochondria need to branch with present-day purple non-sulphur bacterial (or alphaproteobacterial) homologues in phylogenetic trees. Using that lever, one can pry loose a corollary: all genes that do not fulfil those criteria were acquired from other sources. The ensuing procedure for identifying the donor is then simple: we assume that the prokaryotic homologue and the prokaryotic rRNA gene (the basis of naming prokaryotic groups) of the genome within which the homologue of the eukaryotic gene resides, have remained linked — within the same chromosome — from the time that the gene was donated (for plastid and mitochondrial origins, about a third of Earth's history ago) until the present, and we assume that the procedure of inferring gene phylogeny is error-free. Using such assumptions, whether

explicitly stated or not, one can infer that a gene X was donated by organism Y. OK, but to be fair then the same logic needs to apply to all genes, in which case the practice of inferring gene origins directly from trees quickly turns into an affair of one endosymbiont per gene and, if we think it through in full, we would end up assuming that all prokaryotic genes having eukaryotic homologues have remained resident in the same prokaryotic chromosome together with their species name-giving rRNA for the last 1–2 billion years. Now recall that endosymbiotic theory is a lot older than the practise of building gene trees. Alternatively, endosymbiotic theory is fine but it needs to be better integrated into a modern world of microbial genomics, one where we know that the pan-genomes of prokaryotic species are much larger than any individual's genome, and where lateral gene transfer is known to transport genes across chromosomes with little respect for species (or other taxonomic) borders. In summary, we probably need to keep our expectations more relaxed when it comes to the phylogenetic behaviour of genes that eukaryotes acquired from plastids and mitochondria. If we do that, endosymbiotic theory explains a lot as it is.

Acknowledgements

We thank Andrzej Bodyl and John W. Stiller for their constructive comments on the manuscript. This work was funded by the European Research Council (grant no. 232975 to WFM) and by the Deutsche Forschungsgemeinschaft (grant no. GO1825/4-1 to SBG). CK thanks the Deutscher Akademischer Austauschdienst for a PhD stipend.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Mereschkowsky C: **Über Natur und Ursprung der Chromatophoren im Pflanzenreiche**. *Biol Centralbl* 1905, **25**:593-604 (English translation in Martin W, Kowallik KV, *Eur J Phycol* 1999, **34**:287–295).

2. Wallin IE: *Symbionticism and the Origin of Species*. London: Bailliere, Tindall and Cox; 1927, 171.

3. Sapp J: **The dynamics of symbiosis: an historical overview**. *Can J Bot* 2004, **82**:1046-1056.

4. Schnepf E: **Zur Feinstruktur von *Geosiphon pyriforme.* Ein Versuch zur Deutung cytoplasmatischer Membranen und Kompartimente**. *Arch Mikrobiol* 1964, **49**:112-131.

5. John P, Whatley FR: ***Paracoccus denitrificans* and the evolutionary origin of the mitochondrion**. *Nature* 1975, **254**:495-498.

6. Glöckner G, Rosenthal A, Valentin K: **The structure and gene repertoire of an ancient red algal plastid genome**. *J Mol Evol* 2000, **51**:382-390.

7. Burger G, Gray MW, Forget L, Lang BF: **Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists**. *Genome Biol Evol* 2013, **5**:418-438.

8. Kleine T, Maier UG, Leister D: **DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis**. *Annu Rev Plant Biol* 2009, **60**:115-138.

9. Meisinger C, Sickmann A, Pfanner N: **The mitochondrial proteome: from inventory to function**. *Cell* 2008, **134**:22-24.

10. Martin W, Herrmann RG: **Gene transfer from organelles to the nucleus: how much, what happens, and why?** *Plant Physiol* 1998, **118**:9-17.

11. Allen JF: **Control of gene-expression by redox potential and the requirement for chloroplast and mitochondrial genomes**. *J Theor Biol* 1993, **165**:609-631.

12. Allen JF: **The function of genomes in bioenergetic organelles**. *Phil Trans R Soc Lond B: Biol Sci* 2003, **358**:19-37.

13. Maier UG, Zauner S, Woehle C, Bolte K, Hempel F, Allen JF, Martin WF: **Massively convergent evolution for ribosomal protein gene content in plastid and mitochondrial genomes**. *Genome Biol Evol* 2013, **5**:2318-2329.

14. Bogorad L: **Evolution of organelles and eukaryotic genomes**. *Science* 1975, **188**:891-898.

15. Cavalier-Smith T: **The origin of nuclei and of eukaryotic cells**. *Nature* 1975, **256**:463-468.

16. Raff RA, Mahler HR: **The non symbiotic origin of mitochondria**. *Science* 1972, **177**:575-582.

17. Bonen L, Doolittle WF: **Prokaryotic nature of red algal chloroplasts**. *Proc Natl Acad Sci U S A* 1975, **72**:2310-2314.

18. Farrely F, Butow RA: **Rearranged mitochondrial genes in the yeast nuclear genome**. *Nature* 1983, **301**:296-301.

19. Giovannoni S, Turner S, Olsen G, Barns S, Lane D, Pace N: **Evolutionary relationships among cyanobacteria and green chloroplasts**. *J Bacteriol* 1988, **170**:3584-3592.

20. Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR: **Mitochondrial origins**. *Proc Natl Acad Sci U S A* 1985, **82**:4443-4447.

21. Stackebrandt E, Murray RGE, Trüper HG: ***Proteobacteria* classis nov., a name for the phylogenetic taxon that includes the "purple bacteria and their relatives"**. *Int J Syst Bacteriol* 1988, **38**:321-325.

22. Dolezal P, Likic V, Tachezy J, Lithgow T: **Evolution of the molecular machines for protein import into mitochondria**. *Science* 2006, **313**:314-318.

23. McFadden GI, van Dooren GG: **Evolution: red algal genome affirms a common origin of all plastids**. *Curr Biol* 2004, **14**:R514-R516.

24. • Zarsky V, Tachezy J, Dolezal P: **Tom40 is likely common to all mitochondria[SINGLE]**. *Curr Biol* 2012, **22**:R479-R481.
Proposes that one TOM40 unites all extant eukaryotic mitochondria and organelles of mitochondrial origin.

25. Shiflett AM, Johnson PJ: **Mitochondrion-related organelles in eukaryotic protists**. *Annu Rev Microbiol* 2010, **64**:409-429.

26. Bullmann L, Haarmann R, Mirus O, Bredemeier R, Hempel F, Maier UG, Schleiff E: **Filling the gap, evolutionarily conserved Omp85 in plastids of chromalveolates**. *J Biol Chem* 2010, **285**:6848-6856.

27. Shi LX, Theg SM: **The chloroplast protein import system: from algae to trees**. *Biochim Biophys Acta* 2013, **1833**:314-331.

28. Howe CJ, Barbrook AC, Nisbet RER, Lockhart PJ, Larkum AWD: **The origin of plastids**. *Philos Trans R Soc Lond B: Biol Sci* 2008, **363**:2678-2685.

29. Parfrey LW, Lahr DJG, Knoll AH, Katz LA: **Estimating the timing of early eukaryotic diversification with multigene molecular clocks**. *Proc Natl Acad Sci U S A* 2011, **108**:13624-13629.

30. Gould SB, Waller RF, McFadden GI: **Plastid evolution**. *Annu Rev Plant Biol* 2008, **59**:491-517.

31. Lane CE, Archibald JM: **The eukaryotic tree of life: endosymbiosis takes its TOL**. *Trends Ecol Evol* 2008, **23**:268-275.

32. Kowallik KV: **Evolution durch genomische Kombination**. In *Gott oder Darwin.* Edited by Klose J, Oehler J. Berlin, Germany: Springer Verlag; 2008:141-157.

33. Gibbs SP: **The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae**. *Ann N Y Acad Sci* 1981, **361**:193-208.

34. Delwiche CF: **Tracing the thread of plastid diversity through the tapestry of life**. *Am Nat* 1999, **154**:S164-S177.

35. Baurain D, Brinkmann H, Petersen J, Rodriguez-Ezpeleta N, Stechmann A, Demoulin V, Roger AJ, Burger G, Lang BF, Philippe H: **Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles**. *Mol Biol Evol* 2010, **27**:1698-1709.

36. • Petersen J, Ludewig AK, Michael V, Bunk B, Jarek M, Baurain D, Brinkmann H: ***Chromera velia*, endosymbioses and the rhodoplex hypothesis–plastid evolution in cryptophytes, alveolates, stramenopiles, and haptophytes (CASH lineages)[SINGLE]**. *Genome Biol Evol* 2014, **6**:666-684.
A treatise of competing hypotheses for the evolution of red complex plastids. Proposes the 'rhodoplex hypothesis' as a possible alternative to the chromalveolate hypothesis.

37. Bodyl A, Stiller JW, Mackiewicz P: **Chromalveolate plastids: direct descent or multiple endosymbioses?** *Trends Ecol Evol* 2009, **24**:119-121.

38. Sommer MS, Gould SB, Lehmann P, Gruber A, Przyborski JM, Maier UG: **Der1-mediated pre-protein import into the periplastid compartment of chromalveolates?** *Mol Biol Evol* 2007, **24**:918-928.

39. Bolte K, Bullmann L, Hempel F, Bozarth A, Zauner S, Maier UG: **Protein targeting into secondary plastids**. *J Eukaryot Microbiol* 2009, **56**:9-15.

40. Smith MH, Ploegh HL, Weissman JS: **Road to ruin: targeting proteins for degradation in the endoplasmic reticulum**. *Science* 2011, **334**:1086-1090.

41. Gould SB: **Ariadne's thread: Guiding a precursor protein across five membranes in a cryptophyte**. *J Phycol* 2008, **44**:23-26.

42. Felsner G, Sommer MS, Gruenheit N, Hempel F, Moog D, Zauner S, Martin W, Maier UG: **ERAD components in organisms with complex red plastids suggest recruitment of a preexisting protein transport pathway for the periplastid membrane**. *Genome Biol Evol* 2011, **3**:140-150.

43. •• Stork S, Moog D, Przyborski JM, Wilhelmi I, Zauner S, Maier UG: **Distribution of the SELMA translocon in secondary plastids of red algal origin and predicted uncoupling of ubiquitin-dependent translocation from degradation[DOUBLE]**. *Eukaryot Cell* 2012, **11**:1472-1481.
The most comprehensive phylogenetic analysis of the SELMA components, with incisive perspectives on how the function evolved within its constituent components.

44. Cavalier-Smith T: **A six kingdom classification and a unified phylogeny**. In *Endocytobiology II.* Edited by Schwemmler W, Schenk HEA. Berlin, Germany: De Gruyter; 1983:1027-1034.

45. Hirakawa Y, Burki F, Keeling PJ: **Genome-based reconstruction of the protein import machinery in the secondary plastid of a chlorarachniophyte alga**. *Eukaryot Cell* 2012, **11**:324-333.

46. Stöbe B, Kowallik KV: **Gene-cluster analysis in chloroplast genomics**. *Trends Genet* 1999, **9**:344-347.

47. Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D: **Genomic footprints of a cryptic plastid endosymbiosis in diatoms**. *Science* 2009, **324**:1724-1726.

48. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D: **Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus**. *Proc Natl Acad Sci U S A* 2002, **99**:12246-12251.

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

50. Woehle C, Dagan T, Martin W, Gould SB: **Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related *Chromera velia***. *Genome Biol Evol* 2011, **3**:1220-1230.

51. •• Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH, Hirakawa Y *et al.*: **Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs[DOUBLE]**. *Nature* 2012, **492**:59-65.
Reports the genome sequences of two organisms harbouring complex plastids and a nucleomorph, elegantly comparing the independent evolution of complex cell types in both lineages. Suggests an evolutionary rationale behind the retention of nucleomorphs.

52. Dagan T, Martin W: **Microbiology. Seeing green and red in diatom genomes**. *Science* 2009, **324**:1651-1652.

53. Goremykin VV, Hansmann S, Martin WF: **Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast genomes: Revised molecular estimates of two seed plant divergence times**. *Plant Syst Evol* 1997, **206**:337-351.

54. Martin W, Stoebe B, Goremykin V, Hapsmann S, Hasegawa M, Kowallik KV: **Gene transfer to the nucleus and the evolution of chloroplasts**. *Nature* 1998, **393**:162-165.

55. Lockhart PJ, Howe CJ, Barbrook AC, Larkum AWD, Penny D: **Spectral analysis, systematic bias, and the evolution of chloroplasts**. *Mol Biol Evol* 1999, **16**:573-576.

56. Dagan T, Martin W: **Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution**. *Proc Natl Acad Sci U S A* 2007, **104**:870-875.

57. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q *et al.*: **Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis***. *Science* 1998, **282**:754-759.

58. Linka N, Hurka H, Lang BF, Burger G, Winkler HH, Stamme C, Urbany C, Seil I, Kusch J, Neuhaus HE: **Phylogenetic relationships of non-mitochondrial nucleotide transport proteins in bacteria and eukaryotes**. *Gene* 2003, **306**:27-35.

59. Greub G, Raoult D: **History of the ADP/ATP-translocase-encoding gene, a parasitism gene transferred from a Chlamydiales ancestor to plants 1 billion years ago**. *Appl Environ Microbiol* 2003, **69**:5530-5535.

60. Royo J, Gímez E, Hueros G: **CMP-KDO synthetase: a plant gene borrowed from Gram-negative eubacteria**. *Trends Genet* 2000, **16**:432-433.

61. Schmitz-Esser S, Linka N, Collingro A, Beier CL, Neuhaus HE, Wagner M, Horn M: **ATP/ADP translocases: a common feature of obligate intracellular amoebal symbionts related to chlamydiae and rickettsiae**. *J Bacteriol* 2004, **186**:683-691.

62. Brinkman FSL, Blanchard JL, Cherkasov A, Av-Gay Y, Brunham RC, Fernandez RC, Finlay BB, Otto SP, Ouellette BFF, Keeling PJ *et al.*: **Evidence that plant-like genes in *Chlamydia* species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast**. *Genome Res* 2002, **12**:1159-1167.

63. Amiri H, Karlberg O, Andersson SGE: **Deep origin of plastid/parasite ATP/ADP translocases**. *J Mol Evol* 2003, **56**:137-150.

64. Facchinelli F, Colleoni C, Ball SG, Weber AP: **Chlamydia, cyanobiont, or host: who was on top in the ménage à trois?** *Trends Plant Sci* 2013, **18**:673-679.

65. Subtil A, Collingro A, Horn M: **Tracing the primordial Chlamydiae: extinct parasites of plants?** *Trends Plant Sci* 2014, **19**:36-43.

66. Huang J, Gogarten JP: **Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids?** *Genome Biol* 2007, **8**:R99.

67. Moustafa A, Reyes-Prieto A, Bhattacharya D: **Chlamydiae has contributed at least 55 genes to plantae with predominantly plastid functions**. *PLoS ONE* 2008, **3**:e2205.

68. Cencil U, Nitschke F, Steup M, Minassian BA, Colleoni C, Ball SG: **Transition from glycogen to starch metabolism in Archaeplastida**. *Trends Plant Sci* 2014, **19**:18-28.

69. Kneip C, Lockhart P, Voss C, Maier UG: **Nitrogen fixation in eukaryotes — new models for symbiosis**. *BMC Evol Biol* 2007, **7**:55.

70. Raven JA: **Evolution of cyanobacterial symbioses**. In *Cyanobacteria in Symbiosis.* Edited by Rai AN, Bergman B, Rasmussen U. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2002:326-346.

71. Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major P, Gould SB, Goremykin VV, Rippka R, de Marsac NT *et al.*: **Genomes of stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids**. *Genome Biol Evol* 2013, **5**:31-44.

72. Becker B, Hoef-Emden K, Melkonian M: **Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes**. *BMC Evol Biol* 2008, **8**:203.

73. Lange BM, Rujan T, Martin W, Croteau R: **Isoprenoid biosynthesis: The evolution of two ancient and distinct pathways across genomes**. *Proc Natl Acad Sci U S A* 2000, **97**:13172-13177.

74. Rujan T, Martin W: **How many genes in *Arabidopsis* come from cyanobacteria? An estimate from 386 protein phylogenies**. *Trends Genet* 2001, **17**:113-120.

75. Martin WF, Roettger M, Kloesges T, Thiergart T, Woehle C, Gould SB, Dagan T: **Modern endosymbiotic theory: getting lateral gene transfer into the equation**. *J Endocyt Cell Res* 2012, **23**:1-5.

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

77. Stiller JW: **Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer**. *BMC Evol Biol* 2011, **11**:259.

78. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation**. *Nature* 2000, **405**:299-304.

79. Wolf YI, Koonin EV: **Genome reduction as the dominant mode of evolution**. *Bioessays* 2013, **35**:829-837.

80. Suzuki K, Miyagishima S: **Eukaryotic and eubacterial contributions to the establishment of plastid proteome estimated by large-scale phylogenetic analyses**. *Mol Biol Evol* 2010, **27**:581-590.

81. •• Beck C, Knoop H, Axmann IM, Steuer R: **The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms[DOUBLE]**. *BMC Genomics* 2012, **13**:56.

Reports on the nature of the cyanobacterial pangenome showing that the number of newly identified cyanobacterial genes continues to increase with each new cyanobacterial genome sequenced.

82. Lukjancenko O, Wassenaar TM, Ussery DW: **Comparison of 61 sequenced *Escherichia coli* genomes**. *Microb Ecol* 2010, **60**:708-720.

83. •• McInerney JO, O'Connell M, Pisani D: **The hybrid nature of the eukaryota and a consilient view of life on Earth[DOUBLE]**. *Nat Rev Microbiol* 2014, **12**:449-455.

An insightful perspective on the chimaeric nature of eukaryotes as true genomic hybrids of an archaeal host and a bacterial symbiont: the ancestor of mitochondria.

84. •• Williams TA, Foster PG, Cox CJ, Embley TM: **An archaeal origin of eukaryotes supports only two primary domains of life[DOUBLE]**. *Nature* 2013, **504**:231-236.

An incisive overview of current theories for eukaryote origins and tests of their predictions using phylogenetic methods. Presents strong evidence that the host lineage for the origin of mitochondria stems from within the archaea, not as the sister to archaea as in the traditional 'three domains' rRNA tree.

85. Williams TA, Embley M: **Archaeal "dark matter" and the origin of eukaryotes**. *Genome Biol Evol* 2014, **6**:474-481.

86. • Guy L, Saw JH, Ettema TJG: **The archaeal legacy of eukaryotes: a phylogenomic perspective[SINGLE]**. *Cold Spring Harb Perspect Biol* 2014, **6** http://dx.doi.org/10.1101/cshperspect.a016022.

An important overview of the discovery and occurrence of eukaryotic cytoskeletal components in certain groups of archaea currently called the TACK superphylum.

87. Lane N, Martin W: **The energetics of genome complexity**. *Nature* 2010, **467**:929-934.

88. Thrash JC, Boyd A, Huggett MJ, Grote J, Carini P, Yoder RJ, Robbertse B, Spatafora JW, Rappe MS, Giovannoni SJ: **Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade**. *Sci Rep* 2011, **1**:13.

89. Brindefalk B, Ettema TJ, Viklund J, Thollesson M, Andersson SG: **A phylometagenomic exploration of oceanic alphaproteobacteria reveals mitochondrial relatives unrelated to the SAR11 clade**. *PLoS ONE* 2011, **6**:e24457.

90. Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D *et al.*: **A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes**. *Mol Biol Evol* 2004, **21**:1643-1660.

91. Abhishek A, Bavishi A, Bavishi A, Choudhary M: **Bacterial genome chimaerism and the origin of mitochondria**. *Can J Mircobiol* 2011, **57**:49-61.

92. Georgiades K, Raoult D: **The rhizome of *Reclinomonas americana*, *Homo sapiens*, *Pediculus humanus* and *Saccharomyces cerevisiae* mitochondria**. *Biol Direct* 2011, **6**:55.

93. Atteia A, Adrait A, Brugière S, van Lis R, Tardif M, Deusch O, Dagan T, Kuhn L, Gontero B, Martin W *et al.*: **A proteomic survey of *Chlamydomonas reinhardtii* mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the a-proteobacterial mitochondrial ancestor**. *Mol Biol Evol* 2009, **29**:1533-1548.

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

95. Le PT, Pontrarotti P, Raoult D: **Alphaproteobacteria species as a source and target of lateral sequence transfers**. *Trends Microbiol* 2014, **22**:147-156.

96. •• Degli Esposti M, Chouaia B, Comandatore F, Crotti E, Sassera D, Lievens PM, Bandi C: **Evolution of mitochondria reconstructed from the energy metabolism of living bacteria[DOUBLE]**. *PLoS ONE* 2014, **9**:e96566.

A comprehensive comparative survey of membrane bioenergetics in alphaproteobacteria that uncovers new and intruiging links in the evolutionary history of mitochondria.

97. Cavanaugh CM, Wirsen CO, Jannasch HW: **Evidence for methylotrophic symbionts in a hydrothermal vent mussel (bivalvia: mytilidae) from the mid-atlantic ridge**. *Appl Environ Microbiol* 1992, **58**:3799-3803.

98. • Müller M *et al.*: **Biochemistry and evolution of anaerobic energy metabolism in eukaryotes[SINGLE]**. *Microbiol Mol Biol Rev* 2012, **76**:444-495.

A comprehensive review of energy metabolism in eukaryotic anaerobes focusing on the role of mitochondria in well-studied anaerobic model organisms.

99. Stairs CW, Eme L, Brown MW, Mutsaers C, Susko E, Dellaire G, Soanes DM, van der Giezen M, Roger AJ: **A SUF Fe-S cluster biogenesis system in the mitochondrion-related organelles of the anaerobic protist *Pygsuia***. *Curr Biol* 2014, **24**:1176-1186.

100. Hug LA, Stechmann A, Roger AJ: **Phylogenetic distributions and histories of proteins involved in anaerobic pyruvate metabolism in eukaryotes**. *Mol Biol Evol* 2010, **27**:311-324.

101 • Gray MW: **The pre-endosymbiont hypothesis: a new perspective on the origin and evolution of mitochondria[SINGLE]**. *Cold Spring Harb Perspect Biol* 2014, **6** http://dx.doi.org/10.1101/cshperspect.a016097.

A proposal to modify the endosymbiont hypothesis for the origin of mitochondria through the suggestion that a compartment similar to the mitochondrion already existed in the host that acquired mitochondrial endosymbiont, whereby this preexisting compartment contained all the proteins of modern mitochondria that do not branch with alphaproteobacterial homolgues in single gene trees.

102. Melkonian M: **Phylogeny of photosynthetic protists and their plastids**. *Verh Dtsch Zool Ges* 1996, **89**:71-96.

3.2

Plastid origin: Who, when and why?

Chuan Ku*, Mayo Roettger*, Verena Zimorski, Shijulal Nelson-Sathi, Filipa L. Sousa, William F. Martin
Institute of Molecular Evolution, Heinrich-Heine-University of Düsseldorf, Germany
* These authors contributed equally
Corresponding author: bill@hhu.de

The presented manuscript was published in the journal *Acta Societatis Botanicorum Poloniae* in 2015.

Contribution of Chuan Ku (co-first author)
Manuscript writing: 40%

Acta Societatis Botanicorum Poloniae

INVITED REVIEW Acta Soc Bot Pol 83(4):281–289 DOI: 10.5586/asbp.2014.045
Received: 2014-11-20 Accepted: 2014-12-04 Published electronically: 2014-12-31

Plastid origin: who, when and why?

Chuan Ku, Mayo Roettger, Verena Zimorski, Shijulal Nelson-Sathi, Filipa L. Sousa, William F. Martin*
Institute of Molecular Evolution, Heinrich-Heine-University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany

Abstract

The origin of plastids is best explained by endosymbiotic theory, which dates back to the early 1900s. Three lines of evidence based on protein import machineries and molecular phylogenies of eukaryote (host) and cyanobacterial (endosymbiont) genes point to a single origin of primary plastids, a unique and important event that successfully transferred two photosystems and oxygenic photosynthesis from prokaryotes to eukaryotes. The nature of the cyanobacterial lineage from which plastids originated has been a topic of investigation. Recent studies have focused on the branching position of the plastid lineage in the phylogeny based on cyanobacterial core genes, that is, genes shared by all cyanobacteria and plastids. These studies have delivered conflicting results, however. In addition, the core genes represent only a very small portion of cyanobacterial genomes and may not be a good proxy for the rest of the ancestral plastid genome. Information in plant nuclear genomes, where most genes that entered the eukaryotic lineage through acquisition from the plastid ancestor reside, suggests that heterocyst-forming cyanobacteria in Stanier's sections IV and V are most similar to the plastid ancestor in terms of gene complement and sequence conservation, which is in agreement with models suggesting an important role of nitrogen fixation in symbioses involving cyanobacteria. Plastid origin is an ancient event that involved a prokaryotic symbiont and a eukaryotic host, organisms with different histories and genome evolutionary processes. The different modes of genome evolution in prokaryotes and eukaryotes bear upon our interpretations of plastid phylogeny.

Keywords: cyanobacteria; endosymbiosis; evolution; gene transfer; genomics; organelle; photosynthesis; phylogeny

Plastid origin: 110 years since Mereschkowsky

Plastids are eukaryotic metabolic compartments responsible for photosynthesis [1] and a variety of metabolic functions including the biosynthesis of amino acids [2], nucleotides [3], lipids [4], and cofactors [5]. The importance of plastids to photosynthetic lineages of eukaryotes cannot be overstated. In 1905, Mereschkowsky proposed a fully articulated version of endosymbiotic theory positing that plastids originated from cyanobacteria that came to reside as symbionts in eukaryotic cells [6,7]. The theory did not make its way into mainstream biological thinking until it was revived in a synthesis by Margulis [8,9] that also incorporated Wallin's [10] – and Paul Portier's (published in French, cited in Sapp [11]) – ideas about endosymbiotic theory for mitochondrial origin, yet adorned by Margulis' own suggestion of a spirochaete origin of flagella. The spirochaete story never took hold, leaving endosymbiotic theory with three main players: the plastid, the mitochondrion, and its host, which is now understood to be an archaeon [12]. Well into the 1970s, resistance to the concept of endosymbiosis for the origin of plastids (and mitochondria) was stiff [13–15].

* Corresponding author. Email: bill@hhu.de
Handling Editor: Andrzej Bodył

With the availability of protein and DNA sequences, of which Mereschkowsky knew nothing, strong evidence had accumulated by the early 1980s that plastids originated through endosymbiosis from cyanobacteria, rather than autogenously [16]. However, there have been recurrent suggestions, starting with Mereschkowsky [6] and tracing into the 1970's [17], that there were several independent origins of plastids from cyanobacteria. Today it is widely, but not universally [18,19], accepted that plastids had a single origin. The strongest evidence for that view is that the protein import machinery, a good marker for endosymbiotic events [20], in the three lineages of Archaeplastida – eukaryotes with primary plastids [21] – consists of homologous host-originated components [22–24], which would not be the case had plastids in those lineages arisen from independent cyanobacterial symbioses.

Genomics has enriched our understanding of plastid origin. We now know that the plastid genome has undergone extreme reduction while leaving its imprints in the nuclear genome through endosymbiotic gene transfer [25,26] and that plastids have spread across eukaryotes through symbiosis and become secondary and tertiary plastids [27]. Genomic data also supports the case for a single plastid origin. Despite some concerns about incomplete sampling and phylogenetic artifacts [18], recent analyses from different groups, incorporating improved phylogenetic algorithms and sampling of cyanobacteria, provide two lines of evidence, in addition to

the protein import machinery, for a monophyletic origin of plastids involving a single host lineage and a single cyanobacterial lineage. (***i***) All plastids are monophyletic and nested within the cyanobacterial clade [28–34]. (***ii***) Eukaryotes with primary plastids are monophyletic based on both nuclear [29,35] and mitochondrial [36] genomes.

However, much is still left unknown about the origin of plastids. What is the cyanobacterial lineage most closely related to plastids? What was the ancestor of plastids like as it entered into the symbiotic relationship? What might be the reasons that triggered this far-reaching event? Here we consider these aspects and recent advances on issues concerning the when, who and why of plastid origin.

When: early or late branching for the plastid lineage within cyanobacteria?

Cyanobacteria are traditionally classified into five sections according to their morphological and developmental patterns [37]. Section I are unicellular cocci, section II are cocci that aggregate, section III are filamentous, section IV are filamentous with heterocysts, and section V are filamentous with true branching and heterocysts. That taxonomy does not, however, correspond to molecular phylogenetic groupings in any phylogeny reported so far. Although different molecular phylogenies are often themselves mutually inconsistent, some general trends regarding the relationships of extant cyanobacteria have emerged from studies based on rDNA or multiple protein-coding genes [30,32–34,38–43], which are summarized as the backbone tree shown in Fig. 1. There is little dispute that the thylakoid-lacking genus *Gloeobacter* (with which *Aphanothece* is possibly synonymous [40,44]) is the most basal lineage within cyanobacteria, whose closest relatives discovered so far are the nonphotosynthetic melainabacteria [43]. The other cyanobacteria can be further divided into a basal clade of *Synechococcus* strains (e.g., JA-2-3B), a clade of *Synechococcus* sp. PCC 7502 and *Pseudanabaena* strains (e.g., PCC 7367), and a core clade consisting of the other taxa (Fig. 1). The core cyanobacteria contain the majority of the described species and strains and can be further divided into three groups: a clade of section I taxa (e.g., *Thermosynechococcus*; this clade's basal position is not recovered in some analyses [31,38,39,41,43]), a clade of mixed section III and I taxa (including the fast-evolving SynPro clade of *Synechococcus* and *Prochlorococcus*) and finally a clade where all sections are represented, including heterocyst-forming and true-branching taxa.

If we take the last common ancestor of the core cyanobacteria as a reference point, much of the recent debate on the closest extant cyanobacterial neighbor of the plastids is about "early" (point "a" in Fig. 1 [30–33]) versus "late" (points "b–e" in Fig. 1 [29,34,38,39,45,46]) plastid branching. Here instead of "early origin" [47] or "deep origin" [30], the term

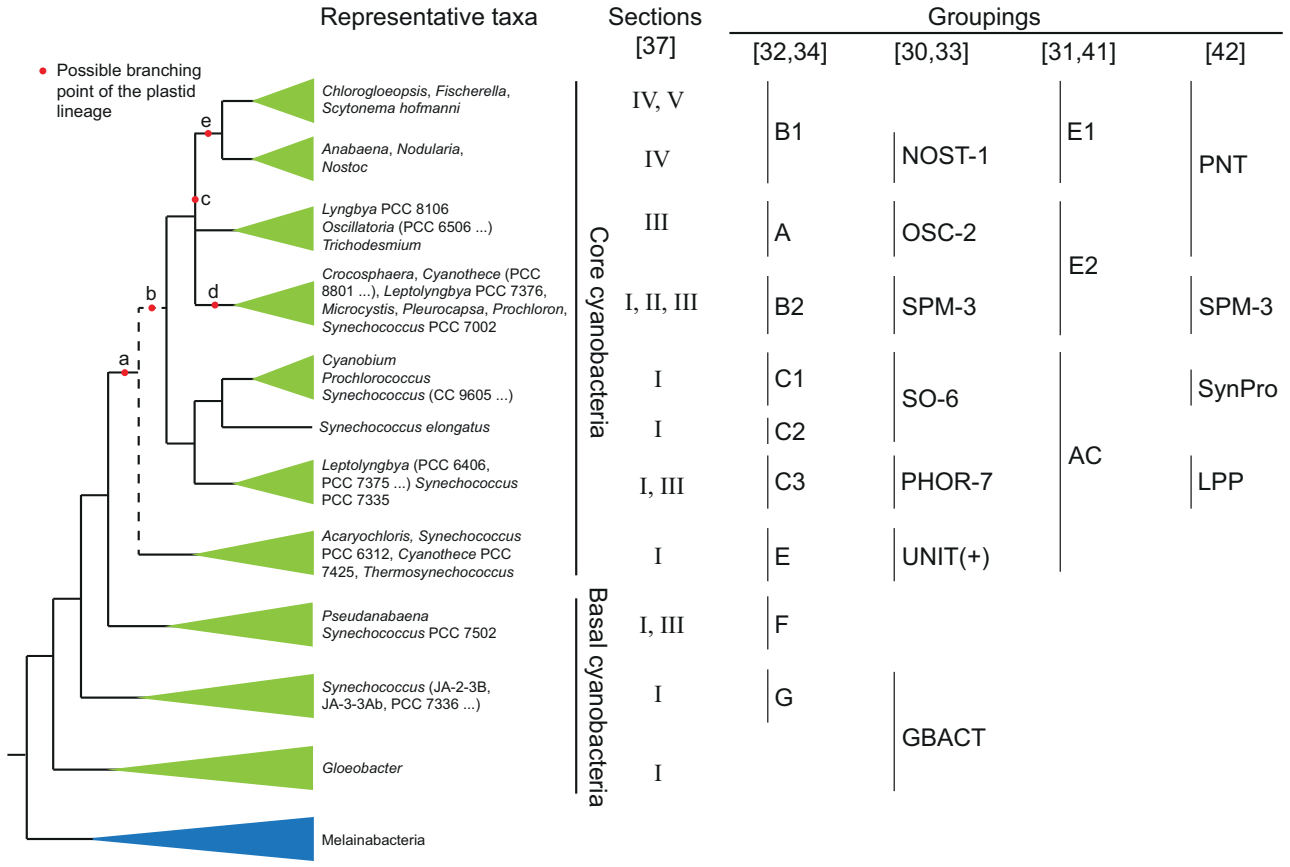


Fig. 1 Possible branching points of the plastid lineage in a consensus cyanobacterial tree based on molecular phylogenetic studies [29–34,38–40,42,43], showing section distribution [37] and groupings in different references. Dotted lines indicate a branching pattern not recovered in some references [31,38,39,41–43]. Branching points: a [30–33]; b [29]; c [34]; d [45] (or within group B2) and [38]; e [39,46] (or within group B1). Numbers in brackets indicate reference numbers for sections and groupings.

"early vs. late branching" is used, because the point where the plastid lineage (the lineage that leads to plastids and does not include other sampled cyanobacteria) branches off the tree might not correspond temporally to the origin of plastids, as the endosymbiosis event itself could have taken place much later than the branching point. Importantly, and as always, discussions of "early" or "late" branching are contingent upon the taxon sample underpinning the tree.

Is there any way to correlate plastid origin to geological time? If a universal molecular clock existed for the cyanobacterial tree, the depth of the branching point could be estimated from sequence divergence. But plastids apparently have a much higher substitution rate than free-living cyanobacteria, among which some taxa such as the group C1 (Fig. 1) also have a higher rate than others. Fossil records indicate that eukaryotes originated no earlier than 1800 million years ago (mya) [48] and complex red algae already existed 1200 mya [49]. These two benchmarks set the upper and lower bounds for the origin of Archaeplastida and plastids, which, according to fossil-calibrated trees that take those benchmarks into account (and is thus not independent of them), is estimated to have taken place around 1500–1600 mya [50,51]. These estimates, however, do not by themselves directly discriminate early- or late-branching scenarios because the branching position of the lineage and the event of plastid origin are two different things.

If the estimated time of the plastid origin is so ancient that it predates the diversification of major cyanobacterial lineages in the core clade (Fig. 1), then a late branching from within the core clade can be ruled out. However, molecular clock estimates suggest the last common ancestor of the core clade existed around 2500–3000 mya [38,41,42]. This makes both the early- and late-branching hypotheses possible. If the former is correct, it would mean that plastids originated from a relatively basal lineage about 1 billion years after the lineage branched off from the main stem of the cyanobacterial tree. If a late-branching hypothesis (points "c", "d" or "e" in Fig. 1) is correct, this would set the branching time to the range 2.3–1.6 billion years ago [38,41,42], much closer in time to the minimum date of plastid origin (1.2 billion years ago) set by *Bangiomorpha* [49,51].

One possible explanation for the hitherto unsettled dispute is that the topologies of trees including genes from plastids (or plastid-derived nuclear genes) and free-living cyanobacteria are highly sensitive to the differences between evolutionary models used in the phylogenetic analyses. In one recently published study, for example, when the model is changed from LG+16Γ to LG+d+CAT, the sister group of plastids changes from the core cyanobacterial clade to the B2 (Fig. 1) clade (supplementary Fig. 3A and 3C in Ochoa de Alda et al. [34]). Additionally, plastid genes and plastid-derived nuclear genes have higher substitution rates than cyanobacterial homologs and this can cause the plastid lineage to branch off at a more basal position [39]. Nucleotide compositional biases of protein-coding genes were also suggested to be misleading in tree reconstruction [33]. These factors and the intrinsic uncertainty associated with deep phylogenetics, especially for prokaryotes, where only some "core" genes can be used for analyses, will likely continue to plague the plastid evolution issue. Indeed, even the earliest studies on large datasets of concatenated plastid proteins revealed that fully resolved, but conflicting trees are obtained [52,53], such that the choice of models determines the result. However, core gene phylogenetics need not be the only way to investigate early plastid evolution.

Plastid origin and the origin of oxygenic photosynthesis

A topic usually circumvented in the origin of plastids issue is the origin of cyanobacteria themselves and the advent of oxygenic photosynthesis. Oxygenic photosynthesis using two photosystems in series only occurs among the cyanobacteria (including plastids). Other phototrophic prokaryotes use only one photosystem (PS), either a homolog of PSII, as in alpha- and gammaproteobacteria and chloroflexi, or a homolog of PSI, as in chlorobia, acidobacteria, and firmicutes (heliobacteria), to carry out anaerobic photosynthesis. Several lines of evidence strongly implicate lateral gene transfer (LGT) in photosystem evolution: (***i***) the photosystems have a very patchy and restricted distribution across distant prokaryotic phyla, (***ii***) the photosystems are found in combination with three unrelated CO_2-fixing pathways (the Calvin cycle, the reverse citric acid cycle and the 3-hydroxypropionate pathway), and (***iii***) the complete machinery for photosystem biogenesis, including chlorophyll biosynthesis, is found on large mobile plasmids in proteobacteria [54]. The main question regarding the origin of oxygenic photosynthesis was whether the two photosystems, which clearly share a common ancestor [55,56], arose within the same genome and were exported as single PSs to other lineages [57] or whether the photosystems had long independent evolutionary histories in bacterial lineages and were merged via LGT in the cyanobacterial ancestor (reviewed in [58]).

These alternatives can be discriminated by studying chlorophyll biosynthesis evolution [59], because the two distinct photosystem types corresponding to reaction center I (RCI) in PSI and reaction center II (RCII) in PSII are clearly related and RC evolution cannot proceed without chlorophyll. Thus, if the divergence of the two RC types reflects lineage divergence [60,61], then chlorophyll biosynthetic pathways supporting RCI- and RCII-based bacterial photosynthesis should reflect a deep dichotomy similar to that observed between the RCs themselves. Conversely, if the two RC types evolved via gene duplication within the same prokaryote [57] – a protocyanobacterium – and were subsequently exported to other lineages, then there should be no deep dichotomy among chlorophyll biosynthetic pathways. Investigation of chlorophyll biosynthesis evolution provided evidence in favor of Allen's protocyanobacterial hypothesis [59], which posits that the first organism that possessed two photosystems was an anaerobe that first used them alternatively, growing either like *Chlorobium* in the presence of sulfide (H_2S photosynthesis using PSI) or like *Rhodobacter* in the absence of sulfide (cyclic electron transport using PSII). A redox-dependent regulator, or redox switch, controlling the mutually exclusive expression of the two photosystems would have allowed the protocyanobacterium to use light in the presence or absence of sulfide. A similar, but not identical, situation is observed

today for *Oscillatoria* [57], which can either express only PSI or both photosystems [62].

The transition from having two photosystems to having oxygenic photosynthesis required the invention of a water splitting apparatus. Blankenship [63,64] has favored a model in which a Mn-dependent enzyme such as manganese catalase was the precursor to the water splitting complex. Other models entail environmentally available Mn^{II} to start [65–67]. Mn^{II} is known to undergo photooxidation to Mn^{III} in the presence of uv light [68], hence in the context of Allen's protocyanobacterial hypothesis, were a mutation to occur in the redox switch, allowing both photosystems to be expressed in an environment where sulfide was lacking but Mn^{II} was available, a flux of electrons from environmental (aqueous) Mn^{II} to ferredoxin would have ensued. The final step would have been the transition from oxidizing environmental Mn^{II} (possibly with the help of uv) one at a time as a substrate, to oxidizing a portable Mn^{III} reserve in the cluster of the water splitting apparatus at the periplasmic side of PSII [69]. Recently, geochemical evidence was reported that is in agreement with both that scenario [70] and the model of Dismukes and colleagues [65,71], which also suggested a role for environmental Mn^{II} while also pointing out a role for high CO_2 in the origin of water splitting. The Mn-oxidizing abilities of RCII from *Rhodobacter* [72] are also compatible with the models deriving water splitting complex from environmental Mn^{II}.

An understanding of the evolution of photosynthesis further highlights the uniqueness and importance of the origin of plastids. Although photosynthetic genes have been exported from ancient cyanobacteria to other prokaryotic lineages, none of them have received and retained both photosystems. The only known case of successful transfer of the two photosystems is the origin of photosynthetic eukaryotes, which involved endosymbiosis of cyanobacterial cells. This made plastid-bearing eukaryotes the only group of organisms other than cyanobacteria that are capable of oxygenic photosynthesis.

Who: the 1% and the 99% of the ancestral plastid genome

There are a number of cyanobacterial genomes sequenced, but only a small fraction of that information is typically used to investigate cyanobacterial or plastid evolution. Studies that aim to infer the backbone phylogeny of plastids and free-living cyanobacteria either use single gene trees (16S rDNA) [31,38,41] or, more commonly, sets of protein-coding genes that included 23 [39], 25 [32], 33 [34], 42 [52], 50 [29] or 75 [33] genes. The number of protein-coding genes in sequenced free-living cyanobacteria ranges from 1716 in *Prochlorococcus marinus* MED4 [73] to 12 356 in *Scytonema hofmanni* PCC 7110, the most gene-rich prokaryote known to date [39]. In other words, the cyanobacterial phylogenies are typically based on about 0.1–1% of the genes present in cyanobacterial genomes.

Based on these phylogenies, we can divide the genomes into clades/lineages (Fig. 1). But what can the lineage labels based on the 1% tell us about the other 99% of the genome? Because of the importance of lateral gene transfer in shaping prokaryotic genomes [74,75], no universal classification can be applied to prokaryotic genomes as a whole [76,77] such that the tree of the shared 1% tells us little, if anything, about how many and what other genes are present in the rest of the genome. Consider, for example, the lineage formed by cyanobacterial sections IV and V (Fig. 1). *Fischerella thermalis* PCC 7521 has 5340 protein-coding genes, not even half as many as *S. hofmanni* in the same lineage [39]. Similarly, different ecotypes of the marine cyanobacterium *Prochlorococcus marinus* show variation in the number of open-reading frames by as much as 40% across strains [78], although their core genes are very similar. Traits such as the filamentous morphology that are determined by multiple genes also cannot be predicted from the core gene phylogeny [32]. In view of these, the core genes can provide a useful taxonomy and convenient genome labels, which however are poor proxies for the number and nature of the genes comprising the other 99%. The debate about the plastid branching point is thus one on how we can best label the plastid ancestor when we try to fit it onto the backbone tree based on 1% of the genome.

Many genes that were present in the cyanobacterial plastid ancestor have been transferred to the nucleus [26,28]. That is, a substantial component of the "other 99%" resides in nuclear chromosomes, but these genes can also be used to address plastid origin. Identification of eukaryote-cyanobacterial homologous genes and sequence similarity comparison suggest that present-day cyanobacteria in the sections IV and V (point "e" in Fig. 1) tend to harbor the most homologs and that they have a higher similarity with those in plastid-bearing eukaryotes [39,46]. This suggests that, in terms of overall genome similarity, extant section IV and V taxa should be most similar to the ancestral plastid, but this does not imply that we would find the plastid ancestor among them. As in the case of mitochondrial endosymbiosis [79], no present-day cyanobacteria would contain the same complement of genes as the plastid ancestor due to the cumulative effect of lateral gene transfer and gene loss.

We know LGT is more frequent among prokaryotes [77] and that the three main mechanisms responsible for prokaryote LGT (transduction, conjugation and transformation) have not been reported to play an important role in bringing genes to eukaryotes from prokaryotes or other eukaryotes. Gene transfers to plastids are also rare [19]. Such findings suggest that 1.5 billion years after plastid origin, even if we know the cyanobacterial lineage most closely related to plastids, it cannot faithfully represent the plastid ancestor in terms of gene content, and cyanobacterial pangenomes figure into this issue. By contrast, the present-day plastid genome, together with plastid-derived nuclear genes, has been more or less "frozen" from the time of endosymbiosis. Therefore, the best reconstruction of the plastid ancestor we can get is probably to identify all nuclear genes of plastid origin by removing the genes contributed by the archaeal host and the mitochondrial endosymbiont [80–82] in addition to the eukaryote-specific gene inventions.

Why: the physiological context of plastid origin

When addressing the ancestral state of microbial physiology that led to the initial advantageous association between the founder endosymbiont and its host, the variety of plastid functions within a eukaryotic cell readily gives rise to different hypotheses. Since Mereschkowsky's initial endosymbiotic theory [6], the production of carbohydrates by the cyanobacterial endosymbiont was thought to be a crucial key for the establishment of the plastids [45]. Additionally, the possible scarcity of freely available oxygen at the time and place where the endosymbiotic event occurred [83–85] led Martin and Müller [80] to propose a syntrophic association between a cyanobacterial symbiont with its heterotrophic, eukaryotic host. In this case, the advantage to the host would have been the coupling of the photosynthetic waste (oxygen) produced by the cyanobacterium with the aerobic respiration that occurred at the host mitochondria. Yet the theory best supported by observations from modern cyanobacterial symbioses is that nitrogen fixation played a crucial role in the establishment of the association between the host and its endosymbiont. In nature, highly diverse taxonomic hosts form symbiotic or endosymbiotic associations with photosynthesizing and N_2-fixing cyanobacteria in various environments. They range from autotrophic algae and plants [86] to heterotrophic fungi where over 1500 species represent cases of lichen symbioses with cyanobacteria [87]. Specifically in plants, cyanobacterial symbioses are present in the four main groups – gymnosperms, angiosperms, pteridophytes and bryophytes [86]. The first confirmed example of a nonfilamentous intracellular endosymbiont with the potential to offer fixed nitrogen to its host is the spheroid body in *Rhopalodia gibba*, a diatom alga [88]. Symbiosis is not restricted to unicellular endosymbionts. The vertically transmitted endosymbiont (*Anabaena*) of the pteridophyte *Azolla*, which was reduced to an organism devoted to nitrogen fixation, is a multicellular cyanobacterium [89]. Another example is the angiosperm *Gunnera manicata* that uses the nitrogen provided by the filamentous symbiont (*Nostoc punctiforme*) and continues growing under N-limited conditions in the presence of the symbiont [90]. In gymnosperms, a nitrogen-fixing cyanobacterial symbiont (*Nostoc*) was found in the roots of most known cycad species [91,92].

It has been shown that nitrogen deprivation stimulates modern symbiotic associations [86,90] and that nitrogen-rich conditions on the contrary facilitate the dissipation of pre-established symbiotic relationships [86]. If nitrogen was the key factor for the host and symbiont to form the initial association, it is congruent with the finding that present-day members of filamentous, heterocyst-forming and N_2-fixing cyanobacterial sections IV and V have a collection of genes most similar to that possessed by the plastid ancestor [39,46]. However, the ability to fix nitrogen was lost over time and modern plant plastids do not perform nitrogen fixation anymore [93]. Nitrogen fixation, including the splitting of the stable triple electron pair bond between the two nitrogen atoms is highly energy expensive. Each mol of fixed nitrogen requires 16 mol of hydrolyzed ATP [94]. The oxidation state of the environment and, as a consequence thereof, the increased availability of nitrate [95] result in a highly decreased need to acquire fixed nitrogen via symbiosis today and shape present-day N_2-fixing cyanobacterial symbioses into a niche solution in N-poor areas [96], whereas at the origin of plastids, it might have played a crucial role.

Interpreting trees for plastid origin

In recent years, there have been repeated claims in the literature that a chlamydial infection had something to do with plastid origin as a kind of a helper symbiont. According to the most recent version of the hypothesis [97], this hypothetical chlamydial endosymbiont, which is found neither in any extant plastid-bearing lineages nor in any contemporary cyanobacterial symbioses, had contributed a significant number of genes to the nuclear genome. Some problems with the chlamydial hypothesis and reasons why it is unlikely to be true have been discussed previously [20]. Among them is the circumstance that gene transfer from endosymbionts to their host has become very popular topics these days, and people continuously find the idea of gene transfer interesting, be it the human genome, where many claims for gene transfer turned out to be artefacts [98], the case of the plastid-bearing slug *Elysia* [99], where the claims for gene transfer also turned out to be unfounded [100,101], the case of trypanosomes [102,103], where claims for gene transfer also turned out not to be true [104], or the case of ciliates, where claims for gene transfer in the context of secondary endosymbiosis also turned out not to be true [105]. Of those examples, the chlamydia case is most similar to the trypanosome and ciliate cases, because the chlamydia story involves small numbers of genes with odd branching patterns that (***i***) do not stand out above the background signals to be expected in such analyses [105] and that (***ii***) were identified as such in earlier studies of plant genomes [28]. Indeed, none of the examples of cyanobacterial symbioses outlined in the previous section involve chlamydia or any other helper bacteria, cyanobacteria do just fine by themselves when it comes to establishing stable symbioses, both intracellular and extracellular.

Since the chlamydial hypothesis is solely based on trees, rather than cellular evidence as in the endosymbiotic theory for plastids, it is important to know what the trees can really tell us. In addition to potential phylogenetic errors in the trees showing transfers from chlamydiae [106], it is notable that many other lineages of prokaryotes also appear to have donated genes to eukaryotes. Anyone interested in proposing a hypothesis similar to the chlamydial one can readily single out another lineage and a eukaryotic compartment or pathways where more proteins have been apparently donated from that lineage. This appearance, however, comes from the circumstance that genes donated to eukaryotes by plastids and mitochondria continue to be transferred among free-living prokaryotes long after the organelle had its origin [79,107–109]. Chlamydiae and cyanobacteria even specifically share a sufficiently large number of genes that they form a common cluster (a module) in gene sharing networks (Fig. 1C in Dagan et al. [75]). Genes forming that module can generate trees in which chlamydiae look as if they donated genes to eukaryotes [75]. Though proponents of the

chlamydial helper symbiont hypothesis tend to overlook this effect, cognoscenti increasingly appreciate how gene transfer among prokaryotes affects inferences of gene origin through the endosymbiotic origin of organelles [109].

Here, a critic might ask whether apparent chimerism such as that observed for nuclear-encoded genes is also observed for organelle-encoded proteins, and if not, why not? The answer is that it is observed, the problem being that very few studies have ever looked for such (apparent) chimerism. For example, the phylogenies of the proteins in the *Reclinomonas* mitochondrial genome [110], long the largest mitochondrial genome known with 63 protein coding genes, were rarely scrutinized. In 2004, Esser et al. [111], did however investigate the phylogeny of the 55 proteins encoded in the *Reclinomonas americana* mitochondrial genome that are sufficiently well conserved for phylogenetic inference. They found that "the *Reclinomonas protein* branched with homologues from *Rickettsia* species in 5 trees, with homologues from *Wolbachia* in 10 trees, basal to *Rickettsia* and *Wolbachia* in 5 trees, with other α-proteobacteria or groupings thereof in 16 trees, and not with homologues from any α-proteobacterium in 19 trees with bootstrap proportions less than 70% for 53 of the 55 proteins studied" (reference [111] p. 1646). They furthermore surmised: "Recalling that the *Reclinomonas* mitochondrion inherited its genome from proteobacteria, rather than having acquired it through lateral acquisition from various donors, such disparate results could mean (1) that a degree of noise exists in the data (for example, due to poor conservation, as in the case of the twelve proteins that were excluded for lack of good homologues); (2) that the phylogenetic method is producing an imperfect estimation of the phylogeny, producing artifacts in some cases, but getting close to the true position in other cases; (3) that any number of problems inherent to phylogeny reconstruction, such as model misspecification or poor sampling, were present; (4) that the eubacteria sampled might be avidly exchanging these genes over time; or (5) any combination of the above." (reference [111] p. 1646). That is very much the same thing as we are saying here.

With regard to plastid encoded proteins, similar studies are lacking to our knowledge, but for nuclear encoded genes in plants, Martin et al. [28] did find that phylogenetic trees for three cyanobacterial genomes "suggest at face value the *Arabidopsis* lineage to have acquired genes not from one cyanobacterium, but from all three sampled [even at a bootstrap probability (BP) ≥ 0.95], whereby that view contradicts independent evidence suggesting a single origin of plastids from one cyanobacterium, not three or more in the *Arabidopsis* lineage". We note that concatenation, which is very popular for the study of organelle genome phylogeny, condenses the many disparate signals that individual genes contain into a single averaged signal [29,30,32–34], while studies that investigate nuclear genes or (n.b.) seek evidence for gene transfer to eukaryotes are based on individual gene phylogenies [97]. Furthermore, concatenation harbors many pitfalls that are not yet well-understood [112]. Supertrees, which summarize the information from many trees in a single tree, support the participation of a plastid, a mitochondrion, and a host in eukaryote evolution, but no other extra partners [113]. Clearly, there is a need for additional critical studies of the phylogenetics of organelle origins.

Conclusion

Based on evidence from protein import machineries and phylogenies of eukaryotes and prokaryotes, plastids had a single origin involving a eukaryotic host and a cyanobacterial endosymbiont. This event is unique in that it is the only known successful transfer of oxygenic photosynthesis from cyanobacteria to another lineage. The origin of plastids might have involved nitrogen fixation during the initial physiological interactions of the symbiosis, as suggested by the overall similarity between genomes of diazotrophic cyanobacteria and plastid-derived nuclear genes in photosynthetic eukaryotes. Genomics and phylogenetics continue to enrich our understanding of plastid origin in the context of cyanobacterial evolution, but even with the copious amounts of data available, "core" genome phylogenies reported by different groups tend to conflict. This is because phylogenies are heavily model-dependent [12]. There is thus a need to look beyond the "1%" core genome phylogenies when addressing plastid origin. And if we keep lateral gene transfer among prokaryotes in mind – and the resulting fluid nature of prokaryotic chromosomes in the context of organelle origins [107–109] – we can make sense of the phylogenetic patterns observed in plant and algal genomes, without the need to infer spirochaete or chlamydial gene donors or the like. Understanding endosymbiosis in eukaryote genome evolution in the context of prokaryote genome evolution is a challenging exercise of keeping the bigger picture in focus.

Acknowledgments

This work was funded by the European Research Council (grant No. 232975 to WFM). CK is grateful to the Deutscher Akademischer Austauschdienst for a PhD stipend.

Authors' contributions

The following declarations about authors' contributions to the research have been made: compiled the literature: CK, MR, VZ, FS, SN-S, WFM; prepared the figure: CK; wrote the manuscript: CK, MR, VZ, FS, SN-S, WFM. CK and MR contributed equally to this work.

Competing interests

No competing interests have been declared.

References

1. Allen JF. Photosynthesis of ATP – electrons, proton pumps, rotors, and poise. Cell. 2002;110(3):273–276. http://dx.doi.org/10.1016/S0092-8674(02)00870-X
2. Hagelstein P, Sieve B, Klein M, Jans H, Schultz G. Leucine synthesis in chloroplasts: leucine/isoleucine aminotransferase and valine aminotransferase are different enzymes in spinach chloroplasts. J Plant Physiol. 1997;150(1–2):23–30. http://dx.doi.org/10.1016/S0176-1617(97)80176-9
3. Zrenner R, Stitt M, Sonnewald U, Boldt R. Pyrimidine and purine biosynthesis and degradation in plants. Annu Rev Plant Biol. 2006;57(1):805–836. http://dx.doi.org/10.1146/annurev.arplant.57.032905.105421

4. Wang Z, Benning C. Chloroplast lipid synthesis and lipid trafficking through ER–plastid membrane contact sites. Biochem Soc Trans. 2012;40(2):457–463. http://dx.doi.org/10.1042/BST20110752
5. Gerdes S, Lerma-Ortiz C, Frelin O, Seaver SMD, Henry CS, de Crecy-Lagard V, et al. Plant B vitamin pathways and their compartmentation: a guide for the perplexed. J Exp Bot. 2012;63(15):5379–5395. http://dx.doi.org/10.1093/jxb/ers208
6. Mereschkowsky C. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol Cent. 1905;25(18):593–604.
7. Martin W, Kowallik K. Annotated English translation of Mereschkowsky's 1905 paper "Über Natur und Ursprung der Chromatophoren imPflanzenreiche." Eur J Phycol. 1999;34(3):287–295. http://dx.doi.org/10.1080/09670269910001736342
8. Sagan L. On the origin of mitosing cells. J Theor Biol. 1967;14(3):225–274. http://dx.doi.org/10.1016/0022-5193(67)90079-3
9. Margulis L. Origin of eukaryotic cells. New Haven, CT: Yale University Press; 1970.
10. Wallin IE. Symbionticism and the origin of species. London: Tindall and Cox; 1927.
11. Sapp J. Evolution by association: a history of symbiosis. New York, NY: Oxford University Press; 1994.
12. Williams TA, Foster PG, Cox CJ, Embley TM. An archaeal origin of eukaryotes supports only two primary domains of life. Nature. 2013;504(7479):231–236. http://dx.doi.org/10.1038/nature12779
13. Raff RA, Mahler HR. The non symbiotic origin of mitochondria. Science. 1972;177(4049):575–582. http://dx.doi.org/10.1126/science.177.4049.575
14. Bogorad L. Evolution of organelles and eukaryotic genomes. Science. 1975;188(4191):891–898. http://dx.doi.org/10.1126/science.1138359
15. Cavalier-Smith T. The origin of nuclei and of eukaryotic cells. Nature. 1975;256(5517):463–468. http://dx.doi.org/10.1038/256463a0
16. Gray MW, Doolittle WF. Has the endosymbiont hypothesis been proven? Microbiol Rev. 1982;46(1):1–42.
17. Raven PH. A multiple origin for plastids and mitochondria: many independent symbiotic events may have been involved in the origin of these cellular organelles. Science. 1970;169(3946):641–646. http://dx.doi.org/10.1126/science.169.3946.641
18. Howe CJ, Barbrook AC, Nisbet RER, Lockhart PJ, Larkum AWD. The origin of plastids. Philos Trans R Soc Lond B Biol Sci. 2008;363(1504):2675–2685. http://dx.doi.org/10.1098/rstb.2008.0050
19. Stiller JW. Toward an empirical framework for interpreting plastid evolution. J Phycol. 2014;50(3):462–471. http://dx.doi.org/10.1111/jpy.12178
20. Zimorski V, Ku C, Martin WF, Gould SB. Endosymbiotic theory for organelle origins. Curr Opin Microbiol. 2014;22:38–48. http://dx.doi.org/10.1016/j.mib.2014.09.008
21. Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, et al. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol. 2005;52(5):399–451. http://dx.doi.org/10.1111/j.1550-7408.2005.00053.x
22. McFadden GI, van Dooren GG. Evolution: red algal genome affirms a common origin of all plastids. Curr Biol. 2004;14(13):R514–R516. http://dx.doi.org/10.1016/j.cub.2004.06.041
23. Steiner JM, Yusa F, Pompe JA, Löffelhardt W. Homologous protein import machineries in chloroplasts and cyanelles. Plant J. 2005;44(4):646–652. http://dx.doi.org/10.1111/j.1365-313X.2005.02559.x
24. Shi LX, Theg SM. The chloroplast protein import system: from algae to trees. Biochim Biophys Acta. 2013;1833(2):314–331. http://dx.doi.org/10.1016/j.bbamcr.2012.10.002
25. Martin W, Brinkmann H, Savonna C, Cerff R. Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. Proc Natl Acad Sci USA. 1993;90(18):8692–8696. http://dx.doi.org/10.1073/pnas.90.18.8692
26. Timmis JN, Ayliffe MA, Huang CY, Martin W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 2004;5(2):123–135. http://dx.doi.org/10.1038/nrg1271
27. Gould SB, Waller RF, McFadden GI. Plastid evolution. Annu Rev Plant Biol. 2008;59(1):491–517. http://dx.doi.org/10.1146/annurev.arplant.59.032607.092915
28. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, et al. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA. 2002;99(19):12246–12251. http://dx.doi.org/10.1073/pnas.182432999
29. Rodríguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Löffelhardt W, et al. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol. 2005;15(14):1325–1330. http://dx.doi.org/10.1016/j.cub.2005.06.040
30. Criscuolo A, Gribaldo S. Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. Mol Biol Evol. 2011;28(11):3019–3032. http://dx.doi.org/10.1093/molbev/msr108
31. Schirrmeister BE, Antonelli A, Bagheri HC. The origin of multicellularity in cyanobacteria. BMC Evol Biol. 2011;11(1):45. http://dx.doi.org/10.1186/1471-2148-11-45
32. Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, et al. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proc Natl Acad Sci USA. 2013;110(3):1053–1058. http://dx.doi.org/10.1073/pnas.1217107110
33. Li B, Lopes JS, Foster PG, Embley TM, Cox CJ. Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. Mol Biol Evol. 2014;31(7):1697–1709. http://dx.doi.org/10.1093/molbev/msu105
34. Ochoa de Alda JAG, Esteban R, Diago ML, Houmard J. The plastid ancestor originated among one of the major cyanobacterial lineages. Nat Commun. 2014;5:4937. http://dx.doi.org/10.1038/ncomms5937
35. Katz LA, Grant JR, Parfrey LW, Burleigh JG. Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. Syst Biol. 2012;61(4):653–660. http://dx.doi.org/10.1093/sysbio/sys026
36. Jackson CJ, Reyes-Prieto A. The mitochondrial genomes of the glaucophytes *Gloeochaete wittrockiana* and *Cyanoptyche gloeocystis*: multilocus phylogenetics suggests a monophyletic archaeplastida. Genome Biol Evol. 2014;6(10):2774–2785. http://dx.doi.org/10.1093/gbe/evu218
37. Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. Generic assignments, strain histories and properties of pure cultures of cyanobacteria. J Gen Microbiol. 1979;111(1):1–61. http://dx.doi.org/10.1099/00221287-111-1-1
38. Falcón LI, Magallón S, Castillo A. Dating the cyanobacterial ancestor of the chloroplast. ISME J. 2010;4(6):777–783. http://dx.doi.org/10.1038/ismej.2010.2
39. Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major P, et al. Genomes of stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. Genome Biol Evol. 2013;5(1):31–44. http://dx.doi.org/10.1093/gbe/evs117
40. Mareš J, Hrouzek P, Kaňa R, Ventura S, Strunecký O, Komárek J. The primitive thylakoid-less cyanobacterium *Gloeobacter* is a common rock-dwelling organism. PLoS ONE. 2013;8(6):e66323. http://dx.doi.org/10.1371/journal.pone.0066323
41. Schirrmeister BE, de Vos JM, Antonelli A, Bagheri HC. Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. Proc Natl Acad Sci USA. 2013;110(5):1791–1796. http://dx.doi.org/10.1073/pnas.1209927110
42. Sánchez-Baracaldo P, Ridgwell A, Raven JA. A neoproterozoic transition in the marine nitrogen cycle. Curr Biol. 2014;24(6):652–657. http://dx.doi.org/10.1016/j.cub.2014.01.041
43. Soo RM, Skennerton CT, Sekiguchi Y, Imelfort M, Paech SJ, Dennis PG, et al. An expanded genomic representation of the phylum Cyanobacteria. Genome Biol Evol. 2014;6(5):1031–1045. http://dx.doi.org/10.1093/gbe/evu073
44. Mareš J, Komárek J, Compère P, Oren A. Validation of the generic name *Gloeobacter* Rippka et al. 1974, Cyanophyceae. Cryptogam Algol. 2013;34(3):255–262. http://dx.doi.org/10.7872/crya.v34.iss3.2013.255
45. Deschamps P, Colleoni C, Nakamura Y, Suzuki E, Putaux JL, Buleon A, et al. Metabolic symbiosis and the birth of the plant kingdom. Mol Biol Evol. 2008;25(3):536–548. http://dx.doi.org/10.1093/molbev/msm280

46. Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, et al. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Mol Biol Evol. 2008;25(4):748–761. http://dx.doi.org/10.1093/molbev/msn022
47. Nelissen B, van de Peer Y, Wilmotte A, de Wachter R. An early origin of plastids within the cyanobacterial divergence is suggested by evolutionary trees based on complete 16S rRNA sequences. Mol Biol Evol. 1995;12(6):1166–1173.
48. Knoll AH. Paleobiological perspectives on early eukaryotic evolution. Cold Spring Harb Perspect Biol. 2014;6(1):a016121. http://dx.doi.org/10.1101/cshperspect.a016121
49. Butterfield NJ. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology. 2000;26(3):386–404. http://dx.doi.org/10.1666/0094-8373(2000)026<0386:BPNGNS>2.0.CO;2
50. Yoon HS. A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol. 2004;21(5):809–818. http://dx.doi.org/10.1093/molbev/msh075
51. Parfrey LW, Lahr DJG, Knoll AH, Katz LA. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci USA. 2011;108(33):13624–13629. http://dx.doi.org/10.1073/pnas.1110633108
52. Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. Gene transfer to the nucleus and the evolution of chloroplasts. Nature. 1998;393(6681):162–165. http://dx.doi.org/10.1038/30234
53. Lockhart PJ, Howe CJ, Barbrook AC, Larkum AWD, Penny D. Spectral analysis, systematic bias, and the evolution of chloroplasts. Mol Biol Evol. 1999;16(4):573.
54. Petersen J, Brinkmann H, Bunk B, Michael V, Päuker O, Pradella S. Think pink: photosynthesis, plasmids and the *Roseobacter* clade: plasmids and phototrophy. Environ Microbiol. 2012;14(10):2661–2672. http://dx.doi.org/10.1111/j.1462-2920.2012.02806.x
55. Schubert WD, Klukas O, Saenger W, Witt HT, Fromme P, Krauß N. A common ancestor for oxygenic and anoxygenic photosynthetic systems. J Mol Biol. 1998;280(2):297–314. http://dx.doi.org/10.1006/jmbi.1998.1824
56. Sadekar S. Conservation of distantly related membrane proteins: photosynthetic reaction centers share a common structural core. Mol Biol Evol. 2006;23(11):2001–2007. http://dx.doi.org/10.1093/molbev/msl079
57. Allen JF. A redox switch hypothesis for the origin of two light reactions in photosynthesis. FEBS Lett. 2005;579(5):963–968. http://dx.doi.org/10.1016/j.febslet.2005.01.015
58. Olson JM, Blankenship RE. Thinking about the evolution of photosynthesis. Photosynth Res. 2004;80(1–3):373–386. http://dx.doi.org/10.1023/B:PRES.0000030457.06495.83
59. Sousa FL, Shavit-Grievink L, Allen JF, Martin WF. Chlorophyll biosynthesis gene evolution indicates photosystem gene duplication, not photosystem merger, at the origin of oxygenic photosynthesis. Genome Biol Evol. 2013;5(1):200–216. http://dx.doi.org/10.1093/gbe/evs127
60. Blankenship RE. Molecular evidence for the evolution of photosynthesis. Trends Plant Sci. 2001;6(1):4–6. http://dx.doi.org/10.1016/S1360-1385(00)01831-8
61. Hohmann-Marriott MF, Blankenship RE. Evolution of photosynthesis. Annu Rev Plant Biol. 2011;62(1):515–548. http://dx.doi.org/10.1146/annurev-arplant-042110-103811
62. Oren A, Padan E. Induction of anaerobic, photoautotrophic growth in the cyanobacterium *Oscillatoria limnetica*. J Bacteriol. 1978;133(2):558–563.
63. Blankenship RE, Hartman H. The origin and evolution of oxygenic photosynthesis. Trends Biochem Sci. 1998;23(3):94–97. http://dx.doi.org/10.1016/S0968-0004(98)01186-4
64. Raymond J, Blankenship R. The origin of the oxygen-evolving complex. Coord Chem Rev. 2008;252(3–4):377–383. http://dx.doi.org/10.1016/j.ccr.2007.08.026
65. Dismukes GC, Klimov VV, Baranov SV, Kozlov YN, DasGupta J, Tyryshkin A. The origin of atmospheric oxygen on Earth: the innovation of oxygenic photosynthesis. Proc Natl Acad Sci USA. 2001;98(5):2170–2175. http://dx.doi.org/10.1073/pnas.061514798
66. Sauer K, Yachandra VK. A possible evolutionary origin for the Mn_4 cluster of the photosynthetic water oxidation complex from natural MnO_2 precipitates in the early ocean. Proc Natl Acad Sci USA. 2002;99(13):8631–8636. http://dx.doi.org/10.1073/pnas.132266199
67. Allen JF, Martin W. Evolutionary biology: out of thin air. Nature. 2007;445(7128):610–612. http://dx.doi.org/10.1038/445610a
68. Hakala M. Photoinhibition of manganese enzymes: insights into the mechanism of photosystem II photoinhibition. J Exp Bot. 2006;57(8):1809–1816. http://dx.doi.org/10.1093/jxb/erj189
69. Kupitz C, Basu S, Grotjohann I, Fromme R, Zatsepin NA, Rendek KN, et al. Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. Nature. 2014;513(7517):261–265. http://dx.doi.org/10.1038/nature13453
70. Johnson JE, Webb SM, Thomas K, Ono S, Kirschvink JL, Fischer WW. Manganese-oxidizing photosynthesis before the rise of cyanobacteria. Proc Natl Acad Sci USA. 2013;110(28):11238–11243. http://dx.doi.org/10.1073/pnas.1305530110
71. Khorobrykh A, Dasgupta J, Kolling DRJ, Terentyev V, Klimov VV, Dismukes GC. Evolutionary origins of the photosynthetic water oxidation cluster: bicarbonate permits Mn^{2+} photo-oxidation by anoxygenic bacterial reaction centers. Chembiochem. 2013;14(14):1725–1731. http://dx.doi.org/10.1002/cbic.201300355
72. Allen JP, Olson TL, Oyala P, Lee WJ, Tufts AA, Williams JC. Light-driven oxygen production from superoxide by Mn-binding bacterial reaction centers. Proc Natl Acad Sci USA. 2012;109(7):2314–2318. http://dx.doi.org/10.1073/pnas.1115364109
73. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, et al. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature. 2003;424(6952):1042–1047. http://dx.doi.org/10.1038/nature01947
74. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000;405(6784):299–304. http://dx.doi.org/10.1038/35012500
75. Dagan T, Artzy-Randrup Y, Martin W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci USA. 2008;105(29):10039–10044. http://dx.doi.org/10.1073/pnas.0800679105
76. Doolittle WF. Phylogenetic classification and the universal tree. Science. 1999;284(5423):2124–2128. http://dx.doi.org/10.1126/science.284.5423.2124
77. Doolittle WF, Bapteste E. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci USA. 2007;104(7):2043–2049. http://dx.doi.org/10.1073/pnas.0610699104
78. Paul S, Dutta A, Bag SK, Das S, Dutta C. Distinct, ecotype-specific genome and proteome signatures in the marine cyanobacteria *Prochlorococcus*. BMC Genomics. 2010;11(1):103. http://dx.doi.org/10.1186/1471-2164-11-103
79. Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol. 2012;4(4):466–485. http://dx.doi.org/10.1093/gbe/evs018
80. Martin W, Müller M. The hydrogen hypothesis for the first eukaryote. Nature. 1998;392(6671):37–41. http://dx.doi.org/10.1038/32096
81. McInerney JO, O'Connell MJ, Pisani D. The hybrid nature of the eukaryota and a consilient view of life on Earth. Nat Rev Microbiol. 2014;12(6):449–455. http://dx.doi.org/10.1038/nrmicro3271
82. Williams TA, Embley TM. Archaeal "dark matter" and the origin of eukaryotes. Genome Biol Evol. 2014;6(3):474–481. http://dx.doi.org/10.1093/gbe/evu031
83. Holland HD. The oxygenation of the atmosphere and oceans. Philos Trans R Soc Lond B Biol Sci. 2006;361(1470):903–915. http://dx.doi.org/10.1098/rstb.2006.1838
84. Johnston DT, Wolfe-Simon F, Pearson A, Knoll AH. Anoxygenic photosynthesis modulated Proterozoic oxygen and sustained Earth's middle age. Proc Natl Acad Sci USA. 2009;106(40):16925–16929. http://dx.doi.org/10.1073/pnas.0909248106

85. Kasting J. Earth's early atmosphere. Science. 1993;259(5097):920–926. http://dx.doi.org/10.1126/science.11536547

86. Rai AN, Söderbäck E, Bergman B. Cyanobacterium-plant symbioses. New Phytol. 2000;147(3):449–481. http://dx.doi.org/10.1046/j.1469-8137.2000.00720.x

87. Rikkinen J. Lichen guilds share related cyanobacterial symbionts. Science. 2002;297(5580):357–357. http://dx.doi.org/10.1126/science.1072961

88. Prechtl J. Intracellular spheroid bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin. Mol Biol Evol. 2004;21(8):1477–1481. http://dx.doi.org/10.1093/molbev/msh086

89. Ran L, Larsson J, Vigil-Stenman T, Nylander JAA, Ininbergs K, Zheng WW, et al. Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. PLoS ONE. 2010;5(7):e11486. http://dx.doi.org/10.1371/journal.pone.0011486

90. Chiu WL. Nitrogen deprivation stimulates symbiotic gland development in *Gunnera manicata*. Plant Physiol. 2005;139(1):224–230. http://dx.doi.org/10.1104/pp.105.064931

91. Rai AN, Bergman B, Rasmussen U, editors. Cyanobacteria in symbiosis. Dordrecht: Kluwer Academic Publishers; 2002.

92. Costa JL, Romero EM, Lindblad P. Sequence based data supports a single *Nostoc* strain in individual coralloid roots of cycads. FEMS Microbiol Ecol. 2004;49(3):481–487. http://dx.doi.org/10.1016/j.femsec.2004.05.001

93. Allen JF, Raven JA. Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. J Mol Evol. 1996;42(5):482–492. http://dx.doi.org/10.1007/BF02352278

94. Kneip C, Lockhart P, Voß C, Maier UG. Nitrogen fixation in eukaryotes – new models for symbiosis. BMC Evol Biol. 2007;7(1):55. http://dx.doi.org/10.1186/1471-2148-7-55

95. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. Science. 2008;320(5879):1034–1039. http://dx.doi.org/10.1126/science.1153213

96. Usher KM, Bergman B, Raven JA. Exploring cyanobacterial mutualisms. Annu Rev Ecol Evol Syst. 2007;38(1):255–273. http://dx.doi.org/10.1146/annurev.ecolsys.38.091206.095641

97. Facchinelli F, Colleoni C, Ball SG, Weber APM. Chlamydia, cyanobiont, or host: who was on top in the ménage à trois? Trends Plant Sci. 2013;18(12):673–679. http://dx.doi.org/10.1016/j.tplants.2013.09.006

98. Salzberg SL. Microbial genes in the human genome: lateral transfer or gene loss? Science. 2001;292(5523):1903–1906. http://dx.doi.org/10.1126/science.1061036

99. Rumpho ME, Worful JM, Lee J, Kannan K, Tyler MS, Bhattacharya D, et al. Horizontal gene transfer of the algal nuclear gene *psbO* to the photosynthetic sea slug *Elysia chlorotica*. Proc Natl Acad Sci USA. 2008;105(46):17867–17871. http://dx.doi.org/10.1073/pnas.0804968105

100. Wagele H, Deusch O, Handeler K, Martin R, Schmitt V, Christa G, et al. Transcriptomic evidence that longevity of acquired plastids in the photosynthetic slugs *Elysia timida* and *Plakobranchus ocellatus* does not entail lateral transfer of algal nuclear genes. Mol Biol Evol. 2011;28(1):699–706. http://dx.doi.org/10.1093/molbev/msq239

101. de Vries J, Christa G, Gould SB. Plastid survival in the cytosol of animal cells. Trends Plant Sci. 2014;19(6):347–350. http://dx.doi.org/10.1016/j.tplants.2014.03.010

102. Hannaert V, Saavedra E, Duffieux F, Szikora JP, Rigden DJ, Michels PAM, et al. Plant-like traits associated with metabolism of *Trypanosoma parasites*. Proc Natl Acad Sci USA. 2003;100(3):1067–1071. http://dx.doi.org/10.1073/pnas.0335769100

103. Martin W, Borst P. Secondary loss of chloroplasts in trypanosomes. Proc Natl Acad Sci USA. 2003;100(3):765–767. http://dx.doi.org/10.1073/pnas.0437776100

104. Berriman M. The genome of the African trypanosome *Trypanosoma brucei*. Science. 2005;309(5733):416–422. http://dx.doi.org/10.1126/science.1112642

105. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, et al. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. 2006;4(9):e286. http://dx.doi.org/10.1371/journal.pbio.0040286

106. Moreira D, Deschamps P. What was the real contribution of endosymbionts to the eukaryotic nucleus? Insights from photosynthetic eukaryotes. Cold Spring Harb Perspect Biol. 2014;6(7):a016014. http://dx.doi.org/10.1101/cshperspect.a016014

107. Martin W. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. Bioessays. 1999;21(2):99–104. http://dx.doi.org/10.1002/(SICI)1521-1878(199902)21:2<99::AID-BIES3>3.0.CO;2-B

108. Esser C, Martin W, Dagan T. The origin of mitochondria in light of a fluid prokaryotic chromosome model. Biol Lett. 2007;3(2):180–184. http://dx.doi.org/10.1098/rsbl.2006.0582

109. Richards TA, Archibald JM. Cell evolution: gene transfer agents and the origin of mitochondria. Curr Biol. 2011;21(3):R112–R114. http://dx.doi.org/10.1016/j.cub.2010.12.036

110. Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, et al. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature. 1997;387(6632):493–497. http://dx.doi.org/10.1038/387493a0

111. Esser C. A genome phylogeny for mitochondria among alphaproteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol. 2004;21(9):1643–1660. http://dx.doi.org/10.1093/molbev/msh160

112. Thiergart T, Landan G, Martin WF. Concatenated alignments and the case of the disappearing tree. BMC Evol Biol. 2015 (in press).

113. Pisani D, Cotton JA, McInerney JO. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol. 2007;24(8):1752–1760. http://dx.doi.org/10.1093/molbev/msm095

3.3

Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes

Chuan Ku[1,*], Shijulal Nelson-Sathi[1,*], Mayo Roettger[1], Sriram Garg[1], Einat Hazkani-Covo[2], William F. Martin[1]

[1] Institute of Molecular Evolution, Heinrich-Heine-University of Düsseldorf, Germany

[2] Department of Natural and Life Sciences, The Open University of Israel, Israel

* These authors contributed equally

Corresponding author: bill@hhu.de

The presented manuscript was published in the journal *Proceedings of the National Academy of Sciences of the United States of America* in 2015.

Contribution of Chuan Ku (co-first author)

Experimental design: 35%

Data analysis: 35%

Manuscript writing: 30%

COLLOQUIUM PAPER

EVOLUTION

Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes

Chuan Ku[a,1], Shijulal Nelson-Sathi[a,1], Mayo Roettger[a], Sriram Garg[a], Einat Hazkani-Covo[b], and William F. Martin[a,2]

[a]Institute of Molecular Evolution, Heinrich Heine University, 40225 Düsseldorf, Germany; and [b]Department of Natural and Life Sciences, The Open University of Israel, Ra'anana 43107, Israel

Edited by John P. McCutcheon, University of Montana, Missoula, MT, and accepted by the Editorial Board February 5, 2015 (received for review December 12, 2014)

Endosymbiotic theory in eukaryotic-cell evolution rests upon a foundation of three cornerstone partners—the plastid (a cyanobacterium), the mitochondrion (a proteobacterium), and its host (an archaeon)—and carries a corollary that, over time, the majority of genes once present in the organelle genomes were relinquished to the chromosomes of the host (endosymbiotic gene transfer). However, notwithstanding eukaryote-specific gene inventions, single-gene phylogenies have never traced eukaryotic genes to three single prokaryotic sources, an issue that hinges crucially upon factors influencing phylogenetic inference. In the age of genomes, single-gene trees, once used to test the predictions of endosymbiotic theory, now spawn new theories that stand to eventually replace endosymbiotic theory with descriptive, gene tree-based variants featuring supernumerary symbionts: prokaryotic partners distinct from the cornerstone trio and whose existence is inferred solely from single-gene trees. We reason that the endosymbiotic ancestors of mitochondria and chloroplasts brought into the eukaryotic—and plant and algal—lineage a genome-sized sample of genes from the proteobacterial and cyanobacterial pangenomes of their respective day and that, even if molecular phylogeny were artifact-free, sampling prokaryotic pangenomes through endosymbiotic gene transfer would lead to inherited chimerism. Recombination in prokaryotes (transduction, conjugation, transformation) differs from recombination in eukaryotes (sex). Prokaryotic recombination leads to pangenomes, and eukaryotic recombination leads to vertical inheritance. Viewed from the perspective of endosymbiotic theory, the critical transition at the eukaryote origin that allowed escape from Muller's ratchet—the origin of eukaryotic recombination, or sex—might have required surprisingly little evolutionary innovation.

endosymbiosis | evolution | mitochondria | lateral gene transfer | plastids

The origin of eukaryotes was one of life's major evolutionary transitions (1, 2). Despite much progress in recent years, the issue is far from being resolved to everyone's satisfaction. There is broad agreement that the last eukaryotic common ancestor (LECA) possessed numerous features that are lacking in prokaryotes, including a mitochondrion, a nucleus, an extensive endomembrane traffic system, meiosis, sex, spliceosomal introns, a eukaryotic flagellum, a cytoskeleton, and the like (2, 3). The order of events that gave rise to those attributes is still debated (3–5), as are issues concerning (*i*) the number and nature of prokaryotic partners that were involved in eukaryotic symbioses, (*ii*) the role of gene transfers from the ancestral mitochondrion, and (*iii*) the possible role of lateral gene transfer (LGT) from donors that were distinct from the mitochondrial (or plastid) endosymbiont, or its host.

Three recent developments have shed new light on the problem of eukaryote origins. The first is the insight that the host for the origin of eukaryotes is now best understood as a garden-variety archaeon, one that branches within the diversity of known archaeal lineages (4, 6–9). An origin of the host from within the TACK superphylum (4, 7, 9) is the position most widely discussed at present, but the TACK superphylum was itself only recently recognized through the discovery of new archaeal lineages (7). It is possible that, as new archaeal lineages become discovered, the phylogenetic arrangement of eukaryotes and archaea might undergo further adjustments still (10).

A second development is the recognition that the origin of the roughly 2,000 gene families that underpinned the origin of eukaryotic-specific traits in the eukaryote ancestor required the biochemical power of internalized bioenergetic membranes that mitochondria provided (3). Mitochondria, not oxygen, made the energetic difference that separates eukaryotes from prokaryotes. That is because anaerobic mitochondria generate about five ATP per glucose and fermentations in eukaryotes generate two to four ATP per glucose (11), such that the meager 5- to 10-fold increase in ATP yield per glucose conferred by oxygen respiration is dwarfed by the 10^4 to 10^5 increase in ATP yield per gene manifest in cells with mitochondria (3). The key to the orders of magnitude increase in energy available for evolutionary invention that mitochondria conferred is the eukaryotic configuration of internal, compartmentalized bioenergetic membranes relative to genes (3, 5). After all, had oxygen been the key to eukaryote complexity, *Escherichia coli* would have become eukaryotic for the same reason. Furthermore, eukaryotic aerobes and anaerobes interleave across eukaryote phylogeny (11), and bioenergetics point to a mitochondrion ancestor with a facultatively anaerobic lifestyle (12). Only those cells became complex that experienced the increased energy per gene afforded by mitochondria, and the long puzzling lack of true intermediates in the prokaryote–eukaryote transition has a bioenergetic cause (3).

A third, and more involved, development is the recognition of genomic chimerism in eukaryotes (13), an issue that has been brewing for some time (13–22). Genome analyses showed that genes of bacterial origin outnumber genes of archaeal origin in yeast (21) and other eukaryotic genomes (23, 24) by a factor of about 3:1 and that roughly 15–20% of the nuclear genes in photosynthetic eukaryotes are acquisitions attributable to the endosymbiotic origin of plastids from cyanobacteria (25–27).

However, many of the gene acquisitions in photosynthetic eukaryotes do not trace, in gene trees, directly to a cyanobacterial, and thus obviously plastid, origin. Fewer still among the threefold

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Symbioses Becoming Permanent: The Origins and Evolutionary Trajectories of Organelles," held October 15–17, 2014, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Symbioses.

Author contributions: C.K., S.N.-S., M.R., and W.F.M. analyzed data; and C.K., S.N.-S., M.R., S.G., E.H.-C., and W.F.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. J.P.M. is a guest editor invited by the Editorial Board.

[1]C.K. and S.N.-S. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: bill@hhu.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1421385112/-/DCSupplemental.

excess of bacterial genes over archaeal genes in eukaryote genomes trace directly via gene trees to proteobacteria. The excess of bacterial genes in eukaryotes continues to generate new thoughts, new explanations, and debate. There are several different schools of thought on the issue of how the excess of bacterial genes in eukaryotes is best explained. Eukaryotic gene acquisitions from resident organelles (plastids and mitochondria), lateral gene transfers from casual bacterial acquaintances, and pitfalls of inferring eukaryotic gene origins from gene trees alone stand in the foreground.

Unexpected Bacterial Genes in Eukaryotic Genomes

Efforts to explain bacterial genes in eukaryotes that have unexpected branching patterns often involve "supernumerary symbionts," hypothetical cellular partners that are distinct from the mitochondrion or its host but that donated genes to eukaryotes as the only remnant of their ephemeral existence. This idea probably goes back to Zillig et al. (28), who found genes of bacterial origin in *Giardia* long before anyone suspected that it possessed reduced mitochondria (29). Zillig et al. suggested that such genes betray the existence of a bacterial symbiont *incertae sedis* that preceded the origin of mitochondria and that brought extra bacterial genes into the eukaryotic lineage. Gupta and Golding (17) reasoned similarly, as did others (30, 31), who favored the view that the nucleus was an archaeal endosymbiont, which the extra bacterium engulfed, and which became the nucleus. Supernumerary symbionts were thus allied with endosymbiotic theory, but with an important twist that all of the genes that branched "unexpectedly" were attributed to the same supernumerary donor, whereby the expectations were too seldom spelled out (19).

Another school invokes gene acquisition from "food bacteria" (32): that is, the ancestral eukaryote was a phagotroph (33) that fed on bacteria and occasionally incorporated genes so ingested. A different suggestion has it that eukaryotes and archaea are directly descended from actinobacteria, but that the cause of higher sequence similarity in eukaryote–bacterial comparisons stems from cataclysmic elevation of the substitution rate in archaea, which are however suggested to have arisen about 800 My ago (33), despite evidence that archaea are far more ancient (34). De Duve argued that the host for the origin of mitochondria was a bacterium, the archaeal genes (and ribosomes) of eukaryotes having been acquired via LGT from archaea (35). More recent is Gray's "premitochondrial hypothesis" (36), which posits that mitochondrial proteins that do not branch with alphaproteobacterial homologues are relicts from a premitochondrion that existed in the host, although no suggestion is offered for why the host had bacterial genes to begin with (they are just "there"), nor is the existence or origin of bacterial proteins in the eukaryotic cytosol addressed.

Similar to the situation for the eukaryote common ancestor, the plant lineage was also found to harbor many nuclear genes whose gene distributions—shared only by plants and prokaryotes—strongly suggest that they are acquisitions via endosymbiotic gene transfer from the plastid ancestor even though they do not all branch with cyanobacteria in phylogenetic trees (25, 37). Other suggestions have appeared in the literature to address the excess plant-specific bacterial genes. The shopping bag model (38) was introduced to explain the observation that plant nuclear genes acquired from plastids do not all branch with the same cyanobacterial donor (25). In a nutshell, the shopping bag model invokes a different donor bacterium for every gene that does not branch as expected although the expectation is not explicitly formulated. In that respect it is similar to Doolittle's food bacteria theory (32) for eukaryotic heterotrophs. At the same time, it entails a distinctly gradualist view of endosymbiotic theory: that is, the gradual accumulation of genes in preparation for obtaining a plastid, such that the actual acquisition of a plastid was a small final step in a long process preparing the host for its endosymbiont, an element that is also contained in Gray's premitochondrion theory (36). A problem with the shopping bag model is that acquired nuclear genes for plastid functions are quite useless for a host that has neither a plastid nor a TIC/TOC protein routing machinery to direct nuclear encoded gene products to the plastid should it finally acquire one, such that gene acquisitions before the acquisition of the plastid itself would hardly have a selectable function and would thus be more likely to be lost than be fixed.

Inherited Chimerism: Cutting Trees a Bit of Slack

As an alternative to supernumerary symbionts, perhaps the too many bacterial genes in eukaryotes are acquisitions, by an archaeal host, via gene transfer from the mitochondrion itself (39), whereby the excess of bacterial genes that do not tend to branch with any bacterial group in particular, including alphaproteobacteria, is best explained as gene acquisitions from the mitochondrion followed by LGT among prokaryotes, in addition to the many technical shortcomings of deep phylogeny (40). In that view, the localization of bacterial proteins in the cytosol of nonphotosynthetic eukaryotes comes mainly from endosymbiotic gene transfer out of the mitochondrion to the host before the origin of a mitochondrial protein import apparatus, giving rise to bacterially related cytosolic proteins encoded by nuclear genes of mitochondrial origin (19, 39, 41). With the advent of the mitochondrial protein import machinery, and some gene tinkering in the nucleus, the same transfer mechanism could also give rise to nuclear encoded mitochondrial proteins. That view, termed here "inherited chimerism," has stressed two main aspects: (*i*) we cannot take single-gene phylogenies that span over a billion years back to the origin of mitochondria (and plastids) at face value; we need to be skeptical of their topologies, especially at the deepest branches; and (*ii*) LGT among prokaryotes complicates things in a manner too seldom appreciated, in that genes acquired via the mitochondrion and the plastid were sequestered in the eukaryotic lineage whereas their homologues in prokaryotes were free to continue undergoing recombination, within and across taxon boundaries (21, 40, 42–44). Pangenomes, which arise from the mechanisms of inheritance in prokaryotes, play an underappreciated role in this issue, as the following brief consideration of recombination in prokaryotes and eukaryotes illustrates.

Prokaryotes vs. Eukaryotes, Pangenomes vs. Lineages

Differences in the mechanisms of inheritance across the prokaryote–eukaryote divide generate, over long time frames, different patterns of variation. In both prokaryotes and eukaryotes, there are clonally propagating species that seem never to undergo recombination. Because mutation is inevitable (45), prokaryotic or eukaryotic species that never undergo recombination will continuously accumulate sublethal mutations, which they cannot purge from their genomes. This process continuously increases genetic load, for which reason they will eventually go extinct, a process known as Muller's ratchet (46–49). Recombination has an important role in evolution in that it rescues genomes from Muller's ratchet.

In prokaryotes, three main mechanisms of recombination introduce new genes or alleles into the genome to counteract Muller's ratchet: conjugation, transduction, and transformation (50), in addition to other mechanisms that are restricted to only some lineages, such as gene-transfer agents (51). Over evolutionary timescales, these mechanisms are superimposed upon the clonal patterns of variation that prokaryotic cell division produces (52), leading to a continuous increase in genome size that eventually must be counterbalanced by gene losses and results in clonally descended clusters of sequences that differ substantially in gene content (Fig. 1*A*). The genes shared by all members of the group are called the core genome, those differentially present across the genomes in question are called the dispensable or

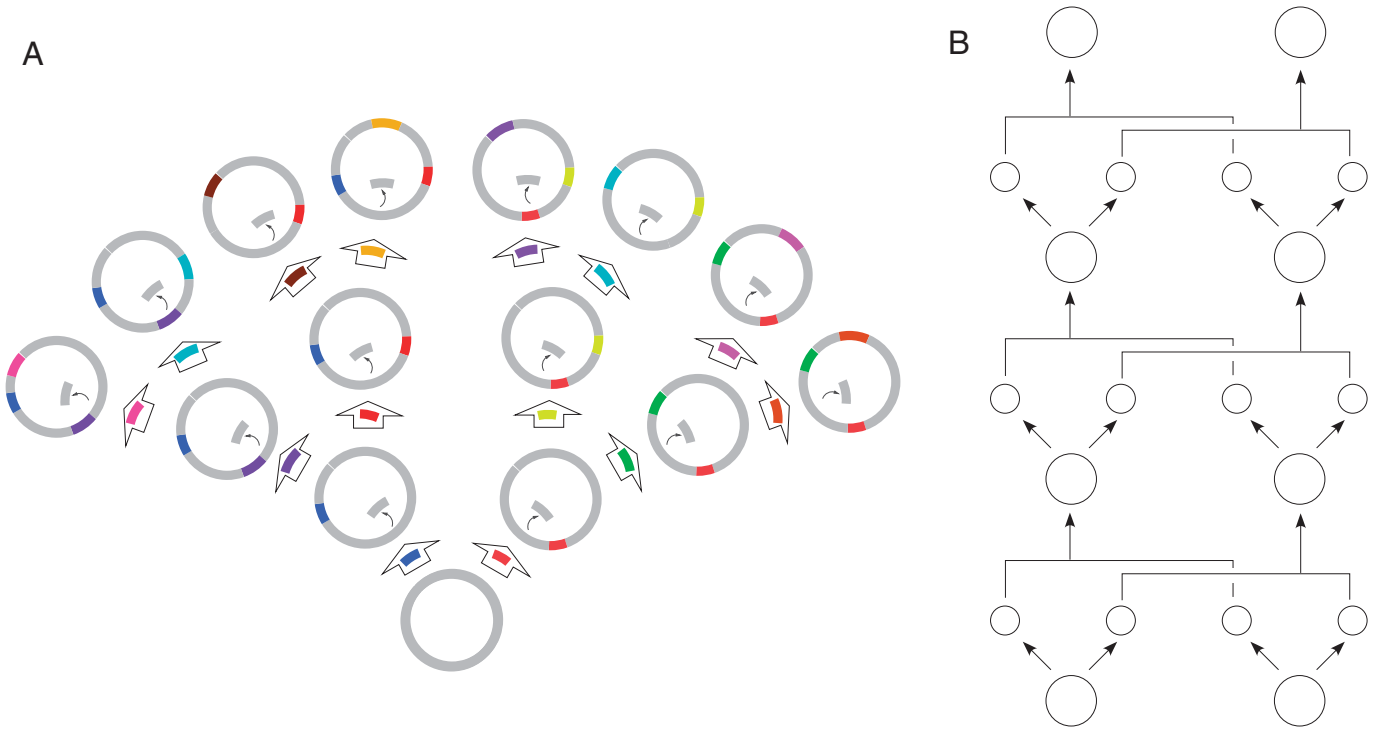


Fig. 1. Recombination and inheritance in prokaryotes and eukaryotes. (*A*) Gene transfer in prokaryotes leads to new genes in different clonally propagating lines. Gene gain (colored segments) is counterbalanced by differential loss. (*B*) Recombination and gamete fusion in eukaryotes (highly schematic) lead to vertically evolving lineages.

accessory genome, and the sum of these components is called the pangenome (53, 54). Importantly, recombination in prokaryotes is not reciprocal, but unidirectional from donor to acceptor, even in archaea that fuse (55). Furthermore, the donor DNA need not come from individuals of the same species; rather, it can come from any taxon or it can even come from dead cells (the environment) (49).

In eukaryotes, the mechanism that counteracts Muller's ratchet is sex. Although there are many variations on the theme (56–59), the underlying principle is that gametes containing different combinations of genes from the same species fuse to produce individuals containing two sets of chromosomes harboring variants (alleles) of the same genes. Meiotic recombination generates new assortments of alleles in the next generation of gametes. Notwithstanding the occasional hybridization, allopolyploidization, or introgression events among closely related species, the process of recombination in eukaryotes produces lineages and patterns that reflect, over geological timescales, vertical descent and new combinations of alleles from within the same gene set (Fig. 1*B*).

It is noteworthy that the mechanisms of recombination in prokaryotes are simultaneously the mechanisms of LGT. Their operation upon clonal lineages over time produces pangenomes whereas the mechanisms of recombination in eukaryotes produce lineages with vertical inheritance. LGT in prokaryotes is just natural variation in action, and microbiologists have always known that there was something like a pangenome out there for prokaryotes because they built 70% DNA–DNA hybridization into the species definition (60, 61), fully aware that 70% hybridization meant 70% shared DNA sequences, not 30% sequence divergence (62).

What Do Pangenomes Look Like?

Pangenomes are collections of genes within the species (or within any taxon) that are or are not uniformly or universally distributed across individual genomes (53), as shown in Fig. 2, where we display the distribution of genes for 54 *E. coli* genomes (Fig. 2*A*), 44 cyanobacterial genomes (Fig. 2*B*), and 208 alphaproteobacterial genomes (Fig. 2*C*). Note that the basic nature of the gene distribution is the same at the species and at the phylum or class level, except for larger numbers of genes at the higher levels, which result from the mechanism in Fig. 1*A* working for greater amounts of time.

Fig. 2 shows only how the genes are shared within the taxa whereas Fig. 3 shows how the genes are distributed across taxa, which is also relevant for the issue of inherited chimerism. This effect is seen for cyanobacteria in Fig. 3*A* and for alphaproteobacteria in Fig. 3*B*. The vast majority of genes found either in this sample of cyanobacteria or in this sample of alphaproteobacteria are not specific to the taxonomic group. Rather, they are shared with other groups. However, they are not shared with all other groups because only about 33 protein-coding genes are universal to all genomes (67), the rest being distributed in some manner. How specifically they are distributed goes beyond the scope of this paper, but it is clear that the distributions mainly entail network-like patterns of sharing (68–70), not tree-like patterns of inclusive hierarchy. The point is this: Were we to reenact endosymbiosis today and allow one of the cyanobacteria in Fig. 3*A* to become the plastid, we would be selecting and sequestering a genome-sized sample of the cyanobacterial pangenome. By putting it into the eukaryotic lineage, we would not affect the ability of the genes shared by the new plastid ancestor and other taxa to undergo LGT and reassortment among the free-living species. If we allow many genes to be relocated to the nucleus while the free-living prokaryotes undergo recombination for the next 1.5 billion years (roughly the age of plastid origin) (71), we might end up with the situation we observe for plants today: Many or most genes that came in with our new plastid will not branch with homologs from a particular cyanobacterial lineage, even if our gene phylogeny is artifact-free. We repeat the experiment for one of the alphaproteobacteria in Fig. 3*B*, which becomes our new mitochondrion, but this time we wait for ~1.8 Ga (roughly

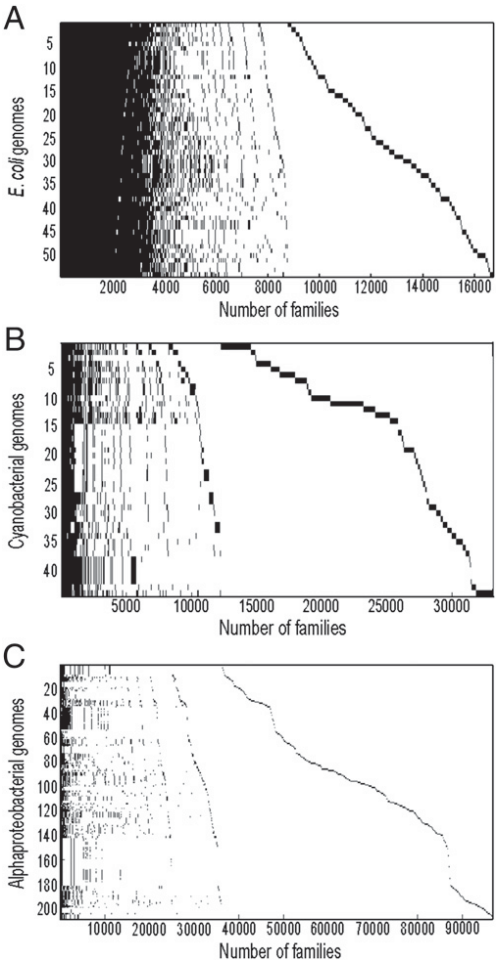


Fig. 2. Bacterial pangenome distribution. The bidirectional best BLAST hit approach (63) was performed on protein sequences of 1,981 complete prokaryotic genomes [see Nelson-Sathi et al. (64) for the full list] that had hits with ≥25% local identity and e-value $<10^{-10}$ in BLAST (65) search. Grouping into protein families was performed using the Markov Chain clustering procedure (66). Patterns of presence (black) and absence (white) of all protein-coding genes are shown. Each genome is represented by a row and gene families by columns. Gene families are sorted in decreasing order (left to right) of their presence in the total genomes. The core genes are shown on the left side and genome-specific genes to the right. (*A*) Distribution of 16,725 genes in 54 *E. coli* genomes, with 8,776 (52.5%) clusters present in at least two genomes and 7,949 (47.5%) unique to individual strains (singletons). Among the singletons, 1,132 genes have at least one homolog in non-*E. coli* species. (*B*) Distribution of 33,118 genes in 44 cyanobacterial genomes, including 12,236 found in at least two genomes and 20,882 singletons. (*C*) A total of 96,916 genes are present in 208 alphaproteobacterial genomes, including 36,176 in at least 2 genomes and 60,740 singletons.

the age of LECA) (71): Many, or even most genes that came in with our new mitochondrion will not branch with a particular alphaproteobacterial lineage, even if our gene trees are free of phylogeny-reconstruction artifacts.

Supernumerary Symbionts or Inherited Chimerism?

Directly from the forest of trees for the excess bacterial genes in eukaryotic genomes, a different category of supernumerary symbionts has emerged that might be called supernumerary phylobionts because their existence is inferred exclusively from phylogenetic trees—trees in which the nearest neighbor of a eukaryotic gene is inferred as the donor. Phylobionts arise directly from observations in gene trees, without independent evidence, and as such their existence and nature are subject to all of the vagaries of phylogenetic methods and lineage sampling. Examples of supernumerary phylobionts include the idea of a supernumerary chlamydiae symbiont that has been repeatedly claimed to have helped the cyanobacterial ancestor of plastids to make the transition from endosymbiont to organelle (72), or various gene-donating bacteria that supposedly helped plants conquer the land (73). The chlamydiae helper symbiont (72, 74) and other hypotheses that summon supernumerary phylobionts from trees are problematic (75, 76)—if we think things through in full, supernumerary phylobionts entail the inference of an additional supernumerary partner for every eukaryotic nuclear gene with prokaryotic homologs, of which there are thousands in eukaryotic genomes (24, 25, 27). As our sample of prokaryotic genomes grows, and as phylogenetic methodologies evolve, it is already evident that, for every eukaryotic gene family, there will eventually be a new and different sister group in phylogenetic trees, and each tree could give rise to some story. In the framework of supernumerary phylobionts, this reasoning will lead to thousands of individual gene donors to the eukaryotic ancestor and the archaeplastidan ancestor. That proposition is untenable. How so? An example illustrates.

What would happen if we were to use the same methodology—single-gene trees—as people have been using to infer the origins of eukaryotic nuclear genes to infer the origin of genes that are still present in the mitochondrion or the plastid? To see, we constructed alignments and single-gene maximum likelihood trees (see *SI Text* for the detailed methods) for those 51 (out of 67) protein-coding genes from the *Reclinomonas americana* mitochondrial genome (77) that are sufficiently well-conserved to make trees and the best conserved 183 out of 209 protein-coding genes in the *Porphyra purpurea* plastid genome (78) in the context of 1,981 prokaryotic genomes (64). The results (Dataset S1 and Figs. S2 and S3) show that, for *Reclinomonas*, 43 different sister groups were obtained, and, in 20 cases, the mitochondrial sister group differs in trees based on the forward and reverse alignments (79) using the same algorithm (Fig. S2). For the *Porphyra* plastid proteins, 124 different sister groups were obtained, and, in 52 cases, the plastid sister group is different in the reverse-alignment trees (Fig. S3).

Using the logic germane to supernumerary phylobiont inference, the findings in Dataset S1 and Figs. S2 and S3 would be interpreted as evidence that neither the mitochondrion nor the plastid arose via endosymbiosis; rather, each would be the product of 43 and 124 independent gene transfers, respectively, from different donors, thus one at a time, to the eukaryotic ancestor and the archaeplastidan ancestor, but the transfers would have to be directed to some kind of preexisting compartment, not dissimilar to Gray's premitochondrion, where rRNA operons and tRNAs also became donated, enabling the result of such transfer to morph into a bioenergetic organelle, but only mimicking a bona fide endosymbiotic origin, the real mechanism being LGT: So say the single-gene trees. We say: That scenario cannot possibly be true. However, why can it not be true? It cannot be true because exactly the same kinds of transfers—one at a time and from independent donors—for exactly the right kinds of genes to support the function of the bioenergetic membrane in mitochondria and the bioenergetic membrane in plastids (in addition to the other biochemical and physiological functions of the organelles) would have to be going on to the nucleus as well, the crux being that, until the whole organelle is assembled through such imaginary LGT, none of the transferred genes have a selectable function. Without selection for function, they would all become pseudogenes, and no organelle would emerge at all. A free-living prokaryote brings along the complete and selectable functional unit, which can then be transferred a chunk at a time to the host, but from a continuously selected and replicating functional source. There is something very wrong with the supernumerary phylobiont stories, and the core of the problem is rooted in trees.

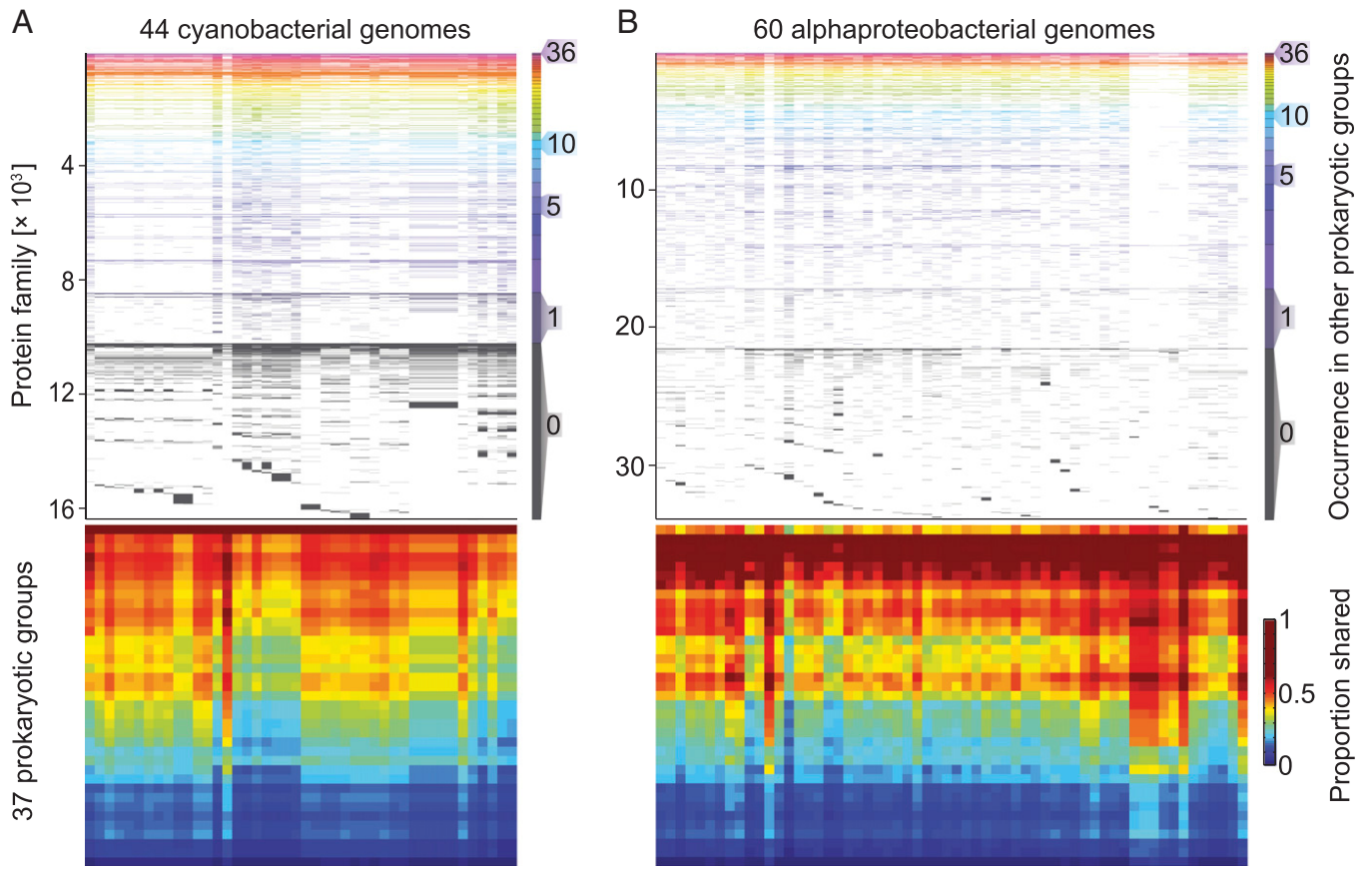


Fig. 3. Gene sharing among prokaryotes. (*A*, *Upper*) Presence/absence of protein families in cyanobacterial genomes, sorted according to the number of taxonomic groups sharing the corresponding gene. (*Lower*) Proportion of genes in each cyanobacterial genome shared with other taxonomic groups. *k*-means clustering was applied to sort taxonomic groups according to pattern similarity. (*B*) Showing the same patterns as in *A* for alphaproteobacteria. (*Lower*) Taxonomic groups are sorted as in *A*. This figure is based on the nonsingleton clusters from those described in Fig. 2. For the complete figure with taxon labels, see Fig. S1.

We have to relax our expectations regarding the ability of single-gene trees to provide a crayon with which we can draw eukaryotic genome history. If we take gene trees at face value, we would have to reject the proposition that plastids and mitochondria descend via endosymbiosis from free-living prokaryotes in favor of a biochemically untenable view of single-gene assembly based on LGTs inferred from gene trees. Endosymbiosis is clearly the better supported alternative, whereby inherited chimerism is a corollary whose function is to help explain odd branches in gene trees so that we do not throw out the baby (endosymbiotic theory) with the bathwater (gene trees). Concatenation is the answer, some might say, but concatenation for prokaryotic genes is very problematic (80), and, if we do combine the eukaryotic trees into categories justifiable by tree-independent methods, what we find is evidence for a plastid, a mitochondrion, and an archaeal host (81).

Where Can We Go Wrong with Trees and Where Will It Lead Us?

Asking all genes that came into the eukaryotic lineage via the mitochondrial and plastid symbiosis, respectively, to branch with homologs from one and the same present-day proteobacterial and one and the same present-day cyanobacterial genome is simply asking too much. If we adopt a vertical, static view of prokaryotic genome evolution, where genes in a prokaryotic lineage can be passed down only within the lineage, then a tree of an endosymbiotically acquired gene would always show a prokaryotic sister group to eukaryotes that consists of only taxa from the lineage to which the organelle belongs, the true donor lineage (Fig. 4*A*). Because of LGT among prokaryotes, however, the prokaryotic homologs of eukaryotic genes almost never show the prokaryotic groups to be monophyletic (Fig. 4*B* and table 1 from ref. 24). Add to that gene loss [which has to be as common in gene evolution as LGT; otherwise genomes would constantly be expanding (82, 83)] and incomplete prokaryote genome sampling, which results in a tree where the prokaryotic sister group is sparsely populated by (Fig. 4*C*) or sometimes even without any representative from the true sister group (Fig. 4*D*). The gene donors we infer from trees today are thus ephemeral (Fig. 4 *B*–*E*). For example, the first plant–chlamydiae gene connection was the plastid ATP/ADP translocase (84), which, until 2007 (72), was found only in Rickettisales, and sparked heated debate on its origin (85). As of November 2014, the *Arabidopsis* plastid ATP/ADP translocase (NP_173003) detects homologs in alphaproteobacteria outside Rickettsiales, in beta-, gamma- and deltaproteobacteria, and in bacteroidetes (Table S1). In 2021, there will be more. These factors (Fig. 4) are sufficient to generate patterns of apparent transfer from prokaryotes to eukaryotes, not to mention tree-building artifacts (38) that can also produce trees showing apparent gene transfer (86). By ignoring such factors, and by naively believing trees at face value, a view is emerging that LGT, not endosymbiosis, is the main mechanism behind the origin of plastids (87, 88). Should we believe that?

Gene Transfer from Organelles to the Nucleus: At Least It's Real

If LGT from prokaryotes to eukaryotes were really as common in genome evolution as such studies would have us believe, then eukaryotic chromosomes should be replete with recently acquired

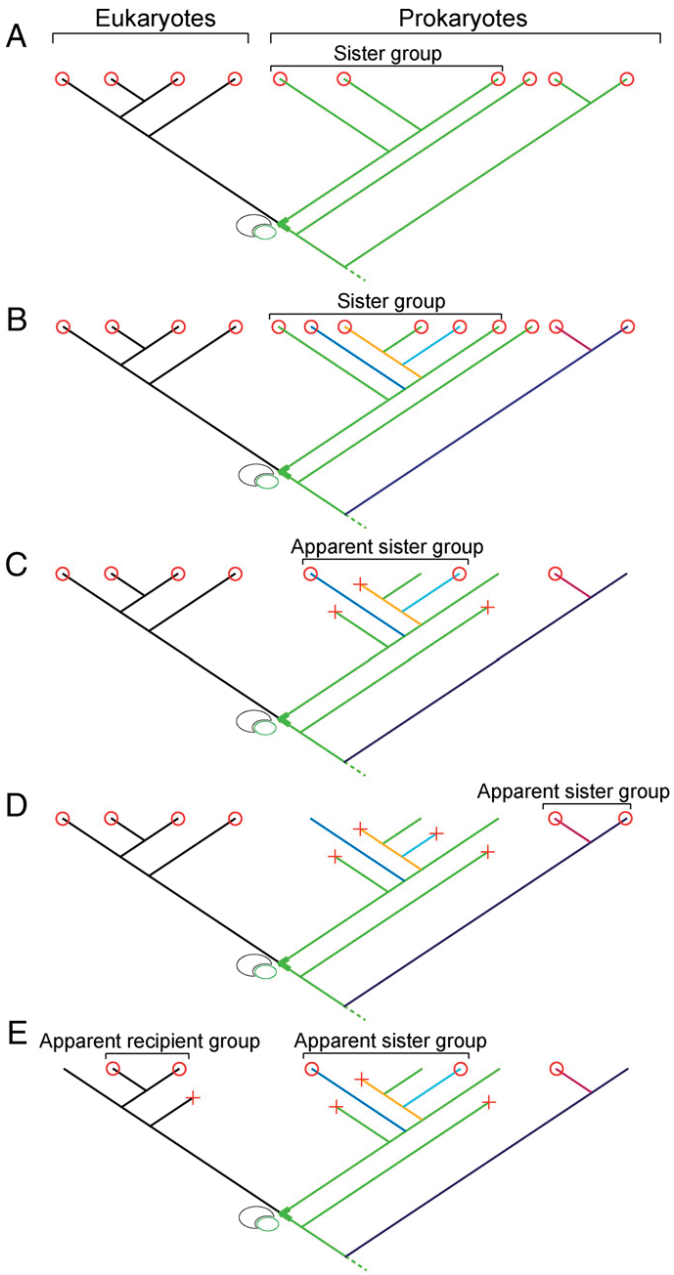


Fig. 4. Histories hidden behind trees. (*A*) In an ideal tree of a gene acquired endosymbiotically from a donor prokaryotic lineage (green), eukaryotes (black) should be nested within present-day representatives of that lineage. (*B*) LGT among prokaryotes results in a prokaryotic sister group consisting of homologs from both donor and nondonor lineages (nongreen colors). Further complicated by gene loss (crosses) and incomplete sampling (only circled homologs are sampled and used for phylogenetic analyses), the sister group observed in the tree is an apparent one that is a subsample of the complete sister group (*C*) or does not contain any representative from the true sister group (*D*). (*E*) The same factors also influence sampling of eukaryotic homologs, resulting in an apparent acquisition of the gene by a subgroup of the eukaryotic clade involved in the endosymbiosis event.

bacterial DNA. However, bona fide recent bacterial gene acquisitions are very rare, and most—but not all—of the bacterial sequences that are reported in genome-sequencing projects are ultimately removed from the databases because they are contaminations from the genome-sequencing process. Important exceptions are the genomes of phloem-feeding insects, which are regularly found to harbor insertions of bacterial DNA that stems from the obligate bacterial endosymbionts that grow in the bacteriome, a specialized organ that houses the symbionts, which provide essential functions to their host, most commonly amino acid biosynthesis. Genome sequences of pea aphids (89), mealybugs (90), psyllids (91), and invertebrates infected by *Wolbachia* (92) have revealed DNA segments that have been integrated from endosymbionts. However, such recent DNA transfers from bacteria are generally quite rare in eukaryotes, which is probably why they get so much attention when such verified cases are reported.

By comparison, the transfer of DNA from organelles to the nuclear genome is ubiquitous among eukaryotic genomes. DNA transfer from organelles to the nucleus occurs in all eukaryote genomes studied to date (93). *Numts*, for nuclear mitochondrial DNA copies (and *nupts* for the plastid) (94), are typical components of eukaryotic genomes (93–95) whereas segments of bacterial chromosomes are not. For example, our genomes harbor 53 *numts* that are specific to the human lineage (96), with 12 *numts* that are polymorphic in human populations (93), and more *numts* continuously being found in the human 1,000 Genomes data (97). Five human *numts* are associated with disease (93), one of which involves a 72-bp numt insertion into exon 14 of the GLI3 gene, causing a premature stop codon, in a rare case of Pallister–Hall syndrome stemming from the Chernobyl incident (98). No human genomes are (yet) known to be polymorphic for recent bacterial DNA insertions.

The mechanism of gene transfer from organelles to the nucleus entails the incorporation of bulk organelle DNA into nuclear chromosomes. Very large copies can be inserted, as the 262-kb mtDNA of *Arabidopsis* (99.91% identical) and the 131-kb complete rice chloroplast genome (99.77% identical) attest (99), suggesting that, during the early phases of organelle origins, large segments or even whole chromosomes were also being transferred, followed by the normal DNA dynamics of mutation, recombination, fixation, and deletion. *Numts* and *nupts* are inserted into double-strand breaks by the nonhomologous end-joining machinery (100, 101) and enter the genome in open chromatin regions (101, 102). *Numts* can be integrated into chromosomes with a short microhomology of 1–7 bp, implicating a submechanism of nonhomologous end joining known as microhomology-mediated repair (103), but insertion can also occur without microhomology—a process known as blunt-end repair.

Analysis of 90 recent *numt* insertions in human and chimpanzee suggests that 35% of the fusion points involve microhomology of at least 2 bp; thus, it seems that repair involving microhomology plays some role in *numt* integration but is not strictly required (103). No analyses of recent insertions of bacterial DNA into the human and chimpanzee lineages have been reported. Notwithstanding the cases of plant-feeding insects and their tightly associated bacteria, why we do not observe recent bacterial transfers, as we do for *numts* and *nupts*? And if all of the prokaryote-to-eukaryote LGT reports are real, then, at some point, we need to see evidence for its long-term effects in terms of different lineages of eukaryotes harboring fundamentally different collections of genes, as we see in prokaryotes (64). However, except for photosynthetic eukaryotes, which acquired the plastid and many genes with it, different eukaryotic lineages tend to possess the very same collections of genes having prokaryotic homologs, which is not true for prokaryotes (Fig. 1). We are saying that prokaryotes recombine via LGT but that eukaryotes have remained genetically isolated from prokaryotes (except at the origins of organelles) because they recombine via sex. Our critics will thus ask: Where did sex come from?

Did Sex Rescue the Ancestral Eukaryote from Muller's Ratchet?

Like eukaryotes, the origin of sex also counts as one of the major evolutionary transitions (1) and remains one of evolutionary biology's toughest problems. Existing theories seek the origin of sex

in a haploid cell with fully fledged eukaryotic mitosis (104), but it is more likely that mitosis and sex arose in a cell that had a mitochondrion (3, 5). During the prokaryote-to-eukaryote transition, eukaryotes seem to have lost the standard mechanisms that prokaryotes use to escape Muller's ratchet—transduction, transformation, and conjugation—because they are lacking in all eukaryotic groups. Had eukaryotes retained one or all three of those mechanisms, it seems unlikely that they would have evolved sex on top of them, and, indeed, cells that never had mitochondria (prokaryotes) never evolved sex. The machinery involved in eukaryotic recombination was surely present at the time of mitochondrial symbiosis because the main enzymes involved are homologous to their prokaryotic counterparts: Spo11, Mre11, Dmc1, Rad51, Mlh1, and Pms1 (105, 106). Did a simple form of eukaryotic recombination, catalyzed by enzymes that are homologous to the enzymes of prokaryotic recombination, rescue nascent eukaryotes from Muller's ratchet? The basic machinery required might have been a property of the host. It is a curiously underpublicized observation that various archaea can fuse their cells (55, 107) and that, in some haloarchaea, fusion is accompanied by recombination (108) whereas, in others, only recombination is observed (109). One needs to be careful not to (over-)state that "archaea have sex," but, in some rare documented examples, they do undergo outright cell fusion (an otherwise curious property of gametes) and, in some rarer cases, recombination and fusion are observed (108).

Thus, it could be that the essentials of the machinery required for sex—fusion of cells from the same species and ability to generate recombinants in fused cells—was present in the host lineage that acquired the mitochondrion. Without such a capability, extinction would have been the alternative. That suggestion would help to ease one more evolutionary transition in the origin of eukaryotes (sex), which would go a long way toward explaining the differences between inheritance in prokaryotes and eukaryotes (Fig. 1), without solving the problem in full or explaining (*i*) how mitosis and meiosis are related to one another, (*ii*) where the cell cycle comes from, or (*iii*) why eukaryotes, in contrast to all prokaryotes, shut down their gene expression at cell division. Such longstanding questions concerning the major evolutionary transition at eukaryote origin (1), are arguably more tractable than ever before, given progress concerning the archaeal nature of the host that acquired mitochondria (4, 7, 81).

Conclusion

Inherited chimerism is an alternative to the problematic practice of conjuring up additional, gene-donating symbionts at organelle origins to explain gene trees. It merely requires a selective force to associate the symbiont (either plastid or mitochondrion) to its host so that the endosymbiosis (one cell living within another) can be established and gene transfer from the symbiont can commence. It places no constraints on the collections of genes that the plastid and the mitochondrial symbionts possessed, other than that it needs to be a genome-sized collection, not tens of thousands of genes, and it allows freely for LGT among prokaryotes before the endosymbionts become organelles and afterward. LGT among prokaryotes has received much attention in the past decades. Inherited chimerism incorporates LGT among prokaryotes into endosymbiotic theory.

ACKNOWLEDGMENTS. We thank Eric Bapteste, Nick Lane, and Eörs Szathmáry for input and constructive comments on the text, and we especially thank James O. McInerney for pointing out papers on archaeal recombination and fusion. This work was supported by grants from the European Research Council (Grant 232975) (to W.F.M.) and from the Open University of Israel Research Fund (to E.H.-C.). C.K. is grateful to the Deutscher Akademischer Austauschdienst for a PhD stipend.

1. Szathmáry E, Smith JM (1995) The major evolutionary transitions. *Nature* 374(6519): 227–232.
2. Lane N (2009) *Life Ascending: The Ten Great Inventions of Evolution* (Norton, New York).
3. Lane N, Martin W (2010) The energetics of genome complexity. *Nature* 467(7318): 929–934.
4. Williams TA, Foster PG, Cox CJ, Embley TM (2013) An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504(7479):231–236.
5. Lane N (2014) Bioenergetic constraints on the evolution of complex life. *Cold Spring Harb Perspect Biol* 6(5):a015982.
6. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM (2008) The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci USA* 105(51):20356–20361.
7. Guy L, Ettema TJG (2011) The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol* 19(12):580–587.
8. Kelly S, Wickstead B, Gull K (2011) Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proc Biol Sci* 278(1708):1009–1018.
9. Koonin EV, Yutin N (2014) The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. *Cold Spring Harb Perspect Biol* 6(4):a016188.
10. Williams TA, Embley TM (2014) Archaeal "dark matter" and the origin of eukaryotes. *Genome Biol Evol* 6(3):474–481.
11. Müller M, et al. (2012) Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol Mol Biol Rev* 76(2):444–495.
12. Degli Esposti M (2014) Bioenergetic evolution in proteobacteria and mitochondria. *Genome Biol Evol* 6(12):3238–3251.
13. McInerney JO, O'Connell MJ, Pisani D (2014) The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat Rev Microbiol* 12(6):449–455.
14. Martin W, Cerff R (1986) Prokaryotic features of a nucleus-encoded enzyme: cDNA sequences for chloroplast and cytosolic glyceraldehyde-3-phosphate dehydrogenases from mustard (Sinapis alba). *Eur J Biochem* 159(2):323–331.
15. Martin W, Brinkmann H, Savonna C, Cerff R (1993) Evidence for a chimeric nature of nuclear genomes: Eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. *Proc Natl Acad Sci USA* 90(18):8692–8696.
16. Brinkmann H, Martin W (1996) Higher-plant chloroplast and cytosolic 3-phosphoglycerate kinases: A case of endosymbiotic gene replacement. *Plant Mol Biol* 30(1): 65–75.
17. Gupta RS, Golding GB (1996) The origin of the eukaryotic cell. *Trends Biochem Sci* 21(5):166–171.
18. Brown JR, Doolittle WF (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev* 61(4):456–502.
19. Martin W, Schnarrenberger C (1997) The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: A case study of functional redundancy in ancient pathways through endosymbiosis. *Curr Genet* 32(1):1–18.
20. Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA* 95(11):6239–6244.
21. Esser C, et al. (2004) A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol* 21(9): 1643–1660.
22. Lane CE, Archibald JM (2008) The eukaryotic tree of life: Endosymbiosis takes its TOL. *Trends Ecol Evol* 23(5):268–275.
23. Cotton JA, McInerney JO (2010) Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proc Natl Acad Sci USA* 107(40):17252–17255.
24. Thiergart T, Landan G, Schenk M, Dagan T, Martin WF (2012) An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome Biol Evol* 4(4):466–485.
25. Martin W, et al. (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA* 99(19):12246–12251.
26. Deusch O, et al. (2008) Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol* 25(4):748–761.
27. Dagan T, et al. (2013) Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biol Evol* 5(1):31–44.
28. Zillig W, et al. (1989) Did eukaryotes originate by a fusion event. *Endocyt Cell Res* 6(1):1–25.
29. Tovar J, et al. (2003) Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. *Nature* 426(6963):172–176.
30. Horiike T, Hamada K, Kanaya S, Shinozawa T (2001) Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria is revealed by homology-hit analysis. *Nat Cell Biol* 3(2):210–214.
31. Forterre P, Philippe H (1999) Where is the root of the universal tree of life? *BioEssays* 21(10):871–879.
32. Doolittle WF (1998) You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet* 14(8):307–311.
33. Cavalier-Smith T (2002) The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int J Syst Evol Microbiol* 52(Pt 1):7–76.
34. Liu Y, Beer LL, Whitman WB (2012) Methanogens: A window into ancient sulfur metabolism. *Trends Microbiol* 20(5):251–258.

35. de Duve C (2007) The origin of eukaryotes: A reappraisal. *Nat Rev Genet* 8(5): 395–403.
36. Gray MW (2014) The pre-endosymbiont hypothesis: A new perspective on the origin and evolution of mitochondria. *Cold Spring Harb Perspect Biol* 6(3):a016097.
37. Rujan T, Martin W (2001) How many genes in *Arabidopsis* come from cyanobacteria? An estimate from 386 protein phylogenies. *Trends Genet* 17(3):113–120.
38. Larkum AWD, Lockhart PJ, Howe CJ (2007) Shopping for plastids. *Trends Plant Sci* 12(5):189–195.
39. Martin W, Müller M (1998) The hydrogen hypothesis for the first eukaryote. *Nature* 392(6671):37–41.
40. Martin W (1999) Mosaic bacterial chromosomes: A challenge en route to a tree of genomes. *BioEssays* 21(2):99–104.
41. Martin W (2010) Evolutionary origins of metabolic compartmentalization in eukaryotes. *Philos Trans R Soc Lond B Biol Sci* 365(1541):847–855.
42. Martin WF (1996) Is something wrong with the tree of life? *BioEssays* 18(7):523–527.
43. Martin W, Herrmann RG (1998) Gene transfer from organelles to the nucleus: How much, what happens, and why? *Plant Physiol* 118(1):9–17.
44. Schnarrenberger C, Martin W (2002) Evolution of the enzymes of the citric acid cycle and the glyoxylate cycle of higher plants: A case study of endosymbiotic gene transfer. *Eur J Biochem* 269(3):868–883.
45. Nei M (2013) *Mutation-Driven Evolution* (Oxford Univ Press, Oxford).
46. Muller HJ (1964) The relation of recombination to mutational advance. *Mutat Res* 106(1):2–9.
47. Felsenstein J (1974) The evolutionary advantage of recombination. *Genetics* 78(2): 737–756.
48. Lynch M, Bürger R, Butcher D, Gabriel W (1993) The mutational meltdown in asexual populations. *J Hered* 84(5):339–344.
49. Takeuchi N, Kaneko K, Koonin EV (2014) Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: benefit of DNA from dead cells and population subdivision. *G3 (Bethesda)* 4(2):325–339.
50. Popa O, Dagan T (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol* 14(5):615–623.
51. Lang AS, Zhaxybayeva O, Beatty JT (2012) Gene transfer agents: Phage-like elements of genetic exchange. *Nat Rev Microbiol* 10(7):472–482.
52. Milkman R (1997) Recombination and population structure in *Escherichia coli*. *Genetics* 146(3):745–750.
53. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pangenome. *Curr Opin Genet Dev* 15(6):589–594.
54. Young JPW, et al. (2006) The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol* 7(4):R34.
55. Naor A, Lapierre P, Mevarech M, Papke RT, Gophna U (2012) Low species barriers in halophilic archaea and the formation of recombinant hybrids. *Curr Biol* 22(15): 1444–1448.
56. Bell G (1988) *Sex and Death in Protozoa* (Cambridge Univ Press, Cambridge, UK).
57. Lee SC, Ni M, Li W, Shertz C, Heitman J (2010) The evolution of sex: A perspective from the fungal kingdom. *Microbiol Mol Biol Rev* 74(2):298–340.
58. Parfrey LW, Katz LA (2010) Dynamic genomes of eukaryotes and the maintenance of genomic integrity. *Microbe* 5(4):156–163.
59. Roach KC, Heitman J (2014) Unisexual reproduction reverses Muller's ratchet. *Genetics* 198(3):1059–1069.
60. Goris J, et al. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 57(Pt 1):81–91.
61. Stackebrandt E, Goebel BM (1994) A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* 44(4):846–849.
62. Beltz GA, Jacobs KA, Eickbush TH, Cherbas PT, Kafatos FC (1983) Isolation of multigene families and determination of homologies by filter hybridization methods. *Methods Enzymol* 100:266–285.
63. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278(5338):631–637.
64. Nelson-Sathi S, et al. (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517(7532):77–80.
65. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
66. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7):1575–1584.
67. Hansmann S, Martin W (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: Influence of excluding poorly alignable sites from analysis. *Int J Syst Evol Microbiol* 50(Pt 4):1655–1663.
68. Dagan T, Artzy-Randrup Y, Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA* 105(29):10039–10044.
69. Bapteste E, et al. (2013) Networks: Expanding evolutionary thinking. *Trends Genet* 29(8):439–441.
70. Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO (2013) Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc Natl Acad Sci USA* 110(17):E1594–E1603.
71. Parfrey LW, Lahr DJG, Knoll AH, Katz LA (2011) Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci USA* 108(33):13624–13629.
72. Huang J, Gogarten JP (2007) Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biol* 8(6):R99.
73. Yue J, Hu X, Sun H, Yang Y, Huang J (2012) Widespread impact of horizontal gene transfer on plant colonization of land. *Nat Commun* 3:1152.
74. Ball SG, et al. (2013) Metabolic effectors secreted by bacterial pathogens: Essential facilitators of plastid endosymbiosis? *Plant Cell* 25(1):7–21.
75. Zimorski V, Ku C, Martin WF, Gould SB (2014) Endosymbiotic theory for organelle origins. *Curr Opin Microbiol* 22C:38–48.
76. Deschamps P (2014) Primary endosymbiosis: Have cyanobacteria and Chlamydiae ever been roommates? *Acta Soc Bot Pol* 83:291–302.
77. Lang BF, et al. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* 387(6632):493–497.
78. Reith M, Munholland J (1995) Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. *Plant Mol Biol Rep* 13(4):333–335.
79. Landan G, Graur D (2007) Heads or tails: A simple reliability check for multiple sequence alignments. *Mol Biol Evol* 24(6):1380–1383.
80. Bapteste E, et al. (2008) Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. *Mol Biol Evol* 25(1):83–91.
81. Pisani D, Cotton JA, McInerney JO (2007) Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol* 24(8):1752–1760.
82. Wolf YI, Koonin EV (2013) Genome reduction as the dominant mode of evolution. *BioEssays* 35(9):829–837.
83. Dagan T, Martin W (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* 104(3): 870–875.
84. Stephens RS, et al. (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* 282(5389):754–759.
85. Schmitz-Esser S, et al. (2004) ATP/ADP translocases: A common feature of obligate intracellular amoebal symbionts related to Chlamydiae and Rickettsiae. *J Bacteriol* 186(3):683–691.
86. Moreira D, Deschamps P (2014) What was the real contribution of endosymbionts to the eukaryotic nucleus? Insights from photosynthetic eukaryotes. *Cold Spring Harb Perspect Biol* 6(7):a016014.
87. Reyes-Prieto A, Moustafa A (2012) Plastid-localized amino acid biosynthetic pathways of Plantae are predominantly composed of non-cyanobacterial enzymes. *Sci Rep* 2:955.
88. Qiu H, et al. (2013) Assessing the bacterial contribution to the plastid proteome. *Trends Plant Sci* 18(12):680–687.
89. Richards S, et al.; International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol* 8(2):e1000313.
90. Husnik F, et al. (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* 153(7):1567–1578.
91. Sloan DB, et al. (2014) Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Mol Biol Evol* 31(4):857–871.
92. Dunning Hotopp JC, et al. (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 317(5845):1753–1756.
93. Hazkani-Covo E, Zeller RM, Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet* 6(2):e1000834.
94. Kleine T, Maier UG, Leister D (2009) DNA transfer from organelles to the nucleus: The idiosyncratic genetics of endosymbiosis. *Annu Rev Plant Biol* 60:115–138.
95. Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 5(2):123–135.
96. Lang M, et al. (2012) Polymorphic NumtS trace human population relationships. *Hum Genet* 131(5):757–771.
97. Dayama G, Emery SB, Kidd JM, Mills RE (2014) The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Res* 42(20): 12640–12649.
98. Turner C, et al. (2003) Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. *Hum Genet* 112(3):303–309.
99. Huang CY, Grünheit N, Ahmadinejad N, Timmis JN, Martin W (2005) Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiol* 138(3):1723–1733.
100. Ricchetti M, Tekaia F, Dujon B (2004) Continued colonization of the human genome by mitochondrial DNA. *PLoS Biol* 2(9):E273.
101. Lloyd AH, Timmis JN (2011) The origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome. *Mol Biol Evol* 28(7):2019–2028.
102. Wang D, Timmis JN (2013) Cytoplasmic organelle DNA preferentially inserts into open chromatin. *Genome Biol Evol* 5(6):1060–1064.
103. Hazkani-Covo E, Covo S (2008) Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genet* 4(10):e1000237.
104. Wilkins AS, Holliday R (2009) The evolution of meiosis from mitosis. *Genetics* 181(1): 3–12.
105. Malik SB, Pightling AW, Stefaniak LM, Schurko AM, Logsdon JM, Jr (2008) An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*. *PLoS ONE* 3(8):e2879.
106. Hörandl E, Hadacek F (2013) The oxidative damage initiation hypothesis for meiosis. *Plant Reprod* 26(4):351–367.
107. Kuwabara T, et al. (2005) *Thermococcus coalescens* sp. nov., a cell-fusing hyperthermophilic archaeon from Suiyo Seamount. *Int J Syst Evol Microbiol* 55(Pt 6): 2507–2514.
108. Naor A, Gophna U (2013) Cell fusion and hybrids in Archaea: Prospects for genome shuffling and accelerated strain development for biotechnology. *Bioengineered* 4(3):126–129.
109. Papke RT, Koenig JE, Rodríguez-Valera F, Doolittle WF (2004) Frequent recombination in a saltern population of *Halorubrum*. *Science* 306(5703):1928–1929.

Supporting Information

Ku et al. 10.1073/pnas.1421385112

SI Text

Taxonomic Distribution of Prokaryotic Sister Taxa to Organelle-Encoded Genes. Protein sequences of the *Reclinomonas americana* mitochondrial (accession no. NC_001823; 67 protein-coding genes) and the *Porphyra purpurea* plastid (accession no. NC_000925; 209 protein-coding genes) genomes were downloaded from the National Center for Biotechnology Information (NCBI) GenBank and blasted (1) against 1,981 completely sequenced prokaryotic genomes [RefSEq. 2012 dataset; see Nelson-Sathi et al. (2) for a complete list]. Prokaryotic homologs of mitochondrial and plastid genes were identified using a combination of blastp identity (30%), e-value ($1 \times e^{-10}$) and query coverage (70%) as thresholds. To avoid redundancy, global identities were calculated using the needle program in the EMBOSS package (3), and identical or nearly identical (>90% global identity) sequences from the same species were removed from the data. Organelle genes with at least four prokaryotic homologs were aligned using MAFFT 7.130 (4) with parameters "–localpair –maxiterate 1000". Maximum likelihood trees were reconstructed using RAxML 7.8.6 (5) under the PROTCATWAG model, except for the mitochondrial gene, *nad10*, which had a special character (U) in a prokaryotic homolog that was not recognizable by RAxML. The nearest neighbor of the query organelle sequence consists of the prokaryotic sequences in the smallest tree bipartition that includes the query and at least one prokaryotic sequence. To test the effect of alignment quality (6, 7) on the inference of nearest neighbors, the above procedure was repeated for reversed sequences. The prokaryotic taxa occurring in the nearest neighbors are summarized in Dataset S1, in Fig. S2 for the mitochondrial genome, and in Fig. S3 for the plastid genome. Gene trees are available in Dataset S1.

1. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
2. Nelson-Sathi S, et al. (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517(7532):77–80.
3. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 16(6):276–277.
4. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 30(4):772–780.
5. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
6. Landan G, Graur D (2007) Heads or tails: A simple reliability check for multiple sequence alignments. *Mol Biol Evol* 24(6):1380–1383.
7. Deusch O, et al. (2008) Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol* 25(4):748–761.

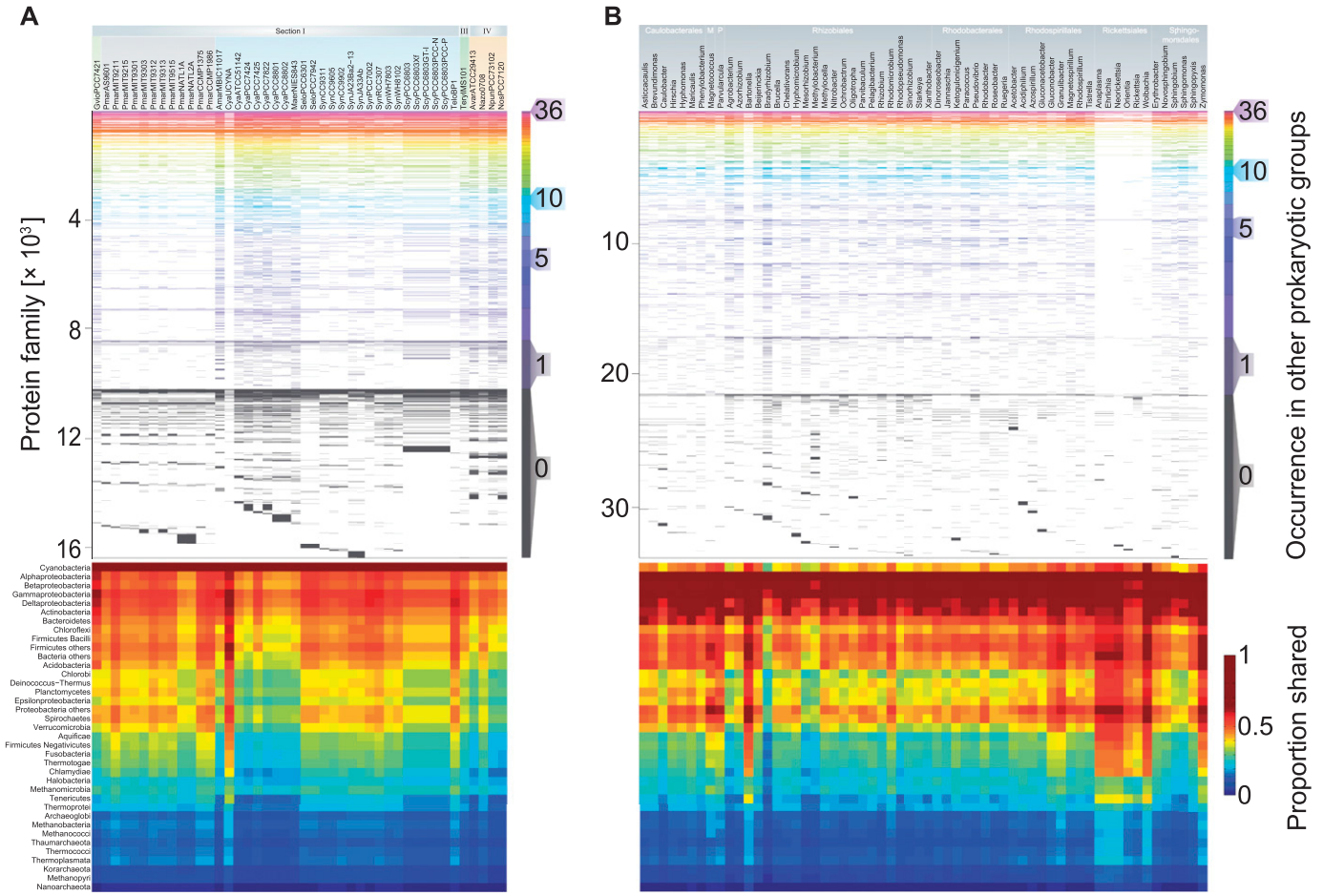


Fig. S1. Gene sharing among prokaryotes, showing detailed taxon labels for Fig. 3. (*A*, *Upper*) Presence of genes in cyanobacterial genomes. Protein families are sorted according to the number of taxonomic groups sharing the corresponding gene. (*A*, *Lower*) Proportion of genes in each cyanobacterial genome shared with other taxonomic groups. *k*-means clustering was applied to sort the groups according to pattern similarity. Gvio, *Gloeobacter violaceus*; Pmar, *Prochlorococcus marinus*; Amar, *Acaryochloris marina*; Cya, *Cyanothece*; Maer, *Microcystis aeroginosa*; Selo, *Synechococcus elongatus*; Syn, *Synechococcus*; Scy, *Synechocystis*; Telo, *Thermosynechococcus elongatus*; Tery, *Trichodesmium erythraeum*; Avar, *Anabaena variabilis*; Nazo, *Nostoc azollae*; Npun, *Nostoc punctiforme*; Nos, *Nostoc* sp. (*B*, *Upper*) Presence of genes in 60 alphaproteobacterial genera (the first genome in each genus in the alphanumerically sorted list was used as the representative). Protein families are sorted according to the number of taxonomic groups sharing the family. (*B*, *Lower*) Proportion of genes in each alphaproteobacterial genome shared with other taxonomic groups, which are sorted as in *A*. M, Magnetococcales; P, Parvularculales. This figure is based on the nonsingleton clusters from those described in Fig. 2.

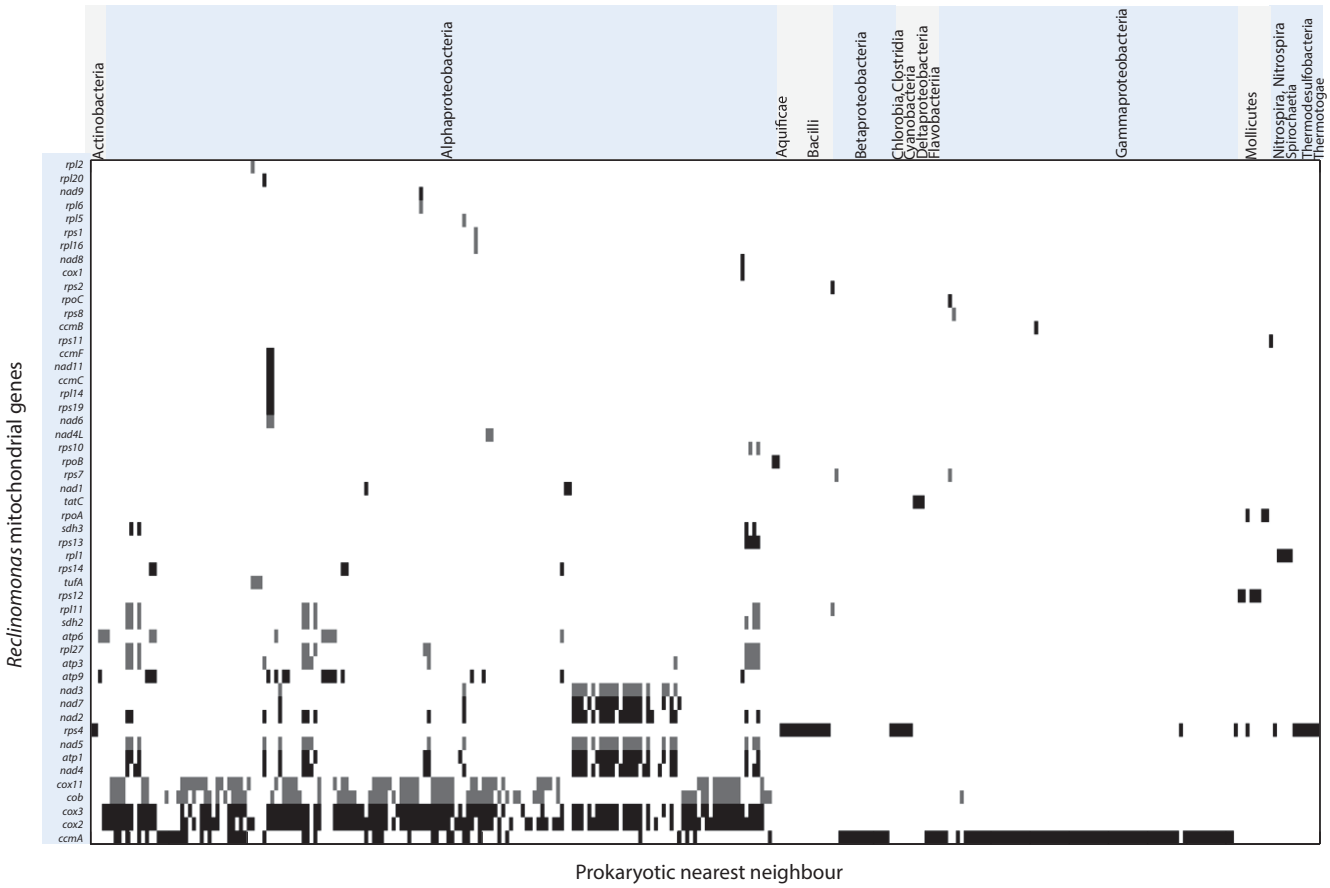


Fig. S2. A summary of the taxonomic distribution of prokaryotic taxa occurring in the nearest neighbor (sister group) of 51 *R. americana* mitochondrial genes in maximum likelihood trees based on the forward alignments. Each tick represents the occurrence of a taxon (detailed names in Dataset S1) from a prokaryotic group (above) in the nearest neighbor of a gene (left). There are 43 different combinations of taxa in the nearest neighbors inferred from the forward alignments, and 20 (indicated with gray ticks) of the 51 genes (39.2%) have a different nearest neighbor in the tree based on the reverse alignment.

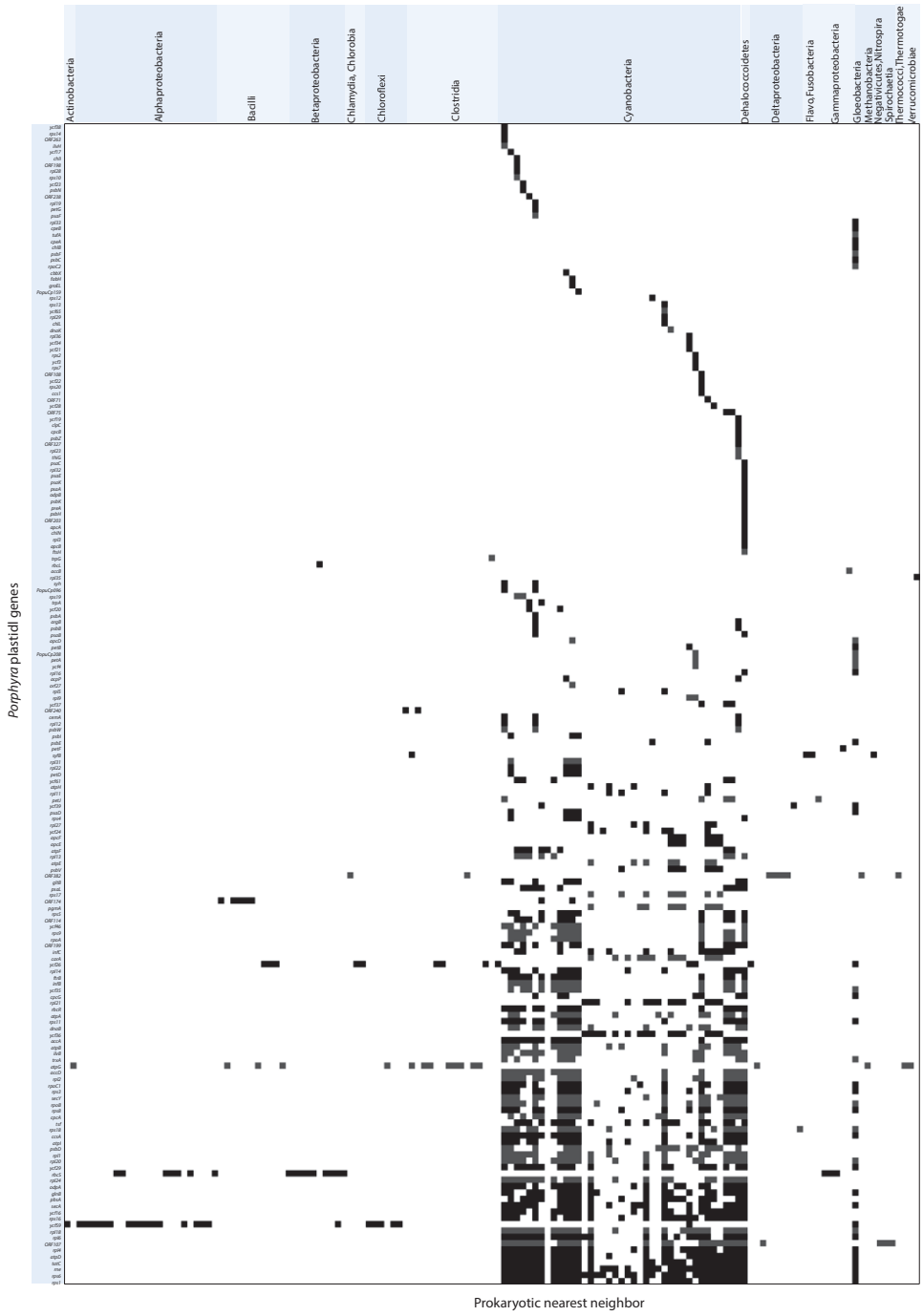


Fig. S3. A summary of the taxonomic distribution of prokaryotic taxa occurring in the nearest neighbor (sister group) of 183 *P. purpurea* plastid genes in maximum likelihood trees based on the forward alignments. Each tick represents the occurrence of a taxon (detailed names in Dataset S1) from a prokaryotic group (above) in the nearest neighbor of a gene (left). There are 124 different combinations of taxa in the nearest neighbors inferred from the forward alignments, and 52 (indicated with gray ticks) of the 183 genes (28.4%) have a different nearest neighbor in the tree based on the reverse alignment.

Table S1. Top 100 prokaryote hits in the NCBI nr database in a BLASTP search on Nov. 26, 2014, using the plastid ATP/ADP translocase of *Arabidopsis* (NP_173003) as query

Lineage	Hit description	Query coverage, %	Score	E value	Identity, %	Accession no.
Chlamydiae	Multispecies: ADP,ATP carrier family protein (*Chlamydia*)	82	498	1E−166	51	WP_020370051.1
Chlamydiae	ADP,ATP carrier protein 1 (*Chlamydia pecorum*)	82	520	2E−175	51	YP_004377432.1
Chlamydiae	ADP/ATP carrier protein (*Chlamydia pecorum*)	82	520	2E−175	51	WP_021757805.1
Chlamydiae	ADP/ATP carrier family protein (*Chlamydia gallinacea*)	82	492	3E−164	49	WP_021828277.1
Chlamydiae	ADP,ATP carrier protein (*Chlamydia trachomatis*)	81	516	6E−174	51	YP_327864.1
Chlamydiae	ADP/ATP carrier protein (*Chlamydophila abortus*)	81	504	2E−169	51	YP_219835.1
Chlamydiae	ADP/ATP carrier protein (*Chlamydia psittaci*)	81	503	1E−168	51	YP_005664131.1
Chlamydiae	ADP/ATP carrier protein (*Chlamydia psittaci*)	81	503	1E−168	51	YP_004422280.1
Chlamydiae	ATPase AAA (*Chlamydia psittaci*)	81	503	2E−168	51	WP_032741515.1
Chlamydiae	ADP,ATP carrier protein 1 (*Chlamydia suis* MD56)	81	508	1E−170	50	ESN89754.1
Chlamydiae	ADP/ATP translocase (*Chlamydophila felis*)	81	499	2E−167	49	YP_515490.1
Deltaproteobacteria	Nucleotide transport protein (*Lawsonia intracellularis*)	81	349	4E−109	42	YP_594385.1
Chlamydiae	ADP,ATP carrier protein (*Parachlamydia acanthamoebae*)	81	365	3E−115	39	YP_004653277.1
Chlamydiae	ADP/ATP carrier protein family (*Chlamydia pneumoniae*)	80	514	4E−173	52	YP_005661992.1
Bacteroidetes	ADP,ATP carrier protein 2 (*Bacteroides fragilis* str. S6L5)	80	483	4E−161	50	EYE60388.1
Gammaproteobacteria	Hypothetical protein (*Legionella shakespearei*)	80	450	5E−148	46	WP_018576123.1
Chlamydiae	ADP,ATP carrier protein 1 (*Simkania negevensis*)	79	511	6E−172	54	YP_004670874.1
Chlamydiae	ADP,ATP carrier protein (*Chlamydia psittaci* 02DC22)	79	493	4E−165	51	EPJ16785.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia akari*)	79	355	1E−111	41	YP_001492929.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia australis*)	79	341	3E−106	41	YP_005414385.1
Alphaproteobacteria (Rickettsiales)	ATP/ADP translocase (*Rickettsia bellii*)	79	347	1E−108	40	YP_001495592.1
Alphaproteobacteria (Rickettsiales)	ATP/ADP translocase (*Rickettsia bellii*)	79	347	2E−108	40	YP_538525.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia felis*)	79	341	2E−106	40	YP_246138.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia canadensis*)	79	340	9E−106	40	YP_005299113.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (*Rickettsia tamurae*)	79	355	1E−111	40	WP_032139269.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia canadensis*)	79	339	2E−105	40	YP_001491788.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (*Rickettsia honei*)	79	349	1E−109	40	WP_016917366.1
Alphaproteobacteria (Rickettsiales)	ADP/ATP translocase 1 (*Rickettsia monacensis*)	79	339	1E−105	40	WP_023507442.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (*Rickettsia aeschlimannii*)	79	335	3E−104	40	WP_032073780.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (*Rickettsia sibirica*)	79	350	1E−109	40	WP_016769753.1
Alphaproteobacteria (Rickettsiales)	ATP/ADP translocase (*Rickettsia africae*)	79	348	3E−109	40	YP_002844794.1
Alphaproteobacteria (Rickettsiales)	ADP/ATP translocase 1 (*Rickettsia* endosymbiont of *Ixodes scapularis*)	79	338	5E−105	40	KDO02891.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia philipii*)	79	337	9E−105	40	YP_005300161.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia rickettsii*)	79	337	1E−104	40	YP_001494197.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia peacockii*)	79	337	1E−104	40	YP_002916372.1
Alphaproteobacteria (Rickettsiales)	ATP/ADP translocase (*Rickettsia slovaca*)	79	336	3E−104	40	YP_005065328.1

Table S1. Cont.

Lineage	Hit description	Query coverage, %	Score	E value	Identity, %	Accession no.
Alphaproteobacteria (Rickettsiales)	ATP/ADP translocase (*Rickettsia parkeri*)	79	335	3E−104	40	YP_005392342.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia japonica*)	79	335	4E−104	40	YP_004884417.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia heilongjiangensis*)	79	335	6E−104	40	YP_004763800.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (*Rickettsia conorii*)	79	336	2E−104	40	WP_029374502.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (*Rickettsia conorii*)	79	335	4E−104	40	WP_029374609.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (*Rickettsia conorii*)	79	347	1E−108	40	WP_029374421.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (*Candidatus Rickettsia gravesii*)	79	333	2E−103	40	WP_017443549.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (*Rickettsia conorii*)	79	333	3E−103	40	WP_010976765.1
Alphaproteobacteria (Rickettsiales)	ATP/ADP translocase (*Candidatus Rickettsia amblyommii*)	79	334	1E−103	40	YP_005364858.1
Alphaproteobacteria (Rickettsiales)	ATP/ADP translocase (*Rickettsia massiliae*)	79	332	6E−103	40	YP_005301542.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia montanensis*)	79	331	1E−102	39	YP_005391808.1
Alphaproteobacteria (Rickettsiales)	ATP/ADP translocase (*Rickettsia massiliae*)	79	331	1E−102	39	YP_001498943.1
Alphaproteobacteria (Rickettsiales)	ATP/ADP translocase (*Rickettsia rhipicephali*)	79	328	2E−101	39	YP_005389935.1
Chlamydiae	ADP,ATP carrier protein 1 (*Simkania negevensis*)	78	402	2E−129	43	YP_004672579.1
Chlamydiae	ADP/ATP translocase (*Criblamydia sequanensis* CRIB−18)	78	347	4E−108	39	CDR34758.1
Chlamydiae	ADP/ATP translocase (*Candidatus Protochlamydia amoebophila*)	77	504	1E−169	55	YP_007249.1
Chlamydiae	ADP/ATP translocase (*Waddlia chondrophila*)	77	513	5E−173	55	YP_003710028.1
Chlamydiae	ATPase AAA (*Chlamydia* sp. "Diamant")	77	502	1E−168	54	WP_032124907.1
Chlamydiae	ADP/ATP carrier family protein (*Chlamydia psittaci* 06–1683)	77	500	7E−168	52	EPJ33718.1
Chlamydiae	ATPase AAA (*Chlamydia pneumoniae*)	77	508	6E−171	52	WP_010882994.1
Chlamydiae	ATPase AAA (*Chlamydia pneumoniae*)	77	506	3E−170	52	WP_010895317.1
Gammaproteobacteria	ATPase AAA (*Endozoicomonas elysicola*)	77	491	1E−164	51	KEI72701.1
Gammaproteobacteria	ATPase AAA (*Candidatus Caedibacter acanthamoebae*)	77	446	6E−147	49	AIL13243.1
Chlamydiae	ADP/ATP translocase (*Waddlia chondrophila*)	77	374	9E−119	44	YP_003708595.1
Chlamydiae	ADP/ATP translocase (*Candidatus Protochlamydia amoebophila*)	77	397	8E−128	42	YP_007240.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (*Rickettsia prowazekii*)	77	355	7E−112	41	WP_004599717.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein (*Rickettsia prowazekii*)	77	355	8E−112	41	YP_005413506.1
Alphaproteobacteria (Rickettsiales)	Hypothetical protein (Rickettsiaceae bacterium Os18)	77	328	3E−101	40	WP_019230826.1
Alphaproteobacteria (Rickettsiales)	ADP/ATP carrier protein 1 (*Rickettsia typhi*)	77	355	1E−111	40	YP_067047.1
Chlamydiae	ADP/ATP translocase 1 (*Chlamydia* sp. "Rubis")	77	340	1E−105	39	CDZ79513.1
Chlamydiae	ADP/ATP carrier protein (*Chlamydia pecorum*)	76	516	5E−174	53	WP_021756588.1
Chlamydiae	ADP/ATP translocase (*Criblamydia sequanensis* CRIB−18)	76	492	9E−165	53	CDR34749.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (endosymbiont of *Acanthamoeba* sp. UWC9)	76	411	4E−133	46	AIK96967.1
Alphaproteobacteria (Rickettsiales)	ATPase AAA (*Candidatus Paracaedibacter symbiosus*)	76	393	3E−126	45	WP_032112829.1
Alphaproteobacteria (Rickettsiales)	Hypothetical protein (*Candidatus Hepatobacter penaei*)	76	382	3E−122	44	WP_031934367.1
Chlamydiae	ADP/ATP translocase (*Waddlia chondrophila*)	76	400	1E−128	42	YP_003710049.1
Chlamydiae	ATPase AAA (*Chlamydia* sp. "Diamant")	76	376	2E−119	42	WP_032124896.1

Table S1. Cont.

Lineage	Hit description	Query coverage, %	Score	E value	Identity, %	Accession no.
Chlamydiae	ATPase AAA (*Chlamydia* sp. "Diamant")	76	372	5E−118	41	WP_032124897.1
Chlamydiae	ADP/ATP translocase (*Candidatus Protochlamydia amoebophila*)	76	376	3E−119	41	YP_007239.1
Chlamydiae	ADP/ATP translocase 1 (*Chlamydia* sp. "Rubis")	76	378	3E−120	41	CDZ79514.1
Alphaproteobacteria (Rickettsiales)	ATP/ADP translocase class 1 (*Candidatus Midichloria mitochondrii*)	76	346	2E−108	41	YP_004679189.1
Alphaproteobacteria (non−Rickettsiales)	ATPase AAA (*Candidatus Liberibacter solanacearum*)	76	329	1E−101	40	KGB27446.1
Alphaproteobacteria (non−Rickettsiales)	ATP/ADP translocase (*Candidatus Liberibacter asiaticus*)	76	340	4E−106	39	YP_003064736.1
Chlamydiae	ADP/ATP translocase (*Criblamydia sequanensis* CRIB−18)	76	348	6E−109	39	CDR34759.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein 1 (*Holospora obtusa*)	76	329	1E−101	39	WP_024161165.1
Chlamydiae	ADP,ATP carrier protein (*Parachlamydia acanthamoebae*)	75	504	2E−169	54	YP_004651022.1
Chlamydiae	ADP,ATP carrier protein (*Chlamydia trachomatis*)	75	509	2E−171	54	YP_002887693.1
Chlamydiae	ADP, ATP carrier protein (*Chlamydia muridarum*)	75	509	3E−171	54	NP_296714.1
Chlamydiae	ADP,ATP carrier protein (*Chlamydia trachomatis*)	75	509	3E−171	54	YP_006360005.1
Chlamydiae	ADP,ATP carrier protein (*Chlamydia trachomatis*)	75	509	3E−171	54	YP_005809267.1
Chlamydiae	ATPase AAA (*Chlamydia trachomatis*)	75	509	3E−171	54	WP_009871413.1
Chlamydiae	ADP/ATP carrier protein family protein (*Chlamydia trachomatis*)	75	509	3E−171	53	YP_007737160.1
Chlamydiae	ATPase AAA (*Chlamydia trachomatis*)	75	508	9E−171	53	WP_009873532.1
Chlamydiae	ADP, ATP carrier protein (Chlamydophila *caviae*)	75	484	1E−161	52	NP_829303.1
Chlamydiae	ADP,ATP carrier protein (*Chlamydia psittaci*)	75	484	2E−161	52	WP_020356125.1
Gammaproteobacteria	ATPase AAA (*Endozoicomonas elysicola*)	75	482	3E−161	52	WP_033402749.1
Chlamydiae	ADP/ATP translocase (*Waddlia chondrophila*)	75	353	4E−111	42	YP_003710048.1
Alphaproteobacteria (Rickettsiales)	ADP,ATP carrier protein homolog (*Orientia tsutsugamushi*)	75	330	3E−102	40	YP_001937491.1
Alphaproteobacteria (Rickettsiales)	ATP/ADP translocase (*Orientia tsutsugamushi*)	75	329	6E−102	40	YP_001248406.1
Chlamydiae	ATPase AAA (*Chlamydia psittaci*)	74	442	7E−146	49	WP_032742269.1
Chlamydiae	ADP/ATP translocase (*Criblamydia sequanensis* CRIB−18)	74	372	4E−118	43	CDR33792.1
Chlamydiae	ADP,ATP carrier protein (*Parachlamydia acanthamoebae*)	73	356	1E−111	42	YP_004653276.1
Alphaproteobacteria (non−Rickettsiales)	ATP/ADP translocase (*Candidatus Liberibacter americanus*)	73	338	8E−106	40	WP_023466217.1
Chlamydiae	ADP,ATP carrier protein (*Chlamydia psittaci* 08–2626_L3)	71	427	4E−140	50	EPP28844.1

Lineage assignment is based on NCBI Taxonomy (www.ncbi.nlm.nih.gov/taxonomy). The hits are sorted according to query coverage. Although there is only one hit from Bacteroidetes among the top 100 hits, another Bacteroidetes taxon, "*Candidatus* Amoebophilus asiaticus," also has ATP/ADP translocase in its completely sequenced genome (1). The presence of ATP/ADP translocase in the genome, which was sequenced using long-insert (8-kb and 40-kb) libraries with a coverage of at least 10×, provides additional, strong support for the existence of ATP/ADP translocase in Bacteroidetes.

1. Schmitz-Esser S, et al. (2010) The genome of the amoeba symbiont "*Candidatus* Amoebophilus asiaticus" reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *J Bacteriol* 192(4):1045–1057.

Other Supporting Information Files

Dataset S1 (TXT)

3.4

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]

[1] Institute of Molecular Evolution, Heinrich-Heine University, 40225 Düsseldorf, Germany.

[2] Institute of Fundamental Sciences, Massey University, Palmerston North 4474, New Zealand.

[3] Departmentof 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] GenomicMicrobiology 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.

Corresponding author: bill@hhu.de

The presented manuscript was published in the journal *Nature* in 2015.

Contribution of Chuan Ku (first author)

Experimental design: 40%

Data analysis: 70%

Manuscript writing: 45%

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.

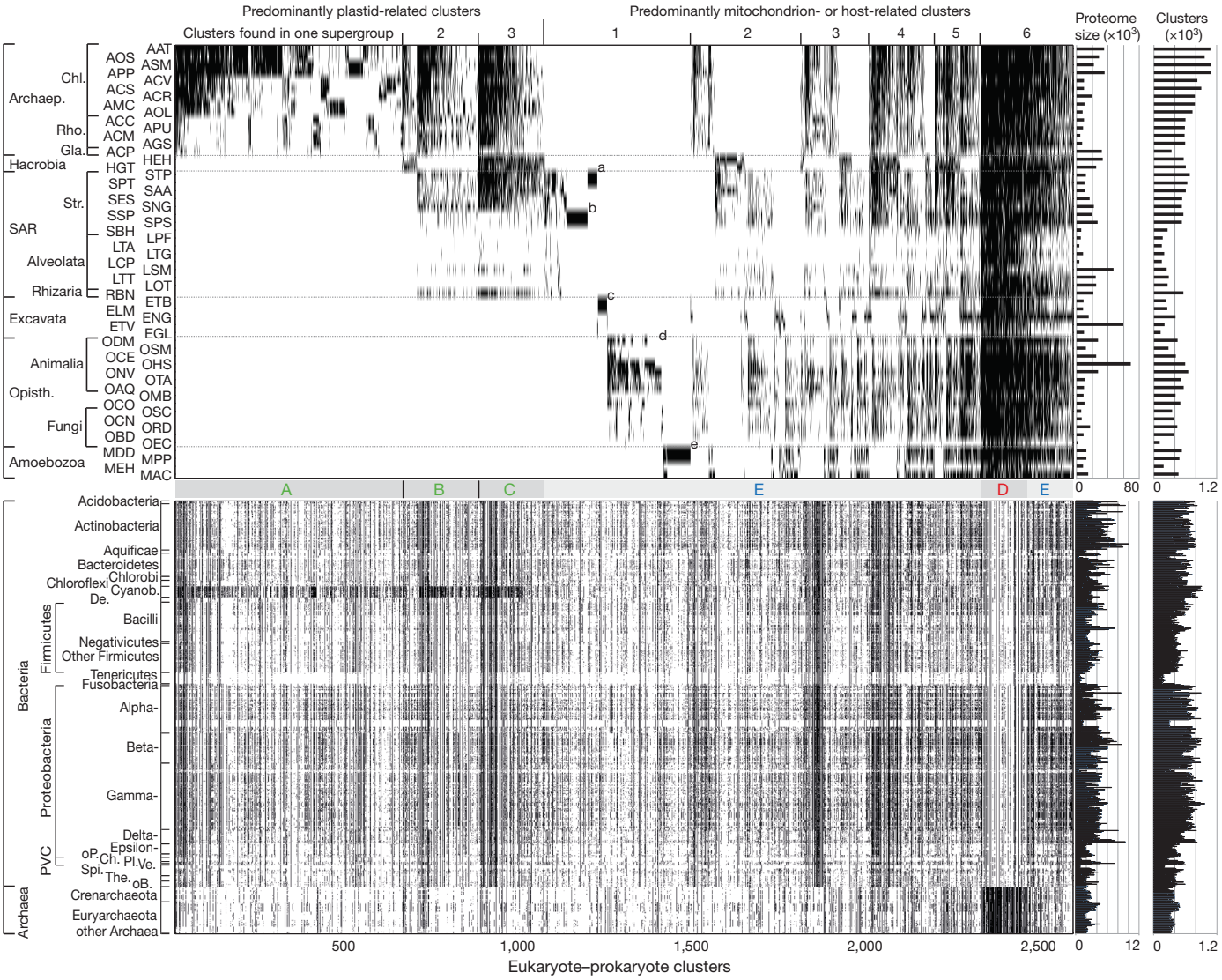


Figure 1 | Distribution of taxa in EPCs. Each black tick indicates gene presence in a taxon. The 2,585 EPCs (*x* axis) are ordered first according to their distribution across six eukaryotic supergroups with clusters specific to lineages with photosynthetic eukaryotes (blocks A–C) on the left, then according 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

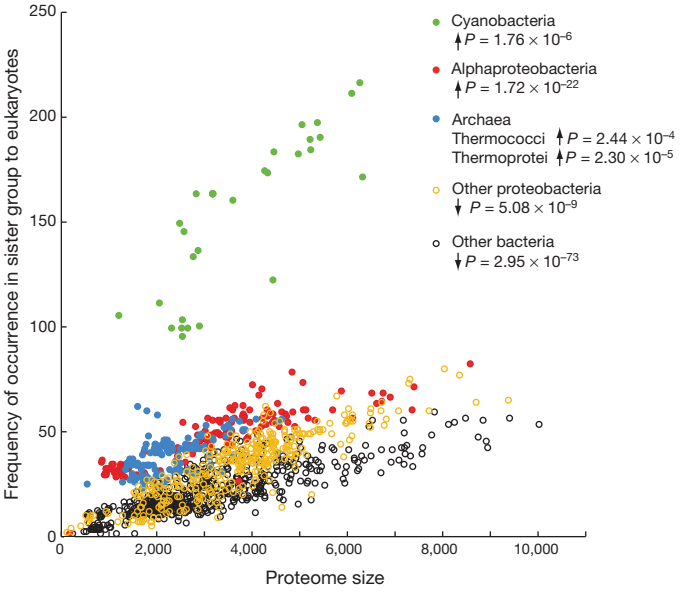


Figure 2 | Occurrence in the sister group versus proteome size. Prokaryotic taxa are plotted according to how frequently they are found in the sister group (defined as the nearest neighbour group) to a monophyletic group of eukaryotes in 1,933 trees against their proteome size. A two-sided Wilcoxon signed-rank 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.

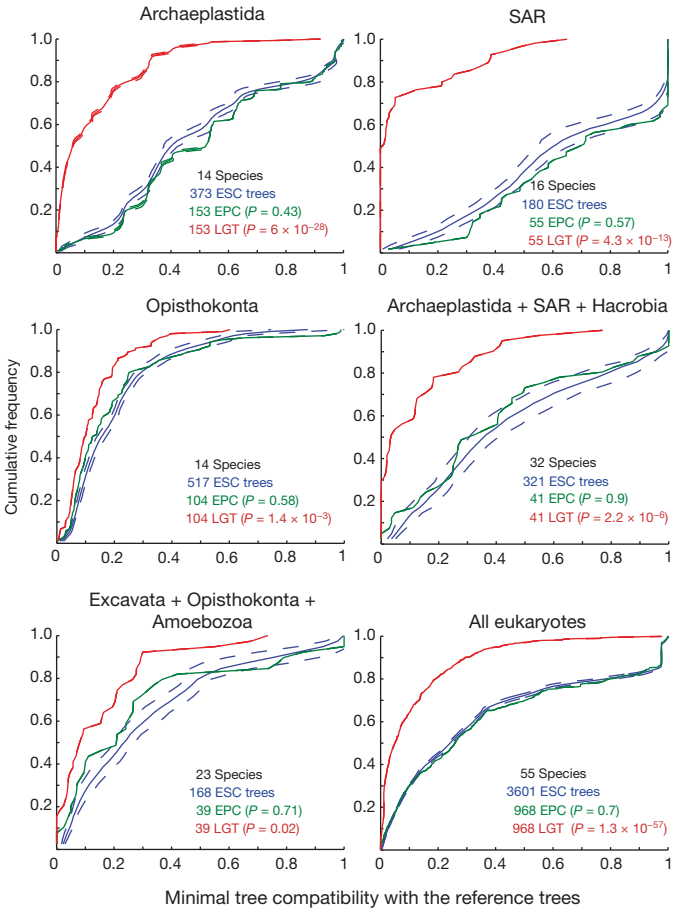


Figure 3 | Comparison of sets of trees for single-copy genes in eukaryotic groups. Cumulative distribution functions (*y* axis) for scores of minimal tree compatibility with the vertical reference data set (*x* axis). Values are number of species, sample sizes, and *P* values of the two-tailed Kolmogorov–Smirnov 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

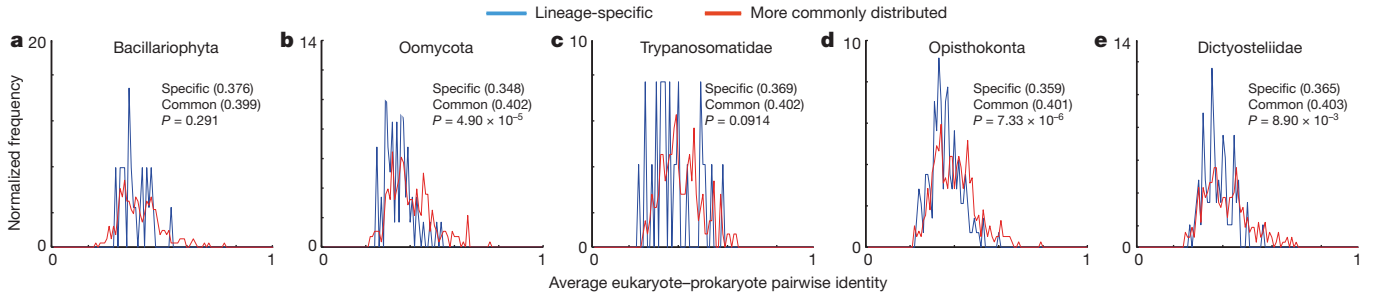


Figure 4 | Eukaryote–prokaryote sequence identities for genes with a tip distribution in eukaryotes versus those whose distributions trace their presence to a more ancient ancestor. a–e, Genes denoted by lower-case letters in Fig. 1 and those found in at least three of five major supergroups. The mean 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.

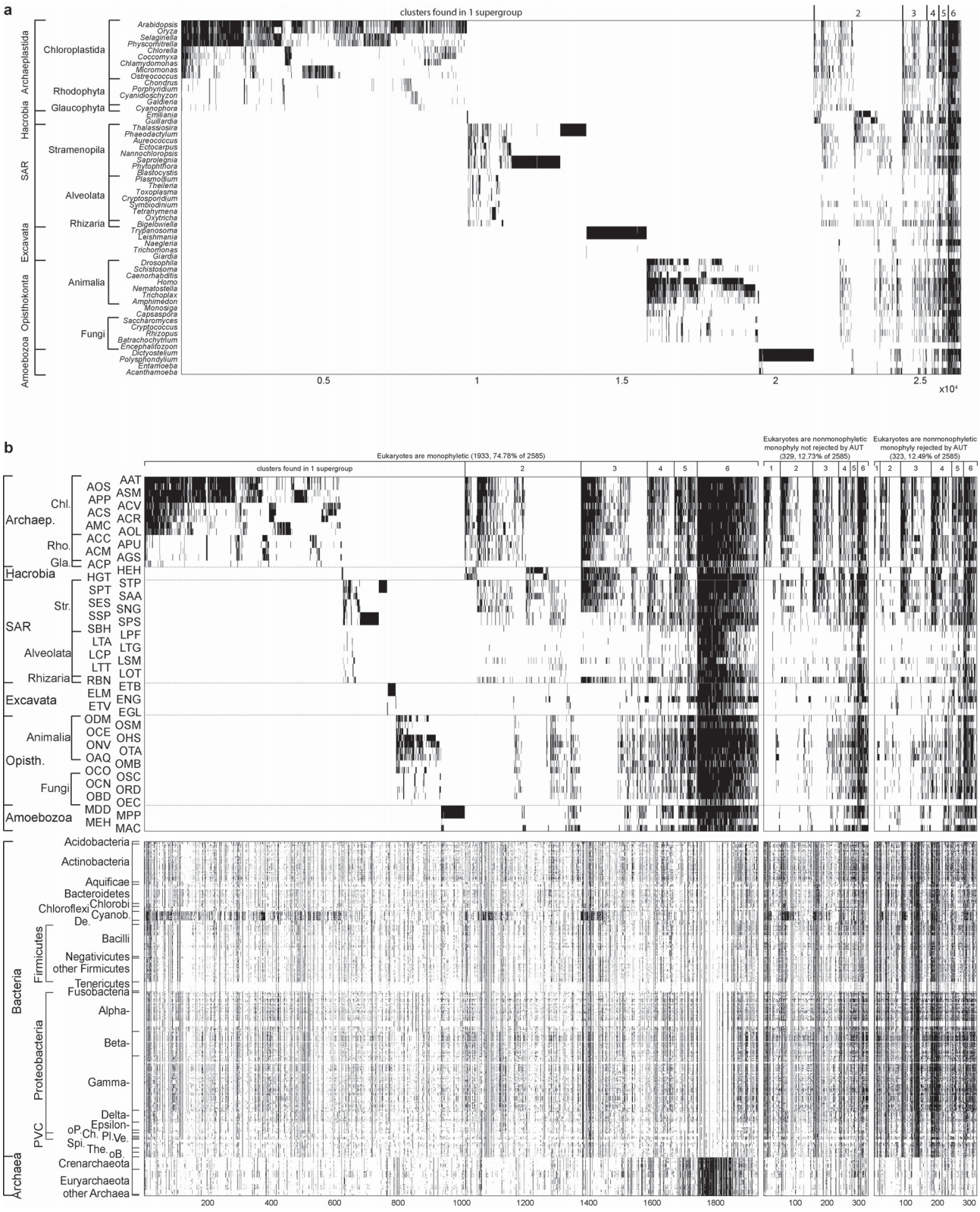
For 134 trees, there was no obvious contamination problem or case of cyanobacteria and plants interleaving. These 134 cases were therefore classified as putative LGT (Supplementary Table 6). But when the 134 cases were compared with the eukaryote monophyletic EPCs, we found significantly more trees than 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%.

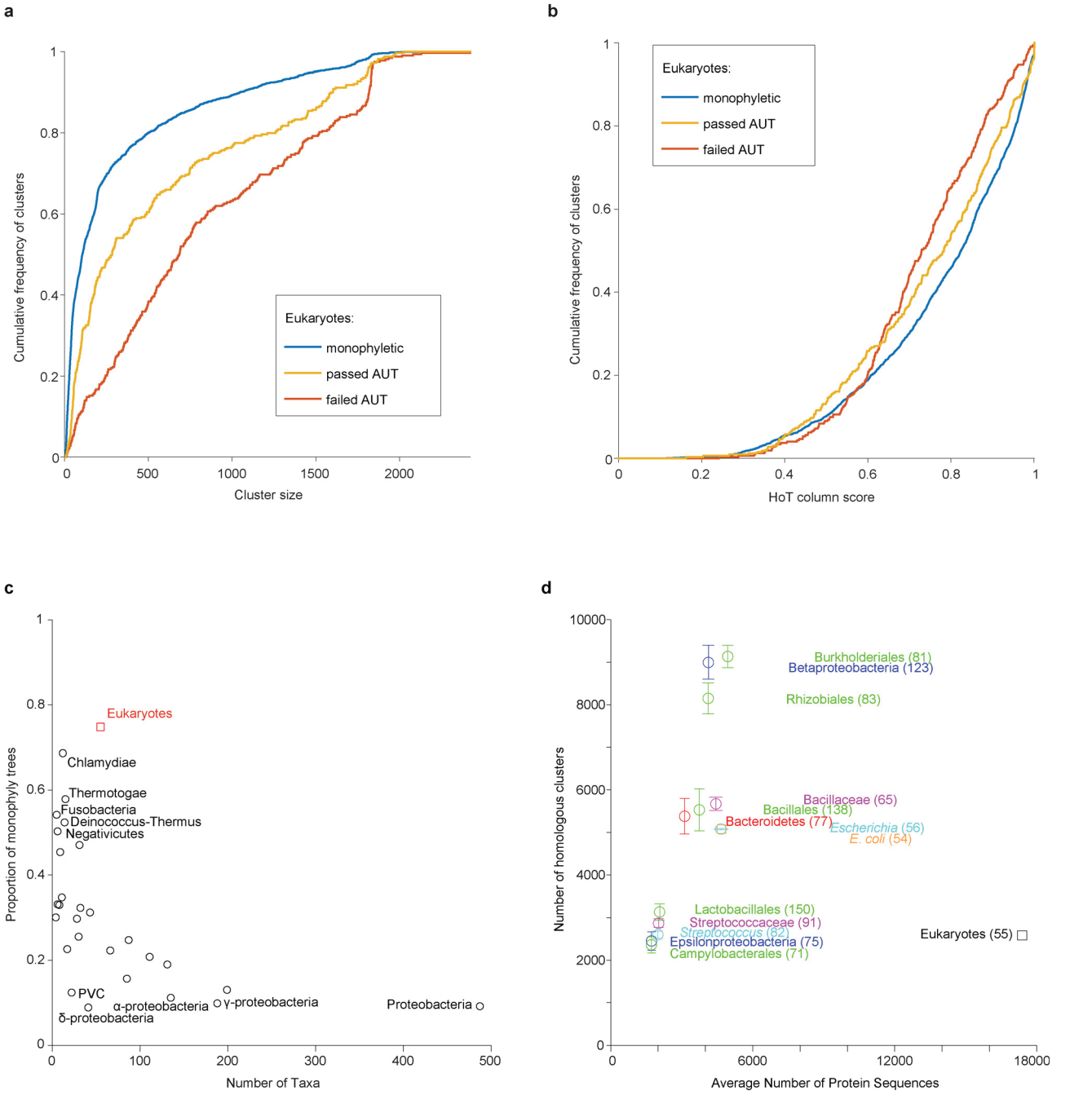
Note that the number for the host contribution is probably an underestimate, as only EPCs with a monophyletic eukaryotic clade in the maximum likelihood tree were counted. For genes of plastid origin, it might be a slight overestimate, since there would also be genes of plastid–host origin that are now 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).

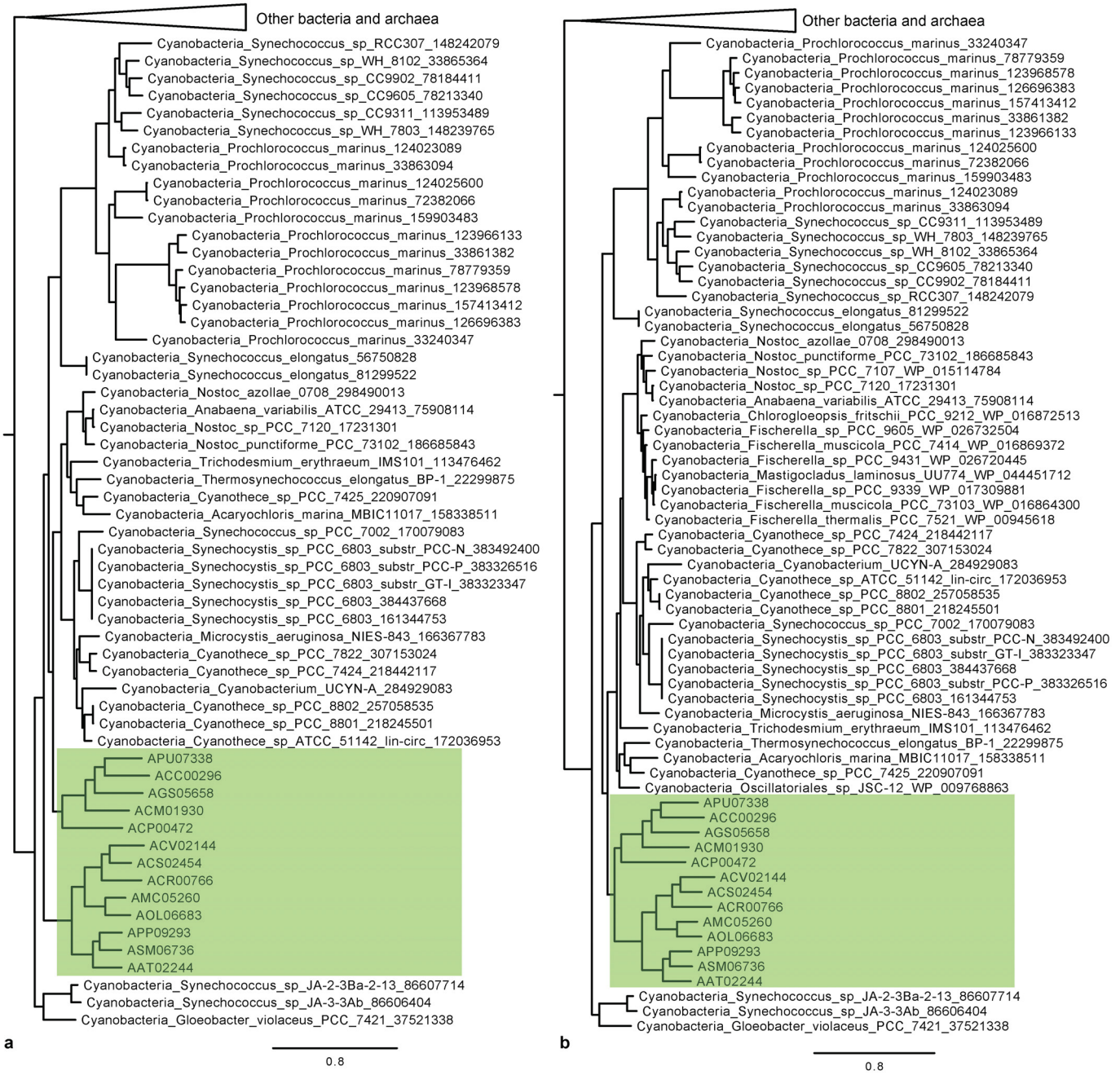
77. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
78. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood 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).



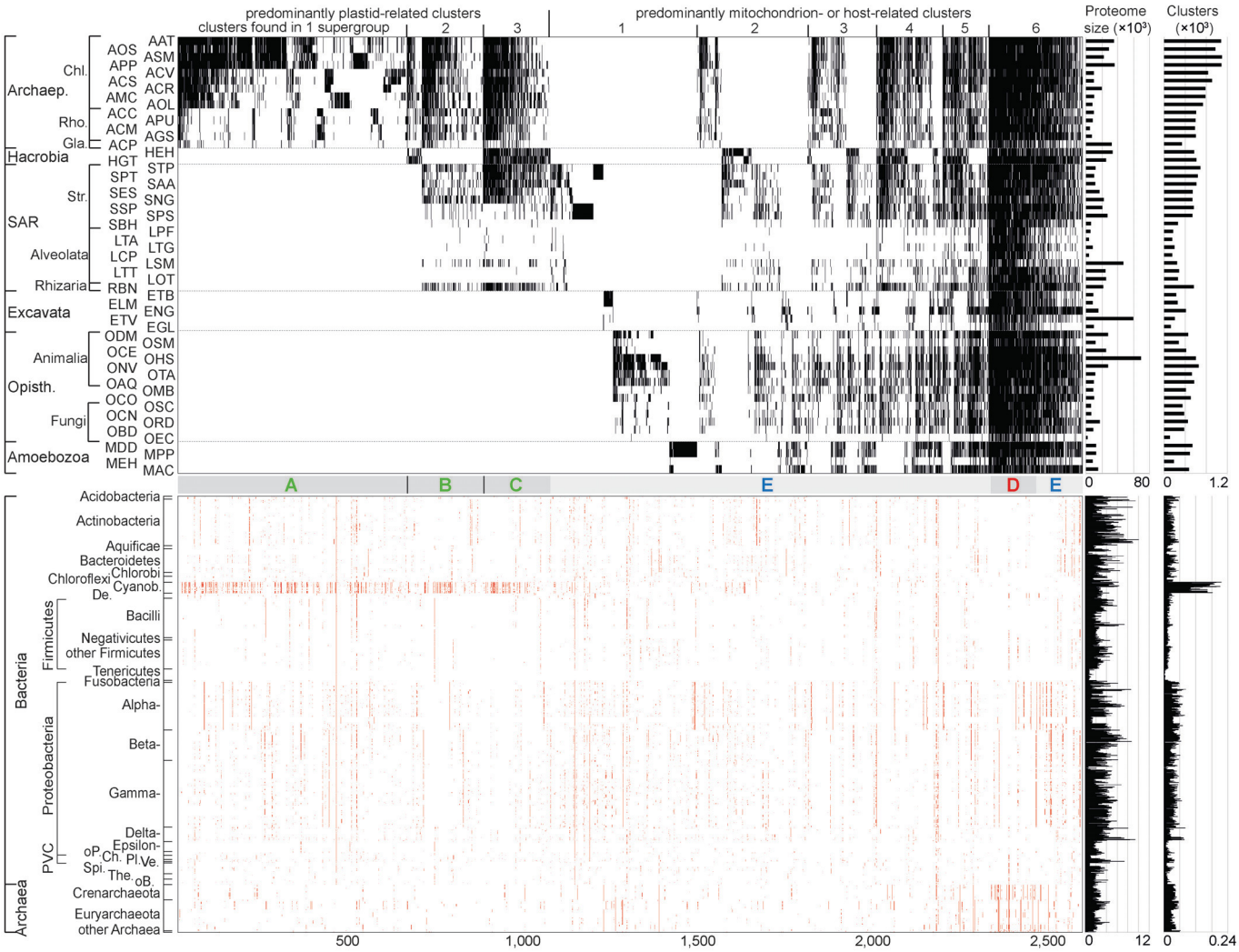
Extended Data Figure 1 | Additional gene distribution patterns. **a**, Distribution of ESCs. Each black tick indicates the presence of a cluster in a taxon. The 26,117 ESCs (*x* axis) from 55 eukaryotic genomes (Supplementary Table 1) are sorted according to their distribution across the six eukaryotic 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.



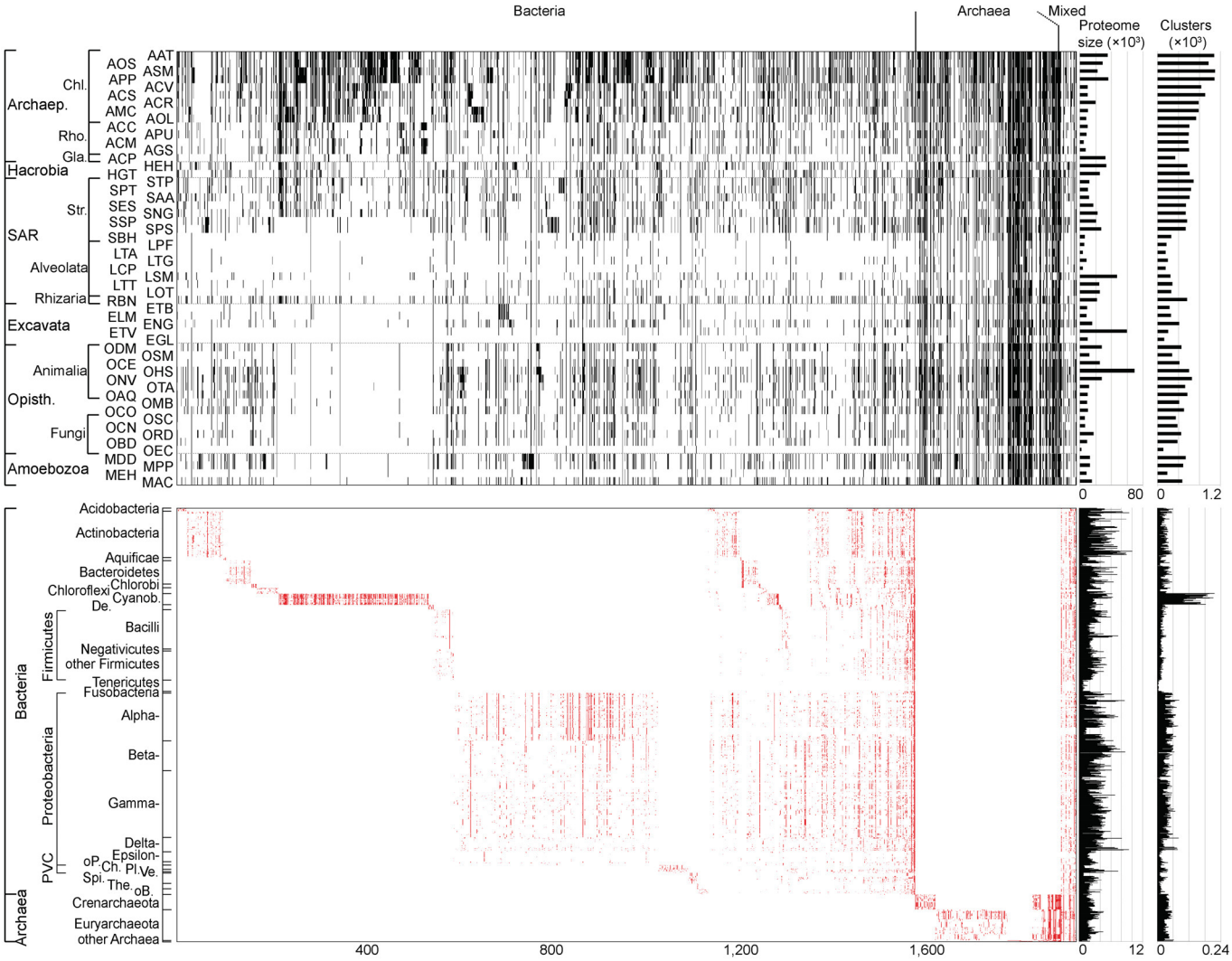
Extended Data Figure 2 | Clustering, monophyly, and gene sharing. **a**, **b**, Monophyly of eukaryotes in maximum likelihood trees, cluster size, and alignment quality. Cumulative frequency of clusters with different cluster size (**a**) or different HoT[72] column scores (**b**) is plotted for three 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.



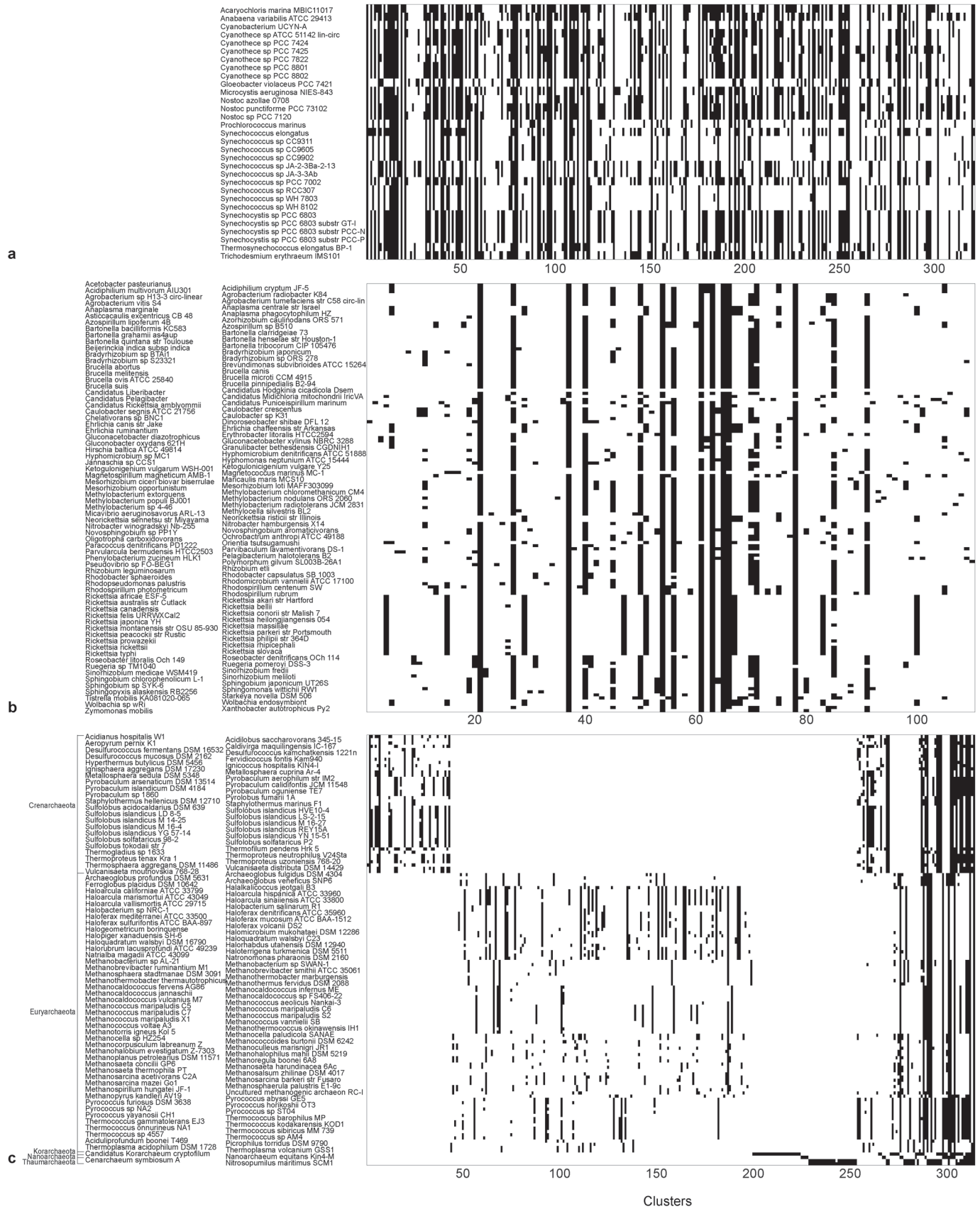
Extended Data Figure 3 | Effect of taxon sampling on eukaryote monophyly in phylogenetic trees. After ten sequences (bold) were added to the original data set (EPC E1689_B206_A295), the relationships among Archaeplastida taxa (highlighted in green) changed from non-monophyly (**a**) to monophyly (**b**). 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).



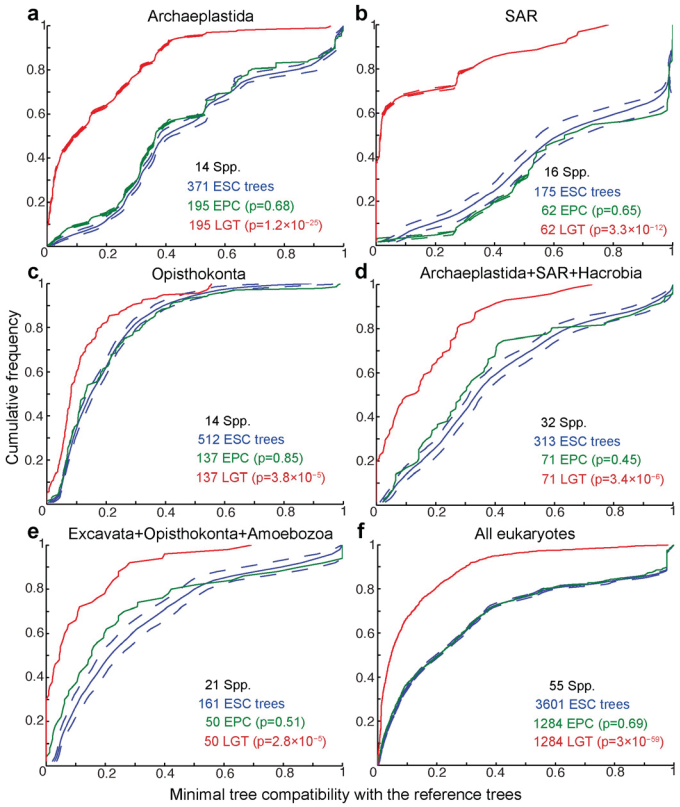
Extended Data Figure 4 | Distribution of prokaryotic taxa in the sister group to eukaryotes, with EPCs sorted by eukaryotic supergroups. Top: each black tick indicates the presence of a eukaryote taxon in one of the 2,585 EPCs. Bottom: each red tick indicates the presence of a prokaryote taxon in the 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.



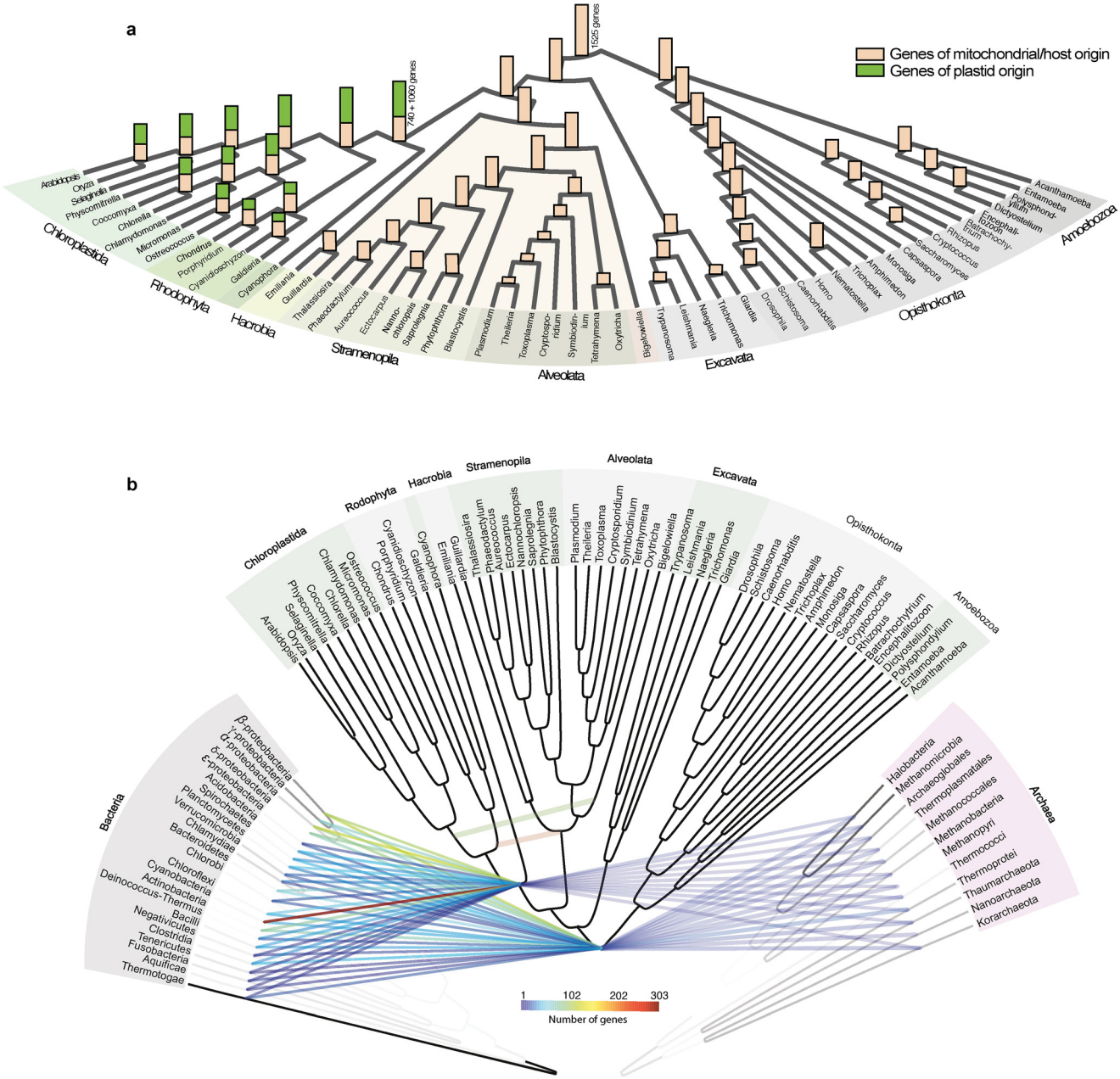
Extended Data Figure 5 | Distribution of prokaryotic taxa in the sister group to eukaryotes, with EPCs sorted by prokaryotic groups. Top: each black tick indicates the presence of a eukaryote taxon in one of the 1,933 EPC maximum likelihood trees where eukaryotes were resolved to be monophyletic. Bottom: 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.



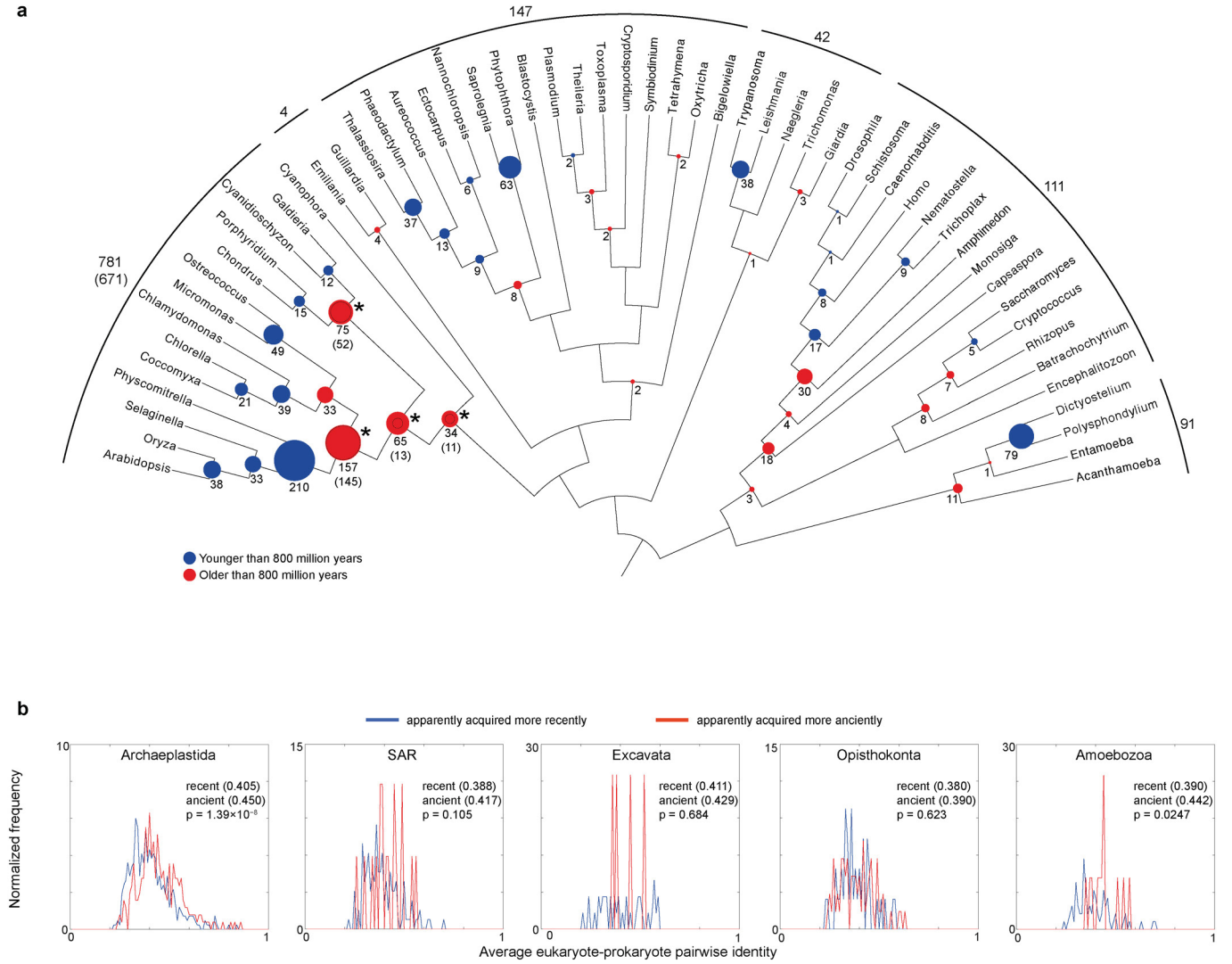
Extended Data Figure 6 | Distribution of taxa in the sister groups consisting purely of cyanobacteria, alphaproteobacteria, or archaea. Each black tick indicates the presence of a prokaryotic taxon in the sister group to eukaryotes in an EPC tree. **a–c**, Distributions of taxa in all pure-cyanobacterial (**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.



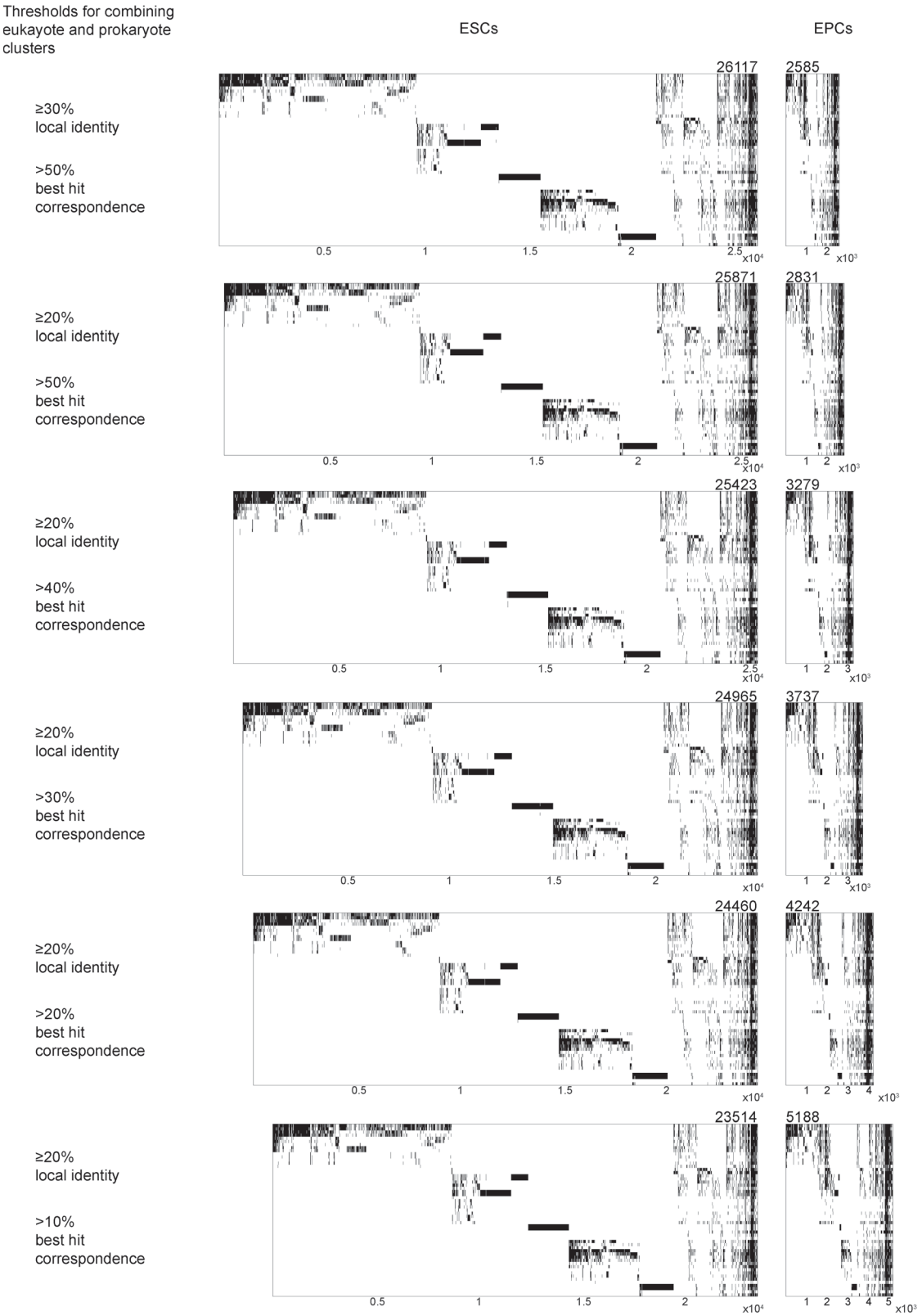
Extended Data Figure 7 | Comparison of sets of trees for single-copy genes in eukaryotic groups, with more inclusive criteria. a–f, Cumulative distribution functions (*y* axis) for scores of minimal tree compatibility with the vertical reference data set (*x* axis). Values are number of species, sample 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.

3.5

A natural barrier to lateral gene transfer from prokaryotes to eukarytoes revealed from genomes: The 70% rule

Chuan Ku, William F. Martin
Institute of Molecular Evolution, Heinrich-Heine University, 40225 Düsseldorf, Germany.
Corresponding author: bill@hhu.de

The presented manuscript was submitted to the journal *BMC Biology* in 2016.

Contribution of Chuan Ku (first author)
Experimental design: 60%
Data analysis: 90%
Manuscript writing: 60%

A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: The 70% rule

A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: The 70% rule

Chuan Ku, William F. Martin*

Institute of Molecular Evolution, Heinrich-Heine University, Düsseldorf, Germany

* Author for correspondence: bill@hhu.de

Abstract

The literature harbours many claims for lateral gene transfer (LGT) from prokaryote to eukaryotes. Such claims are typically founded in analyses of genome sequences. It is undisputed that many genes entered the eukaryotic lineage via the origin of mitochondria and the origin of plastids. Claims for lineage specific LGT to eukaryotes outside the context of organelle origins, or claims of continuous LGT to eukaryotic lineages are more problematic. If eukaryotes acquire genes from prokaryotes continuously during evolution, then sequenced eukaryote genomes should harbour evidence for recent LGT, like prokaryotic genomes do. Here we investigate 30,358 eukaryotic sequences in the context of 1,035,375 prokaryotic homologues among 2,585 phylogenetic trees containing homologs from prokaryotes and eukaryotes. Prokaryote genomes reflect a continuous process of gene acquisition and inheritance, with abundant recent acquisitions showing 80-100% amino acid sequence identity to their phylogenetic sister group homologues from other phyla. By contrast, eukaryote genomes show no evidence for either continuous or recent gene acquisitions from prokaryotes. We find that, in general, genes in eukaryotic genomes that share ≥70% amino acid identity to prokaryotic homologues are are not found outside individual genome assemblies. We propose a 70% rule: coding sequences in eukaryotic sharing more than 70% amino acid sequence identity to prokaryotic homologues are most likely assembly or annotation artifacts. The role of differential loss in eukaryote genome evolution has been vastly underestimated.

Introduction

Few topics in evolutionary biology have received as much attention in the last 20 years as lateral gene transfer (LGT, or horizontal gene transfer [HGT]) [1-3], with over 11,000 papers that have appeared on the topic since 1985, and over 30,000 citations to those paper in 2015 alone (Thomson Reuters Web of Science^{TM} as of April 21, 2016). That is too much literature for anyone to read or review in real time. As such, the LGT literature has almost taken on a life of its own, with reviews citing other reviews as evidence for LGT and claims for LGT being conflated with evidence for same. Cognizant biologists have learned one thing for sure about LGT, namely that not all papers bearing claims for LGT are evidence for the workings of LGT. Two recent tardigrade genome papers being a case in point. One report had it that 16.1% of the genes in the tardigrade genome were recently acquired via LGT from prokaryotes [4], while an independent sequencing project had it that there was virtually no LGT in the tardigrade genome [5], the main difference being that genes probably belonging to associated bacteria were annotated as tardigrade genes in the one study [4] but not in the other [5], whose scaffolds are longer and which is more cautious with contaminations. The human genome initially also suffered from claims for LGT [6], which were quickly refuted [7, 8], though are creeping back into the literature [9] on the basis of the same gene origin estimation software [10] used in the LGT-rich tardigrade genome.

In prokaryotes, LGT is best seen as a way of life. Several naturally occurring mechanisms of LGT among prokaryotes have been known for many decades, as reviewed by Jones and Sneath [11] and more recently by Popa and Dagan [12]: transfer by naked DNA uptake from the environment (transformation) [13], transfer by plasmid transfer (conjugation) [14], transfer via phage particles (transduction) [15] and gene transfer agents [16]. A great deal is known about the genes and proteins that moderate these LGT mechanisms in prokaryotes [17-19]. These LGT mechanisms merely introduce DNA into the prokaryotic cell, whether or not it recombines into the genome or not is governed by the genes and proteins that mediate DNA insertion and/or recombination [20, 21].

Importantly, the mechanisms that introduce DNA into the cell for LGT are the same that introduce DNA into the cell for normal recombination within prokaryotic species [22]. In prokaryotes, recombination is never reciprocal, it is always from donor to recipient. Prokaryotic genomes are highly dynamic in terms of gene content. They are typically replete with constant gains (often through LGT) and losses through deletion [2, 23-25]. Over time these gains and losses lead to pangenome structures not only at the species level but at all taxonomic levels [26-28].

In prokaryotes, acquisition through LGT dwarfs the role of gene duplication in generating gene families within genomes [29]. Prokaryotic LGT is pivotal in the spread of antibiotic resistance [30] and in ecological adaptation [31]. The extent of LGT in prokaryotes has challenged the traditional view of evolution as tree-like processes and prompted the use of more network-like representations [3, 32-34].

Contrary to prokaryotes, eukaryotes use meiosis and sex for recombination, which is always reciprocal [35]. Although eukaryotes are descended from prokaryotes [36, 37], at eukaryote origin they apparently lost the LGT mechanisms typical of prokaryotes, because interspecific or inter-phylum conjugation transformation, transduction and have so far not been observed in eukaryotes. As a consequence, prokaryotes have pangenomes, but eukaryotes lack pangenomes. The only mechanism characterized as a source of new

genes entering nuclear genomes in a natural manner is gene transfers from organelles [38]. Barring targeted gene transfer experiments [39], reports of prokaryote-to-eukaryote LGT are based on gene sequence comparisons and genome sequence annotations. Thus, in contrast to LGT among prokaryotes, which is their natural mechanism to generate new gene combinations, the role of LGT in eukaryote evolution is controversial.

Many recent papers suggest that LGT frequently occurs in phagotrophic, unicellular eukaryotes [40], that there is continuous LGT from prokaryotes to vertebrates and other animals [9] as well as to plants [41] and algae [42]. While the sources of LGT to eukaryotes can be pinpointed in some cases [43, 44], sometimes involving prokaryotes known for their ability to transfer DNA to eukaryotes [45], for the majority of cases reported for prokaryote-to-eukaryote LGT, the mechanisms and specifics (how, when, and between which groups) remain obscure. If the numerous claims for eukaryotes constantly acquiring prokaryotic genes through LGT [46, 47] are true, then there would indeed seem to be no natural barrier for prokaryote-to-eukaryote LGT. We ask: Can such claims be true? And how to find out whether they are true or not via the scientific method using a generally applicable test?

In previous work, we showed the gene acquisitions in eukaryotes correspond to endosymbiotic events [48] and that many of the patterns of "patchy" gene distributions that some reports take as evidence for LGT [46, 47] are in fact the result of differential loss [48] superimposed upon vertical inheritance. If there are really no barriers to LGT from prokaryotes to eukaryotes, as many current papers are claiming [9, 46], then eukaryote genomes should contain both anciently acquired prokaryotic genes and recently acquired prokaryotic genes. Furthermore, it should be possible using robust measures to demonstrate the presence or absence of recently acquired in eukaryotic genomes, using prokaryotic genomes as a test case for comparison.

Results

Our test is simple: Recent LGT in prokaryotes deposits new donor sequences in recipient genomes that show very high sequence identity between donor and recipient lineages [2, 49]. The high sequence identity between donor and recipient (initially 100%) gradually deteriorates over time (amelioration) so that more recent transfers tend to show higher similarity [30, 50]. Thus, if 5.1% [47] or even 16.1% [4] of the genes in a eukaryote come from prokaryotes via constant LGT accumulation over time [41, 46, 51-53] eukaryote genomes should exhibit distributions of donor-recipient sequence identity comparable to those seen in prokaryotes. If not, something is wrong with the eukaryote LGT reports. That prediction can be tested with genome data.

The present data are comprised of 2,585 phylogenetic trees from clusters (alignments) that contain homologs from prokaryotes and eukaryotes, also designated as eukaryotic-prokaryotic clusters (EPCs) [48]. Each of these clusters, generated from 55 eukaryotic and 1981 prokaryotic genomes (S1 Table), contains at least two eukaryotic and at least five prokaryotic sequences, and the sequence similarity threshold in pairwise comparisons is on the order of ≥25% [48]. The criterion of requiring genes to be present in at least two eukaryotic genomes serves to eliminate obvious bacterial contaminations from the data. Yet, as we will see, the two-eukaryote-genome criterion does not remove contaminations that are less obvious. The criterion of having at least five prokaryotic sequences in the cluster is to provide a reference tree framework for the investigation. The 25% amino acid sequence identity criterion is stricter than that employed in many other protein cluster databases, the COG [54] or KOG [55] databases, for example. Our clusters are generated for the purpose of generating alignments and phylogenetic trees, whereby pairwise sequence identity at or below 20% leads to problematic alignment and problematic trees [56].

In trees generated from the COG databases, for example, over 40% of trees exhibit what was once called 'pseudoparalogy', that is, the clusters unite several very distantly related prokaryotic and eukaryotic gene families into the same tree [57], which is fine if functional annotation is the goal (a main goal of many such databases), but problematic if alignments and trees are the objective of investigation. For the present data spanning 2,585 trees in which all sequences are uniquely assigned (no sequence occurs in more than one cluster), the number of taxa in each cluster is shown in S1 Fig, the mean number of eukaryote taxa per tree is 10.6, the mean number of prokaryote taxa per tree is 247.

The simplest way to look for evidence of recent transfer is to compare sequences from a clade of a given taxonomic group (for example, eukaryotes or bacilli) to the sister group of that clade in a maximum likelihood tree (Fig 1). For recent transfers, the proportion of identical amino acid residues for the clade-sister comparison ($I_{C\text{-}S}$) should be high, up to 1.0 (100% amino acid identity) for very recent acquisitions from outside the taxon (Fig 1A). For more ancient transfers (Fig 1B, C), values of $I_{C\text{-}S}$ should be lower, with a lower bound near 0.25 because of the 25%-identity clustering threshold [48]. A taxonomic group can have more than one clade in a tree (Fig 1C, D), and both recent and ancient transfers can be observed in the same tree (Fig 1D).

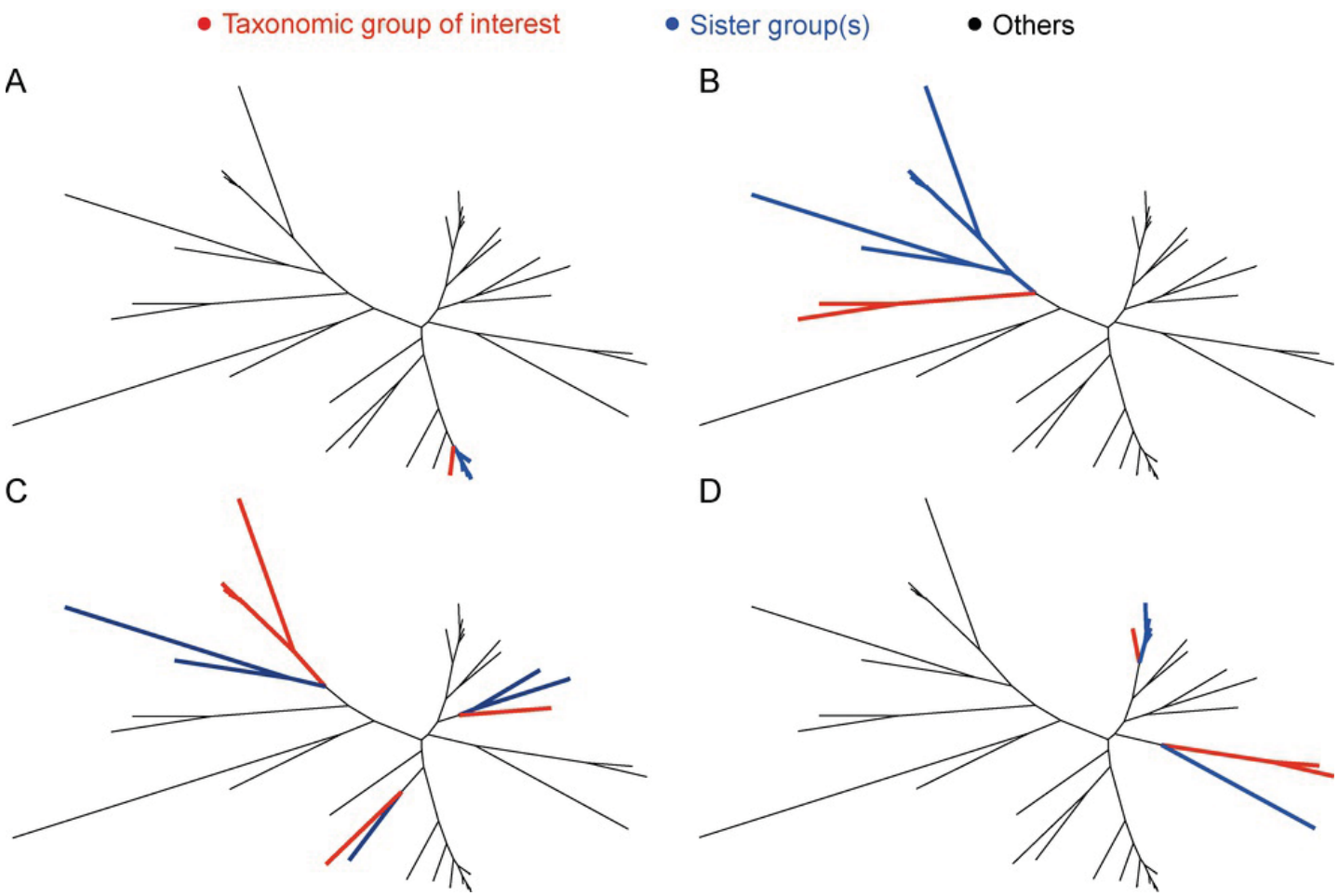


Fig 1. Identification of clades and sister groups. For each tree, largest possible clade(s) and their respective sister group(s) are identified for different taxonomic groups (e.g., eukaryotes or bacilli). One (A, B) or more (C, D) clades can be present for a single taxonomic group, with close (A), divergent (B, C), or both close and divergent sister groups.

For prokaryotic groups (Fig 2 and S2 Fig) and for eukaryotes (Fig 3), we plotted all values $I_{C\text{-}S}$ (Y-axis) that could be extracted from the 2,585 trees against the number of taxa (x-axis) in the clade for each comparison. For bacilli, α-proteobacteria and β-proteobacteria, we observed very recent transfers in the form of $I_{C\text{-}S}$ values of 1.00 (complete identity to the sister group), whereas for the two archaeal groups, the highest $I_{C\text{-}S}$ is only approaching 0.85. To compare quantitatively the relative frequency of high $I_{C\text{-}S}$ values, the singleton clades (i.e., only one taxon) in the respective taxonomic group were used as the reference. For each taxonomic group, a reference value was used as the lower bound of the high sequence identity characteristic of recent LGTs, which was calculated as the average of the singleton $I_{C\text{-}S}$ that are greater than or equal to their third quartile (S2 Table). If the $I_{C\text{-}S}$ of a clade is greater than or equal to this reference value, it is then a high-identity clade (HIC). All prokaryote groups exhibited numerous $I_{C\text{-}S}$ values above their reference line (Fig 2A-H).

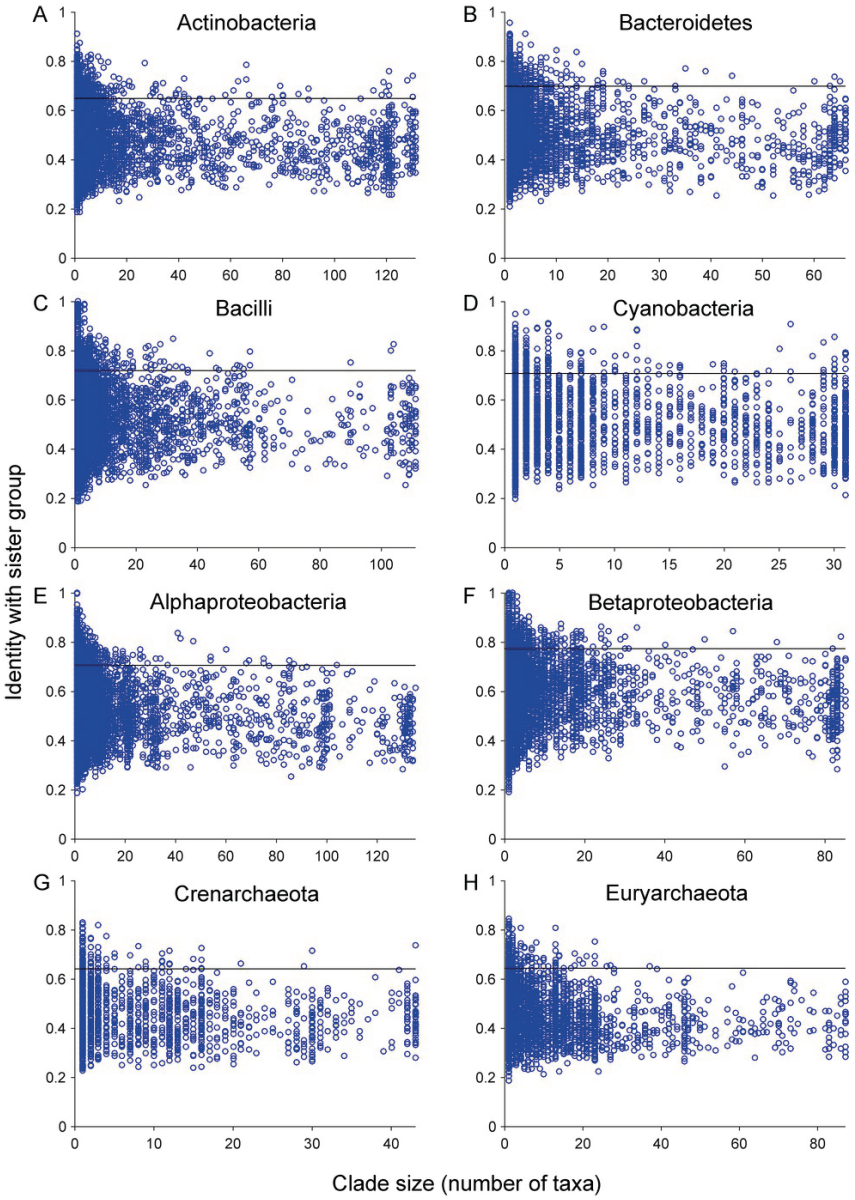


Fig 2. Phylogenomic dissection of major prokaryotic groups. All largest possible clades are plotted for each taxonomic group. Y-axis: average sequence identity between a clade and its sister group ($I_{C\text{-}S}$); x-axis: number of taxa (species in bacteria or genomes in archaea). A horizontal reference line is drawn corresponding to the average of the singleton $I_{C\text{-}S}$ greater than or equal to their third quartile.

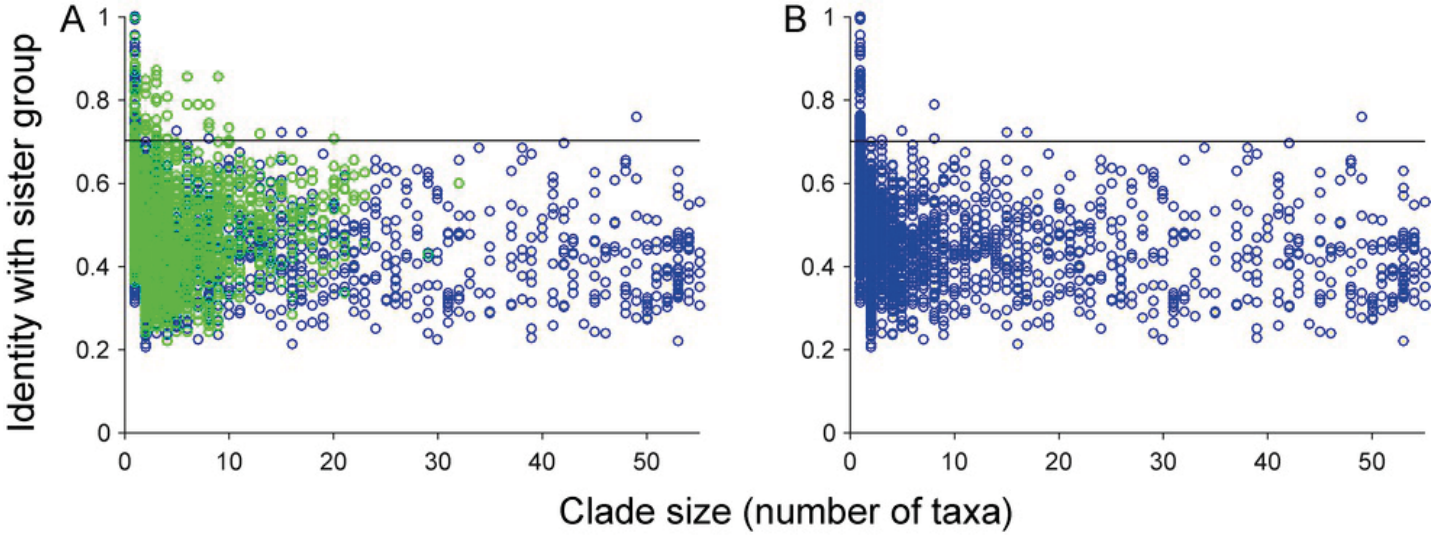


Fig 3. Phylogenomic dissection of eukaryotes. All largest possible eukaryotic clades are plotted. Y-axis: average sequence identity between a clade and its sister group ($I_{C\text{-}S}$); x-axis: number of species. A horizontal reference line is drawn corresponding to the average of the singleton $I_{C\text{-}S}$ greater than or equal to their third quartile. A. All clades. B. Clades of plastid-origin (shown in green in A) are selectively removed.

Deep differences between prokaryotes and eukaryotes

A list of the functional annotations for the top 10 clusters with the most conserved (most recent) acquisitions for each panel in Fig 2 and Fig 3 is given in S3 Table. These recently transferred genes encompass mainly metabolic functions, which is in line with the view that prokaryotes generate diversity mainly through acquisition, rather than through duplication [29].

By contrast, values of $I_{C\text{-}S}$ for eukaryotes were rarely above the reference value 0.70 (Fig 3). Two aspects of the eukaryote comparisons are particularly noteworthy. First, in Fig 3A the points plotted in green indicate $I_{C\text{-}S}$ values for genes of plastid origin (clusters in blocks A-C in [48] and other clades consisting of taxa only from Archaeplastida, from Archaeplastida and SAR or Hacrobia, or from all the three). The green points for $I_{C\text{-}S}$ above the reference value could in principle correspond to recent transfers, yet if we look at the functions involved (S4 Table), they are mainly plastid-related, such as phycobiliproteins, components of the extrinsic photosynthetic antenna complex found in some of the algal lineages. These are not recent acquisitions, rather they were acquired from cyanobacteria at the origin of primary plastids, as earlier investigations have shown [58]. Their high $I_{C\text{-}S}$ values reflect unusually high sequence conservation, not recent acquisition.

If we plot only the eukaryotic $I_{C\text{-}S}$ values for clades not of plastid origin (Fig. 3B), a very remarkable pattern comes to the fore in that only eight non-singleton HICs remain, including the clade E211_B160_0 (49 species; identity 0.76) of the ATP synthase subunit beta and E2540_B5394_A3181_1 (8 species; identity 0.79), where the sea anemone *Nematostella* sequence (jgi|Nemve1|78454|gw.12527.1.1) is nested within a clade otherwise specific to photosynthetic eukaryotes, it is probably a contamination (see below).

After the removal of clades of plastid origin, there are 69 singleton HICs. By singleton we do not mean proteins present only in one eukaryotic genome, because each tree has sequences from at least two eukaryotic genomes. Rather, singleton means that only one eukaryotic taxon is in the clade, separated from the other eukaryotic clade(s) in the tree. The identity and functional annotation of these eukaryotic singletons reveals that they mostly stem from the *Nematostella* and *Amphimedon* genome sequences. The genome sequence of *Nematostella* has an unexpected large number of predicted protein domains [59] and is known to contain many contaminating sequences from bacteria [60], which also seems to be the case for the genome sequence of the sponge *Amphimedon* [61].

That the singletons in the eukaryotic comparisons represent an anomaly is reflected in two further ways. First, if we plot the ratio of non-singleton to singleton HICs (Fig 4; S2 Table), the eukaryotes stand out and are significantly different from the prokaryotes at $p < 0.01$ for all clades or $p < 1 \times 10^{-6}$ when clades of plastid origin are removed (standard Pearson chi-square test; S5 Table). One factor that may influence the numbers of non-singleton and singleton HICs is the different clustering procedures for eukaryotes and prokaryotes [48], especially the different global identity cut-off for sequence pairs to be clustered (40% for eukaryotes and 25% for bacteria or archaea). This could result in a lower reference value in prokaryotes and might influence the ratio. To test this effect, we redid the analyses by clustering sequences of a bacterial group using the procedure for eukaryotes (see Materials and Methods). After the re-analyses (S3 Fig), prokaryotes are still significantly different from eukaryotes at $p < 0.01$ for all clades or $p < 1 \times 10^{6}$ for clades of non-plastid origin (S5 Table).

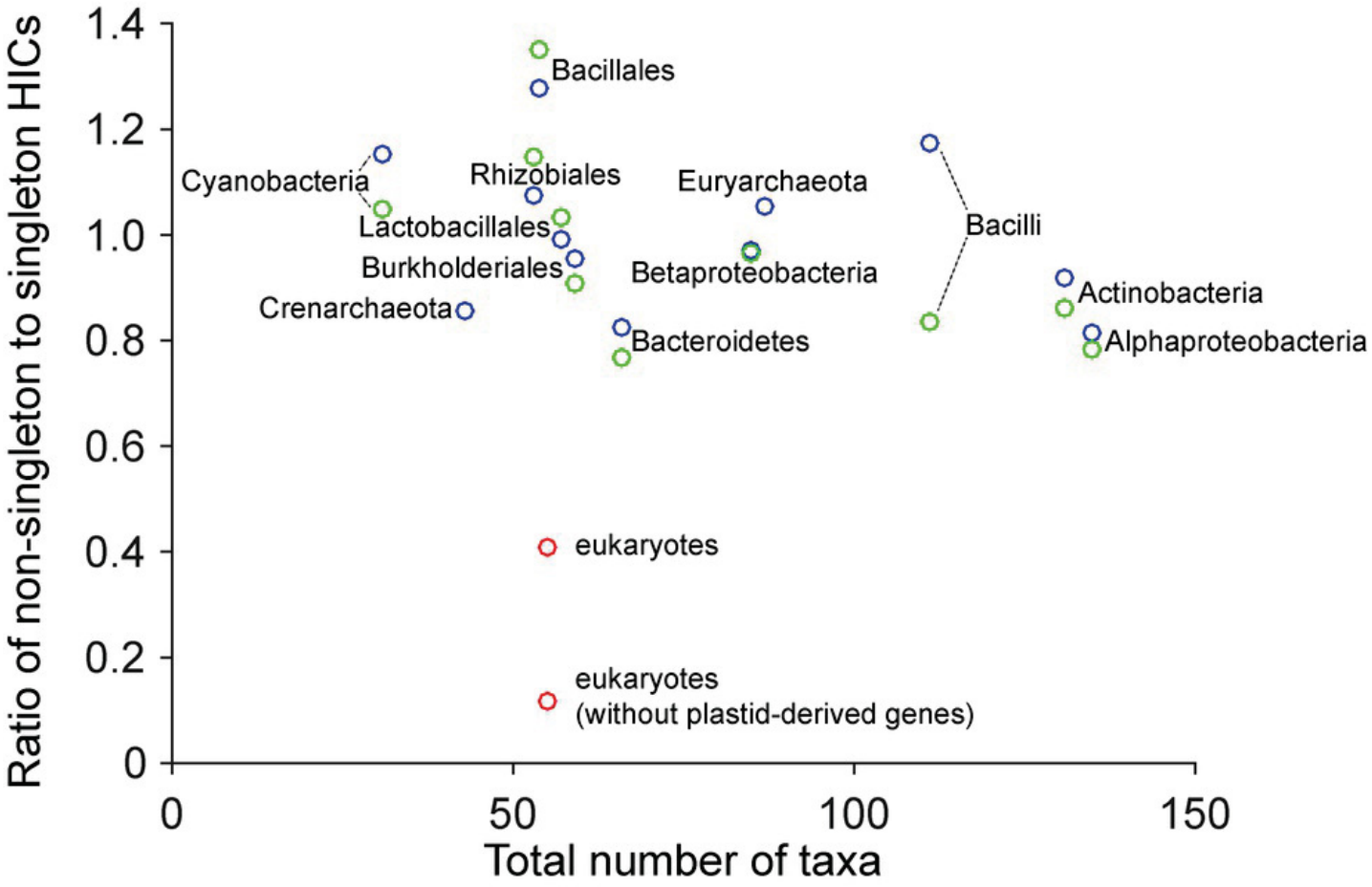


Fig 4. Eukaryotes have relatively fewer non-singleton high-identity clades (HICs) Taxonomic groups are plotted according to their ratio of non-singleton HICs to singleton HICs against their number of taxa. Red: eukaryotes with all clades (Fig 3A) or with clades of plastid-origin removed (Fig 3B); blue: prokaryotic groups based on the original eukaryote-prokaryote clusters (Fig 2; S2 Fig); green: prokaryotic groups based on clusters generated using the same clustering procedure as for eukaryotes (S3 Fig).

Second, if we zoom in on HICs that are up to one-third of the total taxa in size (Fig 5), we see that the prokaryotic acquisitions show a normal and expected tendency to become less similar to their sister group, the more taxa there are in the clade in question. In other words, genes acquired by prokaryotes can be transmitted vertically in the new lineage, and as they do so, they accumulate sequence divergence relative to the sister group, while at the same time lineage diversification takes place, such that the new gene is present in increasingly many descendant lineages (Fig 6). What we see in Fig 5 is basically a snapshot of continuous pangenome formation in prokaryotes, while in eukaryotes nothing of the sort is observed.

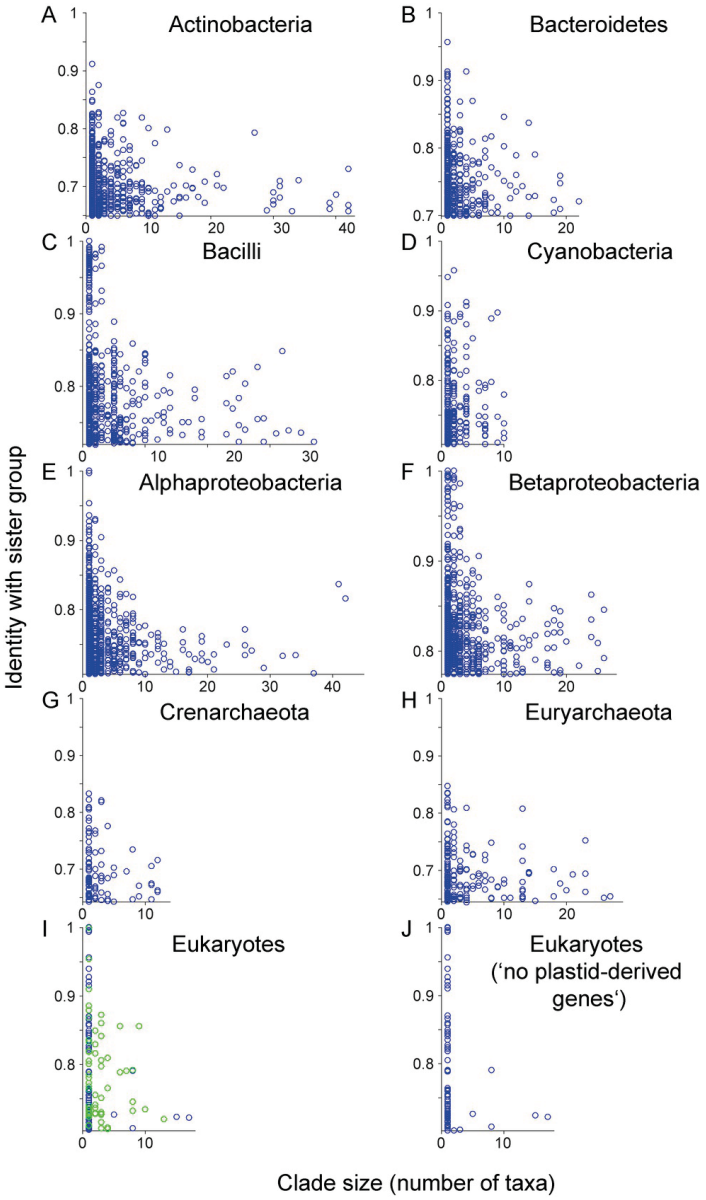


Fig 5. Close-up of the distribution of small-sized high-identity clades (HICs) HICs with up to the one-third of the total taxa are shown for each group in Fig 2 and Fig 3 (with x-axis plotted to the same scale for each group). A-H. Prokaryotic groups. I. All eukaryote clades. J. Eukaryotes with clades of plastid-origin (shown in green in I) selectively removed. The seven proteins having >70% sequence identity to prokaryotic homologues but appearing in more than one eukaryotic genome are annotated as (from left to right): fructose-bisphosphate aldolase, unknown (carbohydrate transport and metabolism), homocitrate synthase, component of cytochrome b6f complex, ribulose-phosphate 3-epimerase, pyridoxal biosynthesis, and adenosylhomocysteinase.

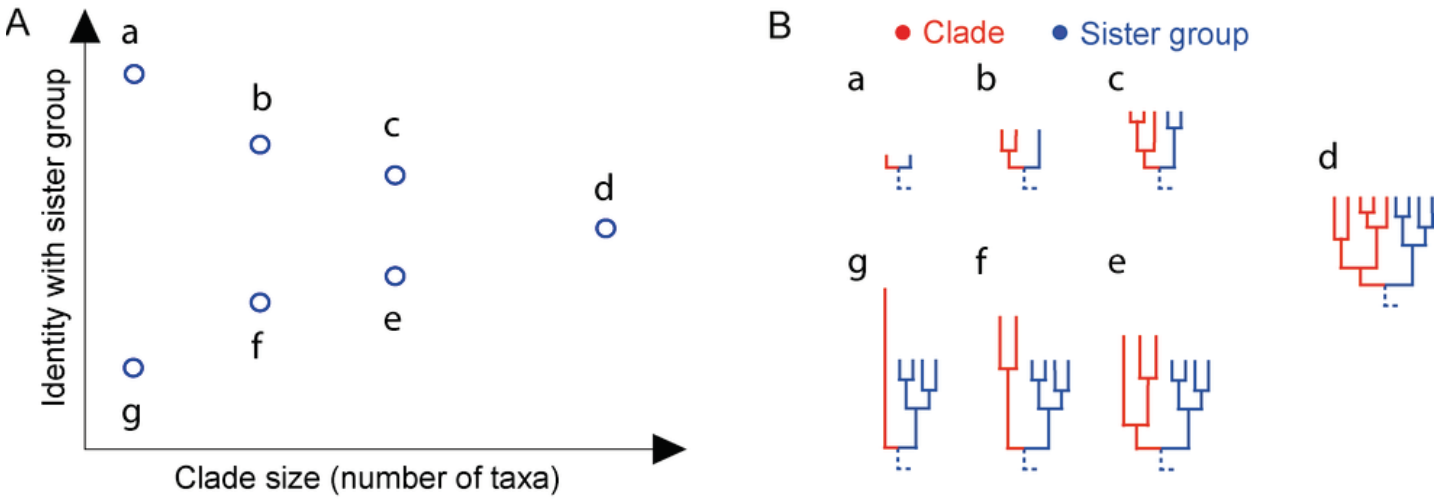


Fig. 6. Distribution of clades in the phylogenomic space.A. Seven representative clades are plotted in the phylogenomic space with clade-sister identity as the y-axis and clade size as the x-axis. B. Phylogenetic trees corresponding to the seven clades illustrate the effects of lineage diversification (a-d), sequence divergence (a-g), and differential gene loss (d-g).

That HICs of non-plastid origin are mainly restricted to singletons can mean one of two things. Either it suggests that eukaryotes do undergo lateral gene acquisition from prokaryotes, but that the acquisitions are very short-lived and do not persist to the lineage diversification stage, in which case they have no evolutionary significance at all. The more likely alternative is, however, that the singletons showing more that 70% amino acid identity to their closest prokaryotic homologue are simply contaminations that, during genome annotation procedures were scored as similar enough to eukaryotic homologues to represent a bona fide eukaryotic gene to be included in the assembly. The 70% amino acid identity threshold seems to be the result of a natural inter-domain barrier to LGT between prokaryotes and eukaryotes. Eukaryotic sequences that share ≥70% amino acid identity to prokaryotic homologues are probably not lateral gene transfers at all, they are probably just contaminants instead.

Discussion

In the present paper, we are asking a fairly simple but very controversial question: Are the many highly publicized claims for LGT from prokaryotes to eukaryotes real, or are they artifacts stemming from some combination of i) genome sequencing contaminations, ii) annotation practice, iii) phylogenetic reconstruction, iv) the underappreciated role of differential gene loss in eukaryote genome evolution or v) a combination of the above. Microbiologists have always known that prokaryotes undergo LGT *among themselves* to some degree [11]. Microbiologists furthermore anticipated the existence of pangenomes in that they built up to 30% difference in gene content into the species definition [62]. Genome sequences, however, have uncovered an extent of LGT among prokaryotes that no one really anticipated. For example, the current estimates for the pangenome size of a single species, *Escherichia coli*, based on 2,085 sequenced strains, are now at 90,000 genes and still climbing, linearly [63]. No mechanism other than LGT will produce pangenomes of that size, and the basic concept of LGT among prokaryotes has never been controversial because it is a natural process and meshes well with what we know about prokaryote biology.

So if we look back to 1998, when the first evidence for substantial LGT from genome sequence analyses was emerging [49], we can now be absolutely certain: Yes, there can be no doubt that LGT in prokaryotes is real, that it is ongoing, and that it reflects a very important aspect of prokaryote biology — natural variation through recombination. At the same time, endosymbiotic theory has always had it that many genes entered the eukaryotic lineage via the endosymbiotic ancestors of mitochondria and chloroplasts, of this we can also be certain [38, 48, 58, 64]. The basic concept of endosymbiotic gene transfer [65] has also never been controversial, because it is a natural process and meshes well with what we know about eukaryote biology.

The aspect of LGT that has been controversial, but perhaps not controversial enough in our view, concerns claims for outright LGT from prokaryotes to eukaryotes outside the context of endosymbiosis. Such claims were put forth in the human genome sequence [6], and they were promptly refuted as artifacts [7, 8]. New claims for prokaryote to eukaryote LGT soon emerged, they became popularized by LGT proponents [66], and soon thereafter many or most eukaryotic genome sequences published in high-profile journals contained reports (or claims) for more LGT [4, 52, 53]. Claims for LGT from chlamydiae to the plant lineage [42, 52, 67, 68] have been tested and rejected [48, 69-73]. Patchy gene distributions in eukaryotes are also often interpreted as evidence for LGT[40], without even considering the alternative: differential loss [48]. The high tide of prokaryote to eukaryote LGT claims might have been reached with the tardigrade showdown, where one group reported that 16.1% of all tardigrade nuclear genes are recent LGTs from prokaryotes [4], while a separate study found almost none at all [5].

If the claims from individual genome sequences for prokaryote-to eukaryote LGT are real, then it means that eukaryotes have indeed been continuously acquiring genes from prokaryotes over evolutionary time. That in turn predicts that we should then see two fundamental patterns in investigations of eukaryotic genome sequences. First, different lineages of eukaryotes should possess fundamentally different collections of genes, just as we see in prokaryotes [35, 74]. Second, genomes should harbour evidence for recently acquired genes in

eukaryote genomes, in addition to the anciently acquired genes that entered eukaryote genomes at the origin of mitochondria and plastids.

Few tests of either prediction have been reported. The obvious test for the first prediction (lineage specific gene acquisitions) is simple: if we investigate gene presence and absence across many different eukaryotic lineages, then genes that eukaryotes share with prokaryotes should reveal patterns of lineage specific acquisition. But the converse is observed, the only evidence for lineage specific gene acquisition in eukaryotes being the mass introduction of bacterial genes in the plant lineage corresponding to the origin of plastids and their subsequent spread during secondary symbiosis [48]. Lineage specific gene losses in eukaryotes are, by contrast, very common [48].

A thorough test of the second prediction (evidence for recent and ancient gene acquisitions) has been lacking. If eukaryotes are acquiring genes from prokaryotes continuously during evolution, then eukaryotic genomes should reveal evidence for recent acquisitions. Here we sought such evidence. We find that prokaryotes do indeed acquire genes from outside their phylum continuously during evolution, while eukaryotes do not. Prokaryotic phyla show a typical pattern of recent acquisitions that show up to 100% amino acid sequence identity to their sister-group homologues (Fig 2). The only examples of such high amino acid sequence identity between prokaryotic and eukaryotic genes are restricted to singleton clades, such as E2190_B358_A1066_1 and E2268_B77_0 from *Nematostella* (S6 Table), which is known to harbour many contaminations [60, 75]. There are a few proteins in plastid-bearing eukaryotes that exhibit >80% amino acid sequence identity to prokaryotic homologues, but these are mostly involved in photosynthetic functions, they are acquisitions that correspond to the origin of plastids (S4 Table).

If we look among the 2,386 clades of non-plastid origin, only very few proteins, such as mitochondrial ATPase, an acquisition corresponding to mitochondrial origin, have ≥70% amino acid sequence identity among proteins present in more than one eukaryotic genome. All other eukaryotic protein sequences showing ≥70% amino acid sequence identity to prokaryotic homologues are either i) acquisitions from the plastid ancestor or ii) are contaminations. Genes shared by prokaryotes and only one eukaryotic genome are suspects for contamination anyway. In the present study, we have queried 2,386 sequence comparisons, such that the paucity or absence of pairwise identity ≥70% between eukaryotic proteins present in more than one genome to prokaryotic homologues might be rather general. We call it the 70% rule.

If lineage specific acquisitions are extremely rare in eukaryotes, as the present data indicate, how can one explain the presence of lineage specific genes that are present in more than one genome? There are two ways to explain sparse gene distribution patterns: lineage specific acquisition or differential loss. If a gene is lost in one lineage, that means that it cannot be essential, hence it is possible for it to be lost in other lineages as well. Furthermore, loss is irreversible process, genes lost in one lineage will be missing in all descendants. If genes are indeed undergoing widespread loss in eukaryotes, as recent studies indicate [48], it follows that some genes will have been lost in all lineages but one. Such genes will have typical eukaryotic attributes, such as normal promoters and introns, and like normal prokaryotic genes they will be distantly related to prokaryotic homologues, but they will be lineage specific (but not genome-specific, like contaminations). This is exactly what is observed for genes that were interpreted as evidence for LGT in the *Galdieria sulphuraria* genome [53], a genome with claims for abundant LGT [76].

Here we aimed to summarize the effects of LGT in prokaryotic and eukaryotic genome evolution. Our findings indicate that eukaryotes do not acquire genes through continual LGT like prokaryotes do. Major gene acquisition do occur in eukaryote evolution, but these correspond to endosymbiotic events [48]. Unlike prokaryotes, where both vertical descent and LGT are major sources of genes, the gene repertoire of eukaryotes was already present in the complex last eukaryotic common ancestor [59], with the origin of plastids, and the spread of plastids via secondary symbiosis, and differential loss determining the distribution of genes across lineages.

Materials and Methods

Eukaryotic, archaeal and bacterial protein sequences were clustered separately and combined into 2,585 eukaryote-prokaryote clusters (EPCs) using the reciprocal best cluster approach as reported in a previous study [48]. Sequences within each cluster were aligned with MAFFT v7.130 [77], followed by maximum-likelihood tree inference using RAxML v7.8.6 [78]. The EPC functional annotations and trees are described in Supplementary Tables 6 and 7 in [48] respectively. For the purpose of this study, we searched across all the EPC trees for the largest possible clades from a taxonomic group (a clade is a largest possible clade if neither of the two neighboring clades consist only of taxa from that taxonomic group). The prokaryotic groups analyzed include two major archaeal subgroups, Euryarchaeota and Crenarchaeota, as well as Cyanobacteria and Alphaproteobacteria, from which the plastids and mitochondria arose, respectively [48, 79, 80]. In addition, other major bacterial phyla or classes and their large orders with a medium number (50 to 150) of taxa were included. For each largest possible clade, the sister group is defined as the neighboring clade with the smaller average branch distance (i.e., nearest neighbor). For the calculation of $I_{C\text{-}S}$ values, identities between all pairs of sequences from the clade and the sister group were calculated using the protdist program of PHYLIP v3.695 [81] and averaged. Standard Pearson chi-square tests were implemented using a script in MATLAB R2015a [82].

To test the effect of clustering procedures, new EPCs were generated for each of the ten bacterial groups analyzed. Their sequences were clustered using the same procedure (40% global identity cutoff; clusters with at least two sequences were retained) for clustering eukaryotic sequences, whereas the sequences from other bacteria were clustered using the original procedure for bacteria (25% global identity cutoff; clusters with at least five sequences were retained) [48]. These two sets of bacterial clusters were then combined into the complete bacterial set using the reciprocal best cluster approach, before it was combined with eukaryotic and archaeal clusters as for the original EPCs. Alignments and phylogenetic analyses were done for each set of reclustered EPCs as described above.

Acknowledgements

Work in the laboratory of WFM is supported by a grant from the European Research Council (AdvGr 666053). CK is supported by a PhD stipend from the German Academic Exchange Service (DAAD).

References

1. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol. 2001;55:709-42.

2. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000;405(6784):299-304. doi: 10.1038/35012500.

3. Doolittle WF. Phylogenetic classification and the universal tree. Science. 1999;284(5423):2124-8. doi: 10.1126/Science.284.5423.2124.

4. Boothby TC, Tenlen JR, Smith FW, Wang JR, Patanella KA, Osborne Nishimura E, et al. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proc Natl Acad Sci USA. 2015;112(52):15976-81. doi: 10.1073/pnas.1510461112.

5. Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, et al. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. Proc Natl Acad Sci USA. 2016;113(18):5053-8. doi: 10.1073/pnas.1600338113.

6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860-921.

7. Salzberg SL, White O, Peterson J, Eisen JA. Microbial genes in the human genome: Lateral transfer or gene loss? Science. 2001;292(5523):1903-6. doi: 10.1126/Science.1061036.

8. Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature. 2001;411(6840):940-4. doi: Doi 10.1038/35082058.

9. 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. 2015;16(1):50.

10. Boschetti C, Carr A, Crisp A, Eyres I, Wang-Koh Y, Lubzens E, et al. Biochemical Diversification through Foreign Gene Expression in Bdelloid Rotifers. PLoS Genet. 2012;8(11). doi: ARTN e1003035

10.1371/journal.pgen.1003035.

11. Jones D, Sneath PHA. Genetic transfer and bacterial taxonomy. Bacteriol Rev. 1970;34(1):40-81.

12. Popa O, Dagan T. Trends and barriers to lateral gene transfer in prokaryotes. Curr Opin Microbiol. 2011;14(5):615-23. doi: 10.1016/J.Mib.2011.07.027.

13. Griffith F. The significance of pneumococcal types. Journal of Hygiene. 1928;27(2):113-59.

14. Tatum EL, Lederberg J. Gene recombination in the bacterium*Escherichia coli*. J Bacteriol. 1947;53(6):673-84.

15. Zinder ND, Lederberg J. Genetic exchange in*Salmonella*. J Bacteriol. 1952;64(5):679-99.

16. Marrs B. Genetic recombination in*Rhodopseudomonas capsulata*. Proc Natl Acad Sci USA. 1974;71(3):971-3. doi: 10.1073/pnas.71.3.971.

17. Grohmann E, Muth G, Espinosa M. Conjugative plasmid transfer in gram-positive bacteria. Microbiol Mol Biol Rev. 2003;67(2):277-+. doi: 10.1128/Mmbr.67.2.277-301.2003.

18. Chen I, Christie PJ, Dubnau D. The ins and outs of DNA transfer in bacteria. Science. 2005;310(5753):1456-60. doi: 10.1126/science.1114021.

19. Lang AS, Zhaxybayeva O, Beatty JT. Gene transfer agents: phage-like elements of genetic exchange. Nat Rev Microbiol. 2012;10(7):472-82. doi: 10.1038/Nrmicro2802.

20. Li X, Heyer WD. Homologous recombination in DNA repair and DNA damage tolerance. Cell Research. 2008;18(1):99-113. doi: 10.1038/cr.2008.1.

21. Krejci L, Altmannova V, Spirek M, Zhao XL. Homologous recombination and its regulation. Nucleic Acids Res. 2012;40(13):5795-818. doi: 10.1093/nar/gks270.

22. Milkman R. Recombination and population structure in*Escherichia coli*. Genetics. 1997;146(3):745-50.

23. Lerat E, Daubin V, Ochman H, Moran NA. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 2005;3(5):807-14. doi: 10.1371/journal.pbio.0030130.

24. Dagan T, Martin W. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc Natl Acad Sci USA. 2007;104(3):870-5. doi: 10.1073/Pnas.0606318104.

25. Puigbo P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biol. 2014;12:66. doi: 10.1186/s12915-014-0066-4.

26. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. Trends Genet. 2009;25(3):107-10. doi: 10.1016/j.tig.2008.12.004.

27. Lobkovsky AE, Wolf YI, Koonin EV. Gene frequency distributions reject a neutral model of genome evolution. Genome Biol Evol. 2013;5(1):233-42. doi: 10.1093/Gbe/Evt002.

28. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol. 2015;23:148-54. doi: 10.1016/j.mib.2014.11.016.

29. Treangen TJ, Rocha EPC. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 2011;7(1):e1001284. doi: 10.1371/Journal.Pgen.1001284.

30. Shoemaker NB, Vlamakis H, Hayes K, Salyers AA. Evidence for extensive resistance gene transfer among Bacteroides spp. and among Bacteroides and other genera in the human colon. Appl Environ Microbiol. 2001;67(2):561-8. doi: 10.1128/Aem.67.2.561-568.2001.

31. Wiedenbeck J, Cohan FM. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev. 2011;35(5):957-76. doi: 10.1111/j.1574-6976.2011.00292.x.

32. Hilario E, Gogarten JP. Horizontal transfer of ATPase genes – the tree of life becomes a net of life. Biosystems. 1993;31(2-3):111-9. doi: 10.1016/0303-2647(93)90038-E.

33. Dagan T, Martin W. The tree of one percent. Genome Biol. 2006;7(10):118. doi: 10.1186/Gb-2006-7-10-118.

34. Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008;36(21):6688-719. doi: 10.1093/Nar/Gkn668.

35. Ku C, Nelson-Sathi S, Roettger M, Garg S, Hazkani-Covo E, Martin WF. Endosymbiotic gene transfer from prokaryotic pangenomes: inherited chimerism in eukaryotes. Proc Natl Acad Sci USA. 2015;112(33):10139-46. doi: 10.1073/pnas.1421385112.

36. McInerney JO, O'Connell MJ, Pisani D. The hybrid nature of the Eukaryota and a consilient view of life on Earth. Nat Rev Microbiol. 2014;12(6):449-55. doi: 10.1038/Nrmicro3271.

37. Williams TA, Foster PG, Cox CJ, Embley TM. An archaeal origin of eukaryotes supports only two primary domains of life. Nature. 2013;504(7479):231-6. doi: 10.1038/Nature12779.

38. Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. 2010;6(2):e1000834. doi: 10.1371/Journal.Pgen.1000834.

39. Sprague GF. Genetic exchange between kingdoms. Curr Opin Genet Dev. 1991;1(4):530-3. doi: 10.1016/S0959-437X(05)80203-5.

40. Andersson JO. Lateral gene transfer in eukaryotes. Cell Mol Life Sci. 2005;62(11):1182-97. doi: 10.1007/s00018-005-4539-z.

41. Yue JP, Hu XY, Sun H, Yang YP, Huang JL. Widespread impact of horizontal gene transfer on plant colonization of land. Nat Commun. 2012;3:1152. doi: 10.1038/Ncomms2148.

42. Qiu H, Yoon HS, Bhattacharya D. Algal endosymbionts as vectors of horizontal gene transfer in photosynthetic eukaryotes. Front Plant Sci. 2013;4. doi: 10.3389/Fpls.2013.00366.

43. Luan JB, Chen WB, Hasegawa DK, Simmons AM, Wintermantel WM, Ling KS, et al. Metabolic coevolution in the bacterial symbiosis of whiteflies and related plant sap-feeding insects. Genome Biol Evol. 2015;7(9):2635-47. doi: 10.1093/gbe/evv170.

44. 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. 2002;99(22):14280-5. doi: 10.1073/Pnas.222228199.

45. Kyndt T, Quispe D, Zhai H, Jarret R, Ghislain M, Liu QC, et al. The genome of cultivated sweet potato contains*Agrobacterium* T-DNAs with expressed genes: An example of a naturally transgenic food crop. Proc Natl Acad Sci USA. 2015;112(18):5844-9. doi: 10.1073/pnas.1419685112.

46. Huang JL. Horizontal gene transfer in eukaryotes: the weak-link model. Bioessays. 2013;35(10):868-75. doi: 10.1002/Bies.201300007.

47. Schönknecht G, Weber APM, Lercher MJ. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. Bioessays. 2014;36(1):9-20. doi: 10.1002/Bies.201300095.

48. Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, et al. Endosymbiotic origin and differential loss of eukaryotic genes. Nature. 2015;524(7566):427-32. doi: 10.1038/nature14963.

49. Lawrence JG, Ochman H. Molecular archaeology of the*Escherichia coli* genome. Proc Natl Acad Sci USA. 1998;95(16):9413-7. doi: 10.1073/pnas.95.16.9413.

50. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a global network of gene exchange connecting the human microbiome. Nature. 2011;480(7376):241-4. doi: 10.1038/nature10571.

51. Doolittle WE. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet. 1998;14(8):307-11.

52. Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber APM, et al.*Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. Science. 2012;335(6070):843-7.

53. Schönknecht G, Chen WH, Ternes CM, Barbier GG, Shrestha RP, Stanke M, et al. Gene Transfer from Bacteria and Archaea Facilitated Evolution of an Extremophilic Eukaryote. Science. 2013;339(6124):1207-10. doi: 10.1126/science.1231707.

54. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 2001;29:22-8.

55. Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003;4(1):41.

56. Thompson JD, Plewniak F, Poch O. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 1999;27(13):2682-90. doi: DOI 10.1093/nar/27.13.2682.

57. Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin EV. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. Nucleic Acids Res. 2005;33(14):4626-38. doi: 10.1093/nar/gki775.

58. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, et al. Evolutionary analysis of*Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA. 2002;99(19):12246-51. doi: 10.1073/pnas.182432999.

59. Zmasek CM, Godzik A. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. 2011;12(1):R4. doi: 10.1186/gb-2011-12-1-r4.

60. Artamonova II, Mushegian AR. Genome sequence analysis indicates that the model eukaryote*Nematostella vectensis* harbors bacterial consorts. Appl Environ Microbiol. 2013;79(22):6868-73. doi: 10.1128/Aem.01635-13.

61. Hentschel U, Piel J, Degnan SM, Taylor MW. Genomic insights into the marine sponge microbiome. Nat Rev Microbiol. 2012;10(9):641-54. doi: Doi 10.1038/Nrmicro2839.

62. Stackebrandt E, Goebel BM. A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol. 1994;44(4):846-9. doi: 10.1099/00207713-44-4-846.

63. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomic. 2015;15(2):141-61. doi: 10.1007/s10142-015-0433-4.

64. Timmis JN, Ayliffe MA, Huang CY, Martin W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 2004;5(2):123-35. doi: 10.1038/nrg1271.

65. Martin W, Brinkmann H, Savonna C, Cerff R. Evidence for a chimeric nature of nuclear genomes: Eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. Proc Natl Acad Sci USA. 1993;90(18):8692-6. doi: 10.1073/Pnas.90.18.8692.

66. Gogarten JP. Gene transfer: Gene swapping craze reaches eukaryotes. Curr Biol. 2003;13(2):R53-R4. doi: 10.1016/S0960-9822(02)01426-4.

67. Huang J, Gogarten J. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biol. 2007;8(6):R99.

68. Ball SG, Subtil A, Bhattacharya D, Moustafa A, Weber APM, Gehre L, et al. Metabolic effectors secreted by bacterial pathogens:

essential facilitators of plastid endosymbiosis? Plant Cell. 2013;25(1):7-21. doi: Doi 10.1105/Tpc.112.101329.

69. Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major P, et al. Genomes of stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. Genome Biol Evol. 2013;5(1):31-44. doi: 10.1093/Gbe/Evs117.

70. Moreira D, Deschamps P. What was the real contribution of endosymbionts to the eukaryotic nucleus? Insights from photosynthetic eukaryotes. Cold Spring Harb Perspect Biol. 2014;6(7):a016014. doi: 10.1101/cshperspect.a016014.

71. Deschamps P. Primary endosymbiosis: have cyanobacteria and Chlamydiae ever been roommates? Acta Soc Bot Pol. 2014;83(4):291-302. doi: 10.5586/asbp.2014.048.

72. Ku C, Roettger M, Zimorski V, Nelson-Sathi S, Sousa FL, Martin WF. Plastid origin: Who, when and why. Acta Soc Bot Pol. 2014;83(4):281-9. doi: 10.5586/asbp.2014.045.

73. Domman D, Horn M, Embley TM, Williams TA. Plastid establishment did not require a chlamydial partner. Nat Commun. 2015;6:6421. doi: 10.1038/ncomms7421.

74. Embley TM, Martin W. Molecular evolution - A hydrogen-producing mitochondrion. Nature. 1998;396(6711):517-9. doi: Doi 10.1038/24994.

75. Artamonova II, Lappi T, Zudina L, Mushegian AR. Prokaryotic genes in eukaryotic genome sequences: when to infer horizontal gene transfer and when to suspect an actual microbe. Environ Microbiol. 2015;17(7):2203-8. doi: 10.1111/1462-2920.12854.

76. Richards TA, Monier A. A tale of two tardigrades. Proceedings of the National Academy of Sciences. 2016. doi: 10.1073/pnas.1603862113.

77. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80. doi: 10.1093/Molbev/Mst010.

78. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22(21):2688-90.

79. Gray MW, Burger G, Lang BF. Mitochondrial evolution. Science. 1999;283:1476-81.

80. Pisani D, Cotton JA, McInerney JO. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol. 2007;24(8):1752-60. doi: 10.1093/Molbev/Msm095.

81. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 1996;266:418-27.

82. Thorvaldsen S, Fla T, Willassen NP. DeltaProt: a software toolbox for comparative genomics. BMC Bioinformatics. 2010;11:573. doi:

4 References

Adl SM, et al. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol 52: 399-451.

Albalat R, Cañestro C 2016. Evolution by gene loss. Nat Rev Genet 17: 379–391.

Allers T, Mevarech M 2005. Archaeal genetics - The third way. Nat Rev Genet 6: 58-73.

Amiri H, Karlberg O, Andersson SGE 2003. Deep origin of plastid/parasite ATP/ADP translocases. J Mol Evol 56: 137-150.

Andersson JO 2009. Gene transfer and diversification of microbial eukaryotes. Annu Rev Microbiol 63: 177-193.

Andersson JO 2005. Lateral gene transfer in eukaryotes. Cell Mol Life Sci 62: 1182-1197.

Archibald JM. 2014. One plus one equals one: symbiosis and the evolution of complex life: Oxford University Press.

Artamonova II, Lappi T, Zudina L, Mushegian AR 2015. Prokaryotic genes in eukaryotic genome sequences: when to infer horizontal gene transfer and when to suspect an actual microbe. Environ Microbiol 17: 2203-2208.

Artamonova II, Mushegian AR 2013. Genome sequence analysis indicates that the model eukaryote *Nematostella vectensis* harbors bacterial consorts. Appl Environ Microbiol 79: 6868-6873.

Avery OT, MacLeod CM, McCarty M 2000. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus Type III. J Exp Med 79: 137-158.

Ball S, Colleoni C, Cenci U, Raj JN, Tirtiaux C 2011. The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis. J Exp Bot 62: 1775-1801.

Ball SG, Bhattacharya D, Qiu H, Weber APM 2016a. Commentary: Plastid establishment did not require a chlamydial partner. Front Cell Infect Microbiol 6: 43.

Ball SG, Bhattacharya D, Weber APM 2016b. Pathogen to powerhouse. Science 351: 659-660.

Ball SG, et al. 2015. Toward an understanding of the function of Chlamydiales in plastid endosymbiosis. Biochim Biophys Acta 1847: 495-504.

Ball SG, Greub G 2015. Blurred pictures from the crime scene: the growing case for a function of Chlamydiales in plastid endosymbiosis. Microbes and Infection 17: 723-726.

Ball SG, et al. 2013. Metabolic effectors secreted by bacterial pathogens: essential facilitators of plastid endosymbiosis? Plant Cell 25: 7-21.

Baum D 2013. The origin of primary plastids: a pas de deux or a ménage a trois? Plant Cell 25: 4-6.

Becker B, Hoef-Emden K, Melkonian M 2008. Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. BMC Evol Biol 8: 203.

Bemm F, Weiß CL, Schultz J, Förster F 2016. Genome of a tardigrade: Horizontal gene transfer or bacterial contamination? Proc Natl Acad Sci USA 113: E3054–E3056.

Bonen L, Doolittle WF 1975. On the prokaryotic nature of red algal chloroplasts. Proc Natl Acad Sci USA 72: 2310-2314.

Boto L 2014. Horizontal gene transfer in the acquisition of novel traits by metazoans. Proc R Soc B 281: 20132450.

Brinkman FSL, et al. 2002. Evidence that plant-like genes in *Chlamydia* species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast. Genome Res 12: 1159-1167.

Carvunis AR, et al. 2012. Proto-genes and de novo gene birth. Nature 487: 370-374.

Cavalier-Smith T 2002. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int J Syst Evol Microbiol 52: 7-76.

Cencil U, et al. 2014. Transition from glycogen to starch metabolism in Archaeplastida. Trends Plant Sci 19: 18-28.

Chen I, Christie PJ, Dubnau D 2005. The ins and outs of DNA transfer in bacteria. Science 310: 1456-1460.

Chen I, Dubnau D 2004. DNA uptake during bacterial transformation. Nat Rev Microbiol 2: 241-249.

Dagan T, Artzy-Randrup Y, Martin W 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci USA 105: 10039-10044.

Dagan T, Martin W 2009. Getting a better picture of microbial evolution en route to a network of genomes. Philos Trans R Soc Lond B Biol Sci 364: 2187-2196.

Dagan T, et al. 2013. Genomes of stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. Genome Biol Evol 5: 31-44.

de Meeus T, Prugnolle F, Agnew P 2007. Asexual reproduction: Genetics and evolutionary aspects. Cell Mol Life Sci 64: 1355-1372.

Degli Esposti M, et al. 2014. Evolution of mitochondria reconstructed from the energy metabolism of living bacteria. PLoS One 9: e96566.

Deschamps P 2014. Primary endosymbiosis: have cyanobacteria and Chlamydiae ever been roommates? Acta Soc Bot Pol 83: 291-302.

Deschamps P, Moreira D 2009. Signal conflicts in the phylogeny of the primary photosynthetic eukaryotes. Mol Biol Evol 26: 2745-2753.

Deusch O, et al. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Mol Biol Evol 25: 748-761.

Domman D, Horn M, Embley TM, Williams TA 2015. Plastid establishment did not require a chlamydial partner. Nat Commun 6: 6421.

Doolittle WE 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet 14: 307-311.

Doolittle WF 1999a. Lateral genomics. Trends Biochem Sci 24: M5-M8.

Doolittle WF 1999b. Phylogenetic classification and the universal tree. Science 284: 2124-2128.

Dubey GP, Ben-Yehuda S 2011. Intercellular nanotubes mediate bacterial communication. Cell 144: 590-600.

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

Espinosa A, et al. 2001. The bifunctional Entamoeba histolytica alcohol dehydrogenase 2 (EhADH2) protein is necessary for amebic growth and survival and requires an

intact C-terminal domain for both alcohol dehydrogenase and acetaldehyde dehydrogenase activity. J Biol Chem 276: 20136-20143.

Esser C, et al. 2004. A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol 21: 1643-1660.

Facchinelli F, Colleoni C, Ball SG, Weber APM 2013a. Chlamydia, cyanobiont, or host: who was on top in the ménage à trois? Trends Plant Sci 18: 673-679.

Facchinelli F, et al. 2013b. Proteomic analysis of the *Cyanophora paradoxa* muroplast provides clues on early events in plastid endosymbiosis. Planta 237: 637-651.

Falcón LI, Magallón S, Castillo A 2010. Dating the cyanobacterial ancestor of the chloroplast. ISME J 4: 777-783.

Felsenstein J. 2004. Inferring phylogeny. Sunderland: Sinauer Associates, Inc.

Fleischmann RD, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496-512.

Fournier GP, Huang JL, Gogarten JP 2009. Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life. Philos Trans R Soc Lond B Biol Sci 364: 2229-2239.

Gelvin SB 2000. Agrobacterium and plant genes involved in T-DNA transfer and integration. Annu Rev Plant Physiol 51: 223-256.

Gogarten JP, Townsend JP 2005. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol 3: 679-687.

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

Gould SB, Waller RR, McFadden GI 2008. Plastid evolution. Annu Rev Plant Biol 59: 491-517.

Graur D. 2016. Molecular and Genome Evolution: Sinauer Associates.

Gray MW 1993. Origin and evolution of organelle genomes. Curr Opin Genet Dev 3: 884-890.

Gray MW 2014. The pre-endosymbiont hypothesis: a new perspective on the origin and evolution of mitochondria. Cold Spring Harb Perspect Biol 6: a016097.

Gray MW, Burger G, Lang BF 1999. Mitochondrial evolution. Science 283: 1476-1481.

Gray MW, Doolittle WF 1982. Has the endosymbiont hypothesis been proven? Microbiol Rev 46: 1-42.

Greub G, Raoult D 2003. History of the ADP/ATP-translocase-encoding gene, a parasitism gene transferred from a Chlamydiales ancestor to plants 1 billion years ago. Appl Environ Microbiol 69: 5530-5535.

Griffith F 1928. The significance of pneumococcal types. Journal of Hygiene 27: 113-159.

Harrison RG, Larson EL 2014. Hybridization, introgression, and the nature of species boundaries. J Hered 105: 795-809.

Hazkani-Covo E, Covo S 2008. Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. PLoS Genet 4: e1000237.

Hazkani-Covo E, Zeller RM, Martin W 2010. Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet 6: e1000834.

Heinemann JA, Sprague GF 1989. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. Nature 340: 205-209.

Horn M 2008. Chlamydiae as Symbionts in Eukaryotes. Annu Rev Microbiol 62: 113-131.

Horn M, et al. 2004. Illuminating the evolutionary history of chlamydiae. Science 304: 728-730.

Hotopp JCD, et al. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317: 1753-1756.

Howe CJ, Barbrook AC, Nisbet RER, Lockhart PJ, Larkum AWD 2008. The origin of plastids. Philos Trans R Soc Lond B Biol Sci 363: 2675-2685.

Huang CY, Grunheit N, Ahmadinejad N, Timmis JN, Martin W 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol 138: 1723-1733.

Huang J, Gogarten J 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biol 8: R99.

Huang J, Gogarten JP 2008. Concerted gene recruitment in early plant evolution. Genome Biol 9: R109.

Huang JL 2013. Horizontal gene transfer in eukaryotes: the weak-link model. BioEssays 35: 868-875.

Husnik F, et al. 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell 153: 1567-1578.

Ikeda H, Tomizawa JI 1965. Transducing fragments in generalized transduction by phage P1: I. Molecular origin of fragments. J Mol Biol 14: 85-109.

Johannsen WL. 1909. Elemente der exakten Erblichkeitslehre. Jena: Gustav Fischer.

Ju YS, et al. 2015. Frequent somatic transfer of mitochondrial DNA into the nuclear genome of human cancer cells. Genome Res 25: 814-824.

Karkar S, Facchinelli F, Price DC, Weber APM, Bhattacharya D 2015. Metabolic connectivity as a driver of host and endosymbiont integration. Proc Natl Acad Sci USA 112: 10208-10215.

Keeling PJ 2013. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annu Rev Plant Biol 64: 583-607.

Keeling PJ, Palmer JD 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9: 605-618.

Kidwell MG 1993. Lateral transfer in natural populations of eukaryotes. Annu Rev Genet 27: 235-256.

Kleine T, Maier UG, Leister D 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annu Rev Plant Biol 60: 115-138.

Konstantinidis KT, Ramette A, Tiedje JM 2006. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci 361: 1929-1940.

Koonin EV, Makarova KS, Aravind L 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55: 709-742.

Koutsovoulos G, et al. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. Proc Natl Acad Sci USA 113: 5053-5058.

Lang AS, Zhaxybayeva O, Beatty JT 2012. Gene transfer agents: phage-like elements of genetic exchange. Nat Rev Microbiol 10: 472-482.

Lang D, Zimmer AD, Rensing SA, Reski R 2008. Exploring plant biodiversity: the *Physcomitrella* genome and beyond. Trends Plant Sci 13: 542-549.

Lange BM, Rujan T, Martin W, Croteau R 2000. Isoprenoid biosynthesis: The evolution of two ancient and distinct pathways across genomes. Proc Natl Acad Sci USA 97: 13172-13177.

Larkum AWD, Lockhart PJ, Howe CJ 2007. Shopping for plastids. Trends Plant Sci 12: 189-195.

Lederberg J, Tatum EL 1946. Gene recombination in *Escherichia coli*. Nature 158: 558-558.

Linka N, et al. 2003. Phylogenetic relationships of non-mitochondrial nucleotide transport proteins in bacteria and eukaryotes. Gene 306: 27-35.

López-García P, Moreira D 1999. Metabolic symbiosis at the origin of eukaryotes. Trends Biochem Sci 24: 88-93.

Margulis L 1996. Archaeal-eubacterial mergers in the origin of Eukarya: Phylogenetic classification of life. Proc Natl Acad Sci USA 93: 1071-1076.

Margulis L. 1970. Origin of Eukaryotic Cells. New Haven, CT: Yale University Press.

Marrs B 1974. Genetic recombination in *Rhodopseudomonas capsulata*. Proc Natl Acad Sci USA 71: 971-973.

Marsh JA, Teichmann SA 2010. How do proteins gain new domains? Genome Biol 11.

Martin W, Brinkmann H, Savonna C, Cerff R 1993. Evidence for a chimeric nature of nuclear genomes: Eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. Proc Natl Acad Sci USA 90: 8692-8696.

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

Martin W, Kowallik KV 1999. Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche'. Eur J Phycol 34: 287-295.

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

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

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

Martin W, et al. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393: 162-165.

Martin WF, et al. 2012. Modern endosymbiotic theory: Getting lateral gene transfer into the equation. Journal of Endocytobiosis and Cell Research 23: 1-5.

McDaniel LD, et al. 2010. High frequency of horizontal gene transfer in the oceans. Science 330: 50-50.

Mereschkowsky C 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol Centralbl 25: 593-604 (English translation in: Martin W, Kowallik KV 1999. Eur J Phycol 1934: 1287-1295).

Moreira D, Deschamps P 2014. What was the real contribution of endosymbionts to the eukaryotic nucleus? Insights from photosynthetic eukaryotes. Cold Spring Harb Perspect Biol 6: a016014.

Moreira D, López-García P 1998. Symbiosis between methanogenic archaea and δ-proteobacteria as the origin of eukaryotes: The syntrophic hypothesis. J Mol Evol 47: 517-530.

Moustafa A, Reyes-Prieto A, Bhattacharya D 2008. Chlamydiae has contributed at least 55 genes to Plantae with predominantly plastid functions. PLoS One 3: e2205.

Mower JP, Bonen L 2009. Ribosomal protein L10 is encoded in the mitochondrial genome of many land plants and green algae. BMC Evol Biol 9: 265.

Naor A, Gophna U 2013. Cell fusion and hybrids in Archaea: prospects for genome shuffling and accelerated strain development for biotechnology. Bioengineered 4: 126-129.

Nass MMK, Nass S 1963. Intramitochondrial fibers with DNA characteristics: I. Fixation and electron staining reactions. J Cell Biol 19: 593-611.

Ohno S. 1970. Evolution by Gene Duplication: Springer Berlin Heidelberg.

Park AK, Kim H, Jin HJ 2009. Comprehensive phylogenetic analysis of evolutionarily conserved rRNA adenine dimethyltransferase suggests diverse bacterial contributions to the nucleus-encoded plastid proteome. Mol Phylogenet Evol 50: 282-289.

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

Pittis AA, Gabaldon T 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature 531: 101-104.

Price DC, et al. 2012. *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. Science 335: 843-847.

Qiu H, et al. 2013a. Assessing the bacterial contribution to the plastid proteome. Trends Plant Sci 18: 680-687.

Qiu H, Yoon HS, Bhattacharya D 2013b. Algal endosymbionts as vectors of horizontal gene transfer in photosynthetic eukaryotes. Front Plant Sci 4.

Raff RA, Mahler HR 1972. The non symbiotic origin of mitochondria. Science 177: 575-582.

Ravenhall M, Skunca N, Lassalle F, Dessimoz C 2015. Inferring horizontal gene transfer. PLoS Comput Biol 11.

Reyes-Prieto A, Moustafa A 2012. Plastid-localized amino acid biosynthetic pathways of Plantae are predominantly composed of non-cyanobacterial enzymes. Sci Rep 2: 955.

Rhymer JM, Simberloff D 1996. Extinction by hybridization and introgression. Annu Rev Ecol Syst 27: 83-109.

Richly E, Leister D 2004. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of Arabidopsis and rice. Gene 329: 11-16.

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

Rogers J, Gibbs RA 2014. Comparative primate genomics: emerging patterns of genome content and dynamics. Nat Rev Genet 15: 347-359.

Rosenshine I, Tchelet R, Mevarech M 1989. The mechanism of DNA transfer in the mating system of an archaebacterium. Science 245: 1387-1389.

Royo J, Gímez E, Hueros G 2000. CMP-KDO synthetase: a plant gene borrowed from Gram-negative eubacteria. Trends Genet 16: 432-433.

Rujan T, Martin W 2001. How many genes in *Arabidopsis* come from cyanobacteria? An estimate from 386 protein phylogenies. Trends Genet 17: 113-120.

Sagan L 1967. On the origin of mitosing cells. J Theor Biol 14: 225-274.

Sager R, Ishida MR 1963. Chloroplast DNA in *Chlamydomonas*. Proc Natl Acad Sci USA 50: 725-730.

Schatz G, Tuppy H, Haslbrun.E 1964. Deoxyribonucleic acid associated with yeast mitochondria. Biochem Biophys Res Commun 15: 127-132.

Schleiff E, Becker T 2011. Common ground for protein translocation: access control for mitochondria and chloroplasts. Nat Rev Mol Cell Bio 12: 48-59.

Schmitz-Esser S, et al. 2004. ATP/ADP translocases: a common feature of obligate intracellular amoebal symbionts related to Chlamydiae and Rickettsiae. J Bacteriol 186: 683-691.

Schönknecht G, Weber APM, Lercher MJ 2014. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. Bioessays 36: 9-20.

Semple C, Steel M. 2003. Phylogenetics. Oxford: Oxford University Press.

Soucy SM, Huang JL, Gogarten JP 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet 16: 472-482.

Stephens RS, et al. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282: 754-759.

Subtil A, Collingro A, Horn M 2014. Tracing the primordial Chlamydiae: extinct parasites of plants? Trends Plant Sci 19: 36-43.

Suzuki K, Miyagishima S 2010. Eukaryotic and Eubacterial Contributions to the Establishment of Plastid Proteome Estimated by Large-Scale Phylogenetic Analyses. Mol Biol Evol 27: 581-590.

Syvanen M 1985. Cross-species gene transfer; Implications for a new theory of evolution. J Theor Biol 112: 333-343.

Syvanen M 2012. Evolutionary implications of horizontal gene transfer. Annu Rev Genet 46: 341-358.

Tatum EL, Lederberg J 1947. Gene recombination in the bacterium *Escherichia coli*. J Bacteriol 53: 673-684.

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

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

Tyra HM, Linka M, Weber AP, Bhattacharya D 2007. Host origin of plastid solute transporters in the first photosynthetic eukaryotes. Genome Biol 8: R212.

Wallin IE. 1927. Symbionticism and the origin of species. London: Baillière, Tindall and Cox.

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

Wilson AC, Carlson SS, White TJ 1977. Biochemical evolution. Annu Rev Biochem 46: 573-639.

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

Wolf YI, Aravind L, Koonin EV 1999. Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange. Trends Genet 15: 173-175.

Wozniak RAF, Waldor MK 2010. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol 8: 552-563.

Yue JP, Hu XY, Sun H, Yang YP, Huang JL 2012. Widespread impact of horizontal gene transfer on plant colonization of land. Nat Commun 3: 1152.

Zinder ND, Lederberg J 1952. Genetic exchange in *Salmonella*. J Bacteriol 64: 679-699.

Danksagung

Mein erster Dank geht an meinen Doktorvater Prof. Dr. William F. Martin. Er ist ein genialer Wissenschaftler, großzügiger Mensch und ein Vorbild für mich, wenn ich je meine eigenen Doktoranden betreue.

Bei Prof. Dr. Martin Lercher möchte ich mich bedanken für die Übernahme der Rolle des Korreferenten.

Ich danke meinen jetzigen und ehemaligen Kollegen beim Institut für Molekulare Evolution für ihre Hilfsbereitschaft und ihr offenes Ohr bei Problemen: Dr. Mayo Röttger, Dr. Shijulal Nelson-Sathi, Dr. Verena Zimorski, Dr. Filipa Sousa, PD Dr. Sven Gould, Dr. Thorsten Thiergart, Dr. Christian Wöhle, Doris Matthée, Ariane Baab, Peter Melzer, Re-Young Yu, Harald Preisner, Dr. Gary Kusdian, Sriram Garg, Jan de Vries, Cessa Rauch, Dr. Gregor Christa, Nabor Lozada Chávez, Dr. Jörn Habicht, Madeline Weis, Sinje Neukirchen, Natalia Mrnjavac. Mein Dank geht auch an Dr. Giddy Landan und Prof. Dr. Tal Dagan an der Christian-Albrechts-Universität Kiel für ihren Rat.

Ich bedanke mich herzlich bei dem Deutschen Akademischen Austauschdienst (DAAD), mit dessen Unterstützung ich 2007 einen Sommerdeutschkurs in Düsseldorf besuchen konnte und jetzt auch hier promoviere. Durch DAAD habe ich Studierende aus aller Welt kennengelernt und ich danke besonders Dennis Daub und dem DAAD-Freundeskreis Köln für die vielfältigen Freizeit- und Kulturveranstaltungen.

Des Weiteren danke ich den anderen taiwanischen Doktoranden in Düsseldorf für die Hilfe im Alltag.

Zum Schluss muss ich mich noch bei meinen Eltern bedanken, ohne deren Unterstützung ich nicht da wäre, wo ich bin.

Eidesstattliche Erklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität“ erstellt worden ist. Die Dissertation wurde weder in der vorgelegten noch in ähnlicher Form bei einer anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 29.06.2016

Chuan Ku