Analysis of early evolutionary events during the transition from prokaryotes to eukaryotes

INAUGURAL DISSERTATION

For the attainment of the title of Doctor rerum naturalium (Dr. rer. nat.)
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

Julia Brückner, M.Sc.

from Oldenburg (Oldb), Germany

Düsseldorf, March 2021

From the Institute of Molecular Evolution
at the Heinrich Heine University Düsseldorf

Published by permission of the
Faculty of Mathematics and Natural Sciences at the
Heinrich Heine University Düsseldorf

Supervisor: Prof Dr. William F. Martin
Co-supervisor: Prof. Dr. Martin J. Lercher

Date of oral examination: June 17th, 2021

Statement of declaration

I hereby certify that this dissertation is the result of my own work. I confirm that no other person's work has been used without due acknowledgement. This work was done wholly while in the candidature for a research degree at the Heinrich Heine University Düsseldorf and has not been submitted in the same or similar form to other institutions. I have not previously failed a doctoral examination procedure.

Düsseldorf, August 10th, 2021

Julia Brückner

An meine Familie und Freunde.

Danke, dass ihr mich durchgehend unterstützt
und mir die wichtigen Dinge im Leben deutlich macht.

Acknowledgements

I never thought I would come this far in research. I didn't even think of studying after school but only wanted to do my apprenticeship, get a good job as a biological laboratory assistant and that's it. But as we all know: 'Life happens to you while you're busy making other plans' (John Lennon). Almost all my experiences in my studies at the Heinrich Heine University were lined with lucky events. One of them being the attendance of the genome analysis Master course by Dr. Mayo Röttger from Prof. Dr. William F. Martin's institute. I was lucky enough to get one of the last waiting list places of this course and found that I liked bioinformatics and the area of studying early evolution. And lucky as I was, there were enough free spaces in the institute at the time and I was able to do my Master thesis, which led to the PhD.

Therefore, I want to especially thank Prof. Martin for giving me the opportunity to work in his group and always encouraging me and helping me out. Without him I would probably have quit early on. Additionally, I want to thank Prof. Lercher for his role as co-supervisor. Mayo was responsible for me finding joy in bioinformatics and without Nils and Michael in our office, I would not have felt as welcomed as I did. The banter and scientific discussions and especially the help with programming that I got from those two (mostly Michael, Nils was the emotional support) helped me out a lot in continuing to work hard. Thank you, Verena and Renate, for helping with questions regarding the bureaucratic side of the PhD process and being there for any and all important matters. I also want to thank all other members of the team and students that I got to know during my four years at the institute, for creating a welcoming and sociable work environment.

Last but most importantly, I want to thank my mother, my brother and my best friends. You always support me, sometimes question me and my decisions (so I can make better ones), and most of all are always there if I need to talk. I feel special gratitude towards you for always believing in me and for believing that I could do it. Guess you were right. I really can do it after all. And I will never forget.

Publications submitted in the course of this thesis:

1. **Brueckner J**, Martin WF. 2020. Bacterial genes outnumber archaeal genes in eukaryotic genomes. *Genome Biol Evol*; 12: 282–292.

https://doi.org/10.1093/gbe/evaa047

2. Nagies FSP*, **Brueckner J***, Tria FDK, Martin WF. 2020. A spectrum of verticality across genes. *PLoS Genet*; 16: e1009200.

https://doi.org/10.1371/journal.pgen.1009200

3. Tria FDK*, **Brueckner J***, Skejo J, Xavier JC, Kapust N, Knopp M, Wimmer JLE, Nagies FSP, Zimorski V, Gould SB, Garg SG, Martin WF. 2021. Gene duplications trace mitochondria to the onset of eukaryote complexity. *Genome Biol Evol*; 13: evab055.

https://doi.org/10.1093/gbe/evab055

4. Xavier JC*, Gerhards R*, Wimmer JLE, **Brueckner J**, Tria FDK, Martin WF. 2021. The metabolic network of the last bacterial common ancestor. *Commun Biol*; 4: 413.

https://doi.org/10.1038/s42003-021-01918-4

* These authors contributed equally to the publication.

Abstract

There are two forms of living cells on earth — prokaryotes and eukaryotes. Prokaryotic cells are very simple organisms while eukaryotes present more complex cells that can organize into multicellular forms. The most widely accepted theory of eukaryote evolution is the endosymbiotic theory, depicting eukaryotes as descendants of archaea through the acquisition of a bacterial endosymbiont into its archaeal host. There has been more than one endosymbiotic event during eukaryote evolution with two major occurrences. The first gave rise to the mitochondrion from an archaeon incorporating a proteobacterium, generating the first eukaryote; the second resulted in the origin of the plant kingdom as a cyanobacterium was enveloped by an existing eukaryotic host. However, the mechanisms of how these events occurred are still mostly unknown and are the basis of many debates to date. Phylogenomic analyses have become the standard tool to investigate these early evolutionary events. Many studies focus on only a few genomes to represent a diverse spectrum of organisms from the three domains of life, some only examine a small number of genes. To obtain a more comprehensive view on how eukaryotes arose, the investigation of a broad spectrum of genomic data from a varied sample of lineages is desirable. Moreover, genome sequencing has become faster and simpler each year, enabling analyses containing a substantial number of organisms and genes. However, this is accompanied by extensively increased computational demands which can become a limiting factor that has to be addressed. This thesis reports the analysis of all protein coding genes from 5,655 prokaryotic and 150 eukaryotic genomes, the construction of their corresponding protein families, their alignments and phylogenetic trees in order to analyze evolutionary events at the border of prokaryotic and eukaryotic life.

Zusammenfassung

Es gibt zwei unterschiedliche Formen von Leben auf der Erde — Prokaryoten und Eukaryoten. Prokaryotische Zellen sind sehr einfache Organismen während Eukaryoten komplexe Zellen darstellen, die sich in multizelluläre Formen organisieren können. Die am weitesten verbreitete Hypothese der Evolution von Eukaryoten ist die Endosymbiontentheorie: die Aufnahme eines bakteriellen Endosymbionten in seinen archaeellen Wirt. Im Laufe der Evolution von Eukaryoten gab es mehrere endosymbiontische Ereignisse — als erstes entstand das Mitochondrion durch Aufnahme eines Proteobakteriums in ein Archaeon und später ist das Pflanzenreich durch den Einschluss eines Cyanobakteriums in einen bestehenden Eukaryoten entstanden. Allerdings sind die Mechanismen, welche sich hinter diesen Vorgängen befinden, immer noch unzureichend untersucht und die Basis fortlaufender Debatten. Mittlerweile sind phylogenomische Analysen eine der Standardmethoden, um diese frühen Evolutionsvorgänge zu erforschen. Viele Untersuchungen verwenden nur wenige Genome, um die drei Domänen des Lebens zu repräsentieren, oder fokussieren sich auf eine geringe Anzahl an Genen. Um jedoch einen umfassenderen Überblick von der Entwicklung der Eukaryoten zu bekommen, sollte ein breites Spektrum von Sequenzdaten aus Genomen aller bekannten taxonomischen Gruppen herangezogen werden. Da die Sequenzierung von Genomen jedes Jahr einfacher und schneller wird, können nun auch Untersuchungen mit einer umfangreichen Anzahl an Organismen und Genen durchgeführt werden. Allerdings ist damit eine gesteigerte Nutzung von Rechenkapazitäten verbunden, welches schnell ein limitierender Faktor wird, der beachtet werden muss. Diese Arbeit stellt die Analyse aller proteinkodierenden Gene aus 5,655 prokaryotischen und 150 eukaryotischen Genomen dar. Die daraus rekonstruierten Proteinfamilien und phylogenetische Stammbäume wurden zur Untersuchung von evolutionären Ereignissen an der Grenze von prokaryotischem und eukaryotischem Leben verwendet.

Table of contents

1 Introduction

It is human nature to be curious and wonder where we came from. Following this train of thought for any living organism at any taxonomic level brings forth the same question at every preceding generation of how each ancestor originated. The recursive leads backwards in time and, because all life forms share the same genetic code and therefore a single origin [Koonin and Novozhilov 2009], ultimately arrives at questions about early evolution. In this thesis, early evolution refers to the time span from the origin of life that took place about 4 billion years ago, to the great oxidation event facilitated by the origin of oxygenic photosynthesis about 2.5 billion years ago [Fischer *et al*. 2016] and the origin of eukaryotes roughly 1.6 billion years ago [Betts *et al*. 2018]. Questions about the nature of the first cells can be addressed by studying this phase of evolutionary history: how did the first cells live, which environments did they colonize and what innovations were required for their adaptations? Such questions are central to understanding the course of early evolution and the relationship between the Earth's environment and life [Arndt and Nisbet 2012; Nisbet and Sleep 2001; Sleep *et al*. 2011; Sleep 2018].

The first organisms that inhabited the earth were prokaryotes, morphologically simple cells with vast biochemical diversity. Prokaryotes are ubiquitously distributed around the earth — living in varying habitats of ocean floors, open ocean, terrestrial soil and terrestrial subsurface. They are estimated to comprise up to 14% of the global biomass [Bar-On *et al*. 2018; Kallmeyer *et al*. 2012; Whitman *et al*. 1998]. The first prokaryotes arose more than 3.8 billion years ago [Sleep 2018], possibly at submarine hydrothermal vents, as genomic reconstructions of the lifestyle and habitat of the last universal common ancestor (LUCA) suggest [Weiss *et al*. 2016]. Following roughly 2 billion years of biochemical diversification, prokaryotes gave rise to eukaryotes roughly 1.6 billion years ago [Betts *et al*. 2018]. Eukaryotes are complex cells possessing a nucleus and mitochondria. The question of how eukaryotes arose is still debated, but many lines of evidence indicate that eukaryote origin involved the endosymbiotic acquisition of a proteobacterial endosymbiont [Martin *et al*. 2015] into an archaeal host [Imachi *et al*. 2020].

There are several main approaches currently used to investigate questions about prokaryotic evolution. Isotopes [Arndt and Nisbet 2012; Nisbet and Sleep 2001; Sleep *et al*. 2011; Sleep 2018], microfossils [Javaux 2019], molecular clocks [Betts *et al*. 2018], and phylogenetic trees [Nelson-Sathi *et al*. 2012; Woese 1987] are the most widely used data employed to investigate prokaryote and early evolution. At present, phylogenetic trees are the most frequently used resource to study prokaryote evolution, owing to the vast amounts of information that genome sequencing technologies have generated. Gene and protein sequence comparisons are the most common way to investigate evolutionary relationships among molecules and genomes [Graur 2016]. Protein structure offers a more sensitive way to detect protein homology than protein sequences alone [Alva *et al*. 2015; Lupas and Alva 2017], but methods that are able to quantify differences in the three-dimensional structure in a way to scale evolutionary divergence are currently lacking. Therefore, structural comparisons are not, at present, a generally applicable tool for inference of phylogenetic differences.

Since the pioneering work of Fitch and Margoliash [1967], protein sequences have been used for reconstructing phylogenetic trees. Trees can be applied in different ways. Single gene trees or concatenation of proteins and subsequent reconstruction of concatenated gene trees are the traditional ways to analyze protein phylogeny [Adam *et al*. 2018]. With the ongoing growth of whole sequenced genomes and constant improvement of bioinformatics tools for phylogenetic research, it is now possible to investigate the information encoded in all protein coding genes of a given genome set [Coleman *et al*. 2020; Ku *et al*. 2015; Weiss *et al*. 2016; Williams *et al*. 2017]. Yet, phylogenetic trees constructed for different proteins almost always conflict. This can be due to different proteins having fundamentally different evolutionary histories [Popa and Dagan 2011], or it can be due to phylogenetic differences generated by inaccuracies in the process of phylogenetic reconstruction [Semple and Steel 2003]. Though phylogenetic analyses no longer require that all phylogenetic signals readily fit on one single tree

[Coleman *et al*. 2020; Williams *et al*. 2017], the issue of how to extract evolutionary insights out of conflicting phylogenetic trees is still a topic of investigation.

For that reason, investigating early evolution, prokaryotic evolution, or physiological evolution from the standpoint of gene trees alone carries caveats. The biological implications of deep phylogenetic inference typically hinge upon one branch in a phylogenetic tree [Martin *et al*. 1998; Nelson-Sathi *et al*. 2012], or the specific placement of a root among a set of short internal branches [Nelson-Sathi *et al*. 2015]. In both cases, it is becoming increasingly evident that with large data sets the results of such investigations tend to depend more upon specific procedures of phylogenetic inference than upon the data themselves [Fan *et al*. 2020]. Such circumstances lead to debates about phylogenetic methodology [Da Cunha *et al*. 2018], which are useful, but do not alleviate the problem that sequence data have fundamental limitations for the study of early evolution.

The present work was designed to process a large dataset of completely sequenced genomes from prokaryotic and eukaryotic organisms in order to reconstruct and investigate all phylogenetic trees that can be reconstructed from this data. For this, the first step was to detect sequence homologs — genes that are similar by virtue of shared ancestry — that could be sorted into protein families of the various organisms for phylogenetic inference by using protein sequence clustering methods [Enright *et al*. 2002]. The purpose of the work was to harness all the information available in those genomes with the goal of analyzing evolutionary events that speak to the nature of genetic interactions between organisms at the threshold of eukaryote origin.

1.1 Protein sequences, gene sequences and genome sequences for phylogenetics

Historically, the first phylogenetic analyses were constructed on the basis of comparative morphology [Darwin 1859; Haeckel 1866, 1874]. That restricted the investigation of evolution to organism groups that possessed sufficient morphological characters for comparison: plants, animals and fungi. Isolated attempts were made to reconstruct the evolution of microbes using comparative physiology [Chatton 1925; Lippmann 1965; Martin and Kowallik 1999; Mereschkowsky 1905], but there was no way to objectively test any phylogenetic scheme for microbes that was constructed from physiological data. The discovery of the linear structure DNA and its ability to store the genetic information of an organism [Watson and Crick 1953] and the development of protein sequencing technologies [Edman 1950; Sanger 1945; Sanger and Thompson 1953] gave rise to the analysis of evolutionary traits based on sequence homology and later the delineation of families of related proteins. Zuckerkandl and Pauling [1965] were the first to suggests that gene sequences might be used to analyze the relationships between organisms. By investigating the most suitable molecules as a basis for the reconstruction of phylogenetic trees, they pointed the way to exploration of molecular similarities and differences for phylogenetic inference. Margaret Dayhoff [Dayhoff and Eck 1968] laid the groundwork for generating alignments of homologous sequences by generating matrices that scored not just the presence of identical amino acids at homologous sites in a sequence, but also similar amino acids based on their physical and chemical properties.

The development of DNA sequencing techniques [Sanger *et al*. 1977] and its subsequent application to whole genome sequencing projects enabled molecular phylogenetic analyses of proteins that could not be directly sequenced by protein sequencing methods [Graur and Li 2000]. In 1977, the use of phylogenetic trees for the nucleotide sequence data from ribosomal RNA (rRNA) was employed to introduce the classification of prokaryotes into archaebacteria (archaea) and eubacteria (bacteria) in addition to urkaryotes (eukaryotes) [Woese and Fox 1977].

Due to the similarity of the ribosome sequences, archaea were recognized as relatives of the host that acquired mitochondria, and later plastids, during the early evolution of eukaryotes. Using trees for rRNA sequences, Woese, Kandler and Wheelis [1990] introduced the 'three domains of life'. Following this proposal, the reconstruction of phylogenetic trees according to the rRNA sequences became the standard method to reconstruct one universal tree of life.

In 2006, Ciccarelli *et al*. developed an automated procedure to reconstruct 'the' tree of life, which identified a set of 31 universally present proteins from 191 species that were included in their analysis. However, these 31 proteins were already in wide use at that time [Charlebois and Doolittle 2004; Hansmann and Martin 2000]. In the study presented by Ciccarelli *et al*., lateral gene transfer (LGT) events were detected and eliminated in order to reconstruct a highly resolved phylogenetic tree as LGT events are not tree-like and confuse phylogenetic signals [Ciccarelli *et al*. 2006; Hansmann and Martin 2000]. This 'tree of life' was generated by concatenating the orthologs of the detected 31 genes from all 191 genomes employed in the study. Because an average prokaryotic genome encodes about 3,000 protein sequences and the mean genome size of eukaryotes across all supergroups is roughly 23,000 protein coding genes, a tree of life reconstructed from 31 genes only represents about 1% of all prokaryotic and about 0.1% of eukaryotic genes in an average genome [Dagan and Martin 2006].

The detection of some evolutionary processes such as lateral gene transfer or gene duplication is only possible with a sample encompassing a large number of genes or genomes. Therefore, the reconstruction of a tree of life from 31 genes presents a very narrow view of the evolution of organisms. Contrasting to endeavors to reconstruct a tree-like structure of evolutionary events, Maria Rivera and James Lake [2004] introduced the ring of life. They figured that an acyclic graph better represents the evolutionary processes between eukaryotes and prokaryotes. Accounts of LGT and genome fusion are represented in their analyses which shows that eukaryotes may be a sister group to archaea and bacteria simultaneously. Thus, a more complete picture of the evolution of organisms can be obtained by regarding all available information inherent in sequence data and reconstructing protein families from those sequences.

1.2 Large-scale protein family reconstruction

Automated sequencing techniques have advanced quickly and become more affordable each year [Zhao and Grant 2011], which has enabled large-scale sequencing of whole genomes in a fast and dependable manner. The rapidly increasing number of sequenced genomes now widely accessible in databases, paired with the development of computerized calculations of alignments of thousands of homologous sequences simultaneously [Adams *et al*. 1992], initiated the age of evolutionary genome analyses or phylogenomics. This permitted one, in principle, to group organisms according to the sum of evolutionary processes that are recorded in their genomes [Tatusov *et al*. 1997], as opposed to groupings by morphological and biochemical traits [Darwin 1859; Haeckel 1866, 1874]. But the availability of genome data did not solve the problem of how to deal with conflicting signals from different genes. Part of this problem is rooted in the process of identifying sequence similarities and the subsequent decision of whether the sequence similarity is based in random processes, gene duplication (paralogy), gene transfer (xenology) or vertical evolution (orthology) [Fitch 1970; Kristensen *et al*. 2011; Roth *et al*. 2008].

1.2.1 The importance of detecting orthologous proteins

During speciation, the genetic information inherent in nucleotide sequences — and consequently in the corresponding amino acid sequences — diverge over time. Orthologous genes from different genomes can be used for the reconstruction of the phylogenies of those organisms, facilitating the classification of species [Fitch 1970; Fitch *et al*. 1995]. The first step in reconstructing protein families from molecular sequences involves the detection of orthologous genes while ideally removing paralogous sequences from the dataset [Fitch 1970]. Paralogs are sequences that are derived from gene duplication events in the same genome whereas orthologs represent speciation events [Fitch 1970]. For the purpose of constructing reliable gene trees for phylogenic inference, orthologous genes represent the best way to generate clusters of homologous sequences that can be applied for protein function inference or annotation and phylogenetic analyses [Roth *et al*. 2008]. But investigating gene duplication events enables the analysis of different evolutionary questions such as the transition from the first eukaryotic common ancestor (FECA) to the last eukaryotic common ancestor (LECA) [Tria *et al*. 2021] as gene duplication is one of the major forces of evolution [Ohno 1970]. However, in prokaryotes, gene duplication is much less common than in eukaryotes [Treangen and Rocha 2011].

The question of gene orthology is extremely crucial for phylogenetic inference studies and the functional annotation of genes, therefore several research groups formed the 'Quest for Orthologs' consortium in 2009 [Gabaldón *et al*. 2009] to define benchmark approaches for the existing and new algorithms for the detection of orthologous sequences and to determine standardized protein sets. These include methods that identify phylogenies by comparing gene trees with species trees, such as NOTUNG [Chen *et al*. 2000], Orthostrapper [Storm and Sonnhammer 2002], RIO [Zmasek and Eddy 2002], PhyOP [Goodstadt and Ponting 2006], PhiGs [Dehal and Boore 2006]), LOFT [van der Heijden *et al*. 2007], and Ensembl Compara [Hunt *et al*. 2018; Vilella *et al*. 2009] as well as graph-based methods, in which the first step is the identification of pairwise sequence alignments (BBH [Mushegian and Koonin 1996], COG [Tatusov *et al*. 1997], InParanoid [Remm *et al*. 2001], Panther [Thomas *et al*. 2003], OrthoMCL [Li *et al*. 2003], OMA [Dessimoz *et al*. 2005], eggNOG [Jensen *et al*. 2008], and HomoloGene [Wheeler *et al*. 2008]). Additionally, there are ortholog detection methods that combine multiple orthology prediction models by introducing a confidence score (MetaPhOrs [Pryszcz *et al*. 2011]). These algorithms can be either manual or fully automated and may have an optional step for protein family reconstruction and annotation. The large number of methods indicates that no optimal solution has been found. There is a trade-off for specificity and sensitivity values in benchmarking analyses of graph-based methods compared to tree-based models. Graph-based algorithms mostly perform well in sensitivity but have low specificity (high false positive, but low false negative rates in ortholog detection) which results in large and very inclusive (paralog-rich) protein clusters, whereas tree-based algorithms generally have higher specificity scores and low sensitivity and generate smaller clusters [Chen *et al*. 2007; Hulsen *et* *al*. 2006] in which ancient paralogs fall into different clusters. These trade-offs need to be considered before setting up a routine that culminates in clusters destined for multiple alignments and phylogenetic analysis.

Graph-based detection models can be separated into two categories based on the inference of orthologous sequences by 1) pairwise or 2) multi-species gene comparisons. Algorithms such as Roundup [DeLuca *et al*. 2006; DeLuca *et al*. 2012], InParanoid [Remm *et al*. 2001; Sonnhammer and Östlund 2015] or the detection of reciprocal (bidirectional) best BLAST (basic local alignment search tool) hits (rBBH) [Wolf and Koonin 2012] fall into the first category while multi-species comparisons are employed by methods like OMA (orthologous matrix) [Dessimoz *et al*. 2005; Zahn-Zabal *et al*. 2020], COG (clusters of orthologous groups) [Galperin *et al*. 2019; Tatusov *et al*. 1997], eggNOG (evolutionary genealogy of genes: non-supervised orthologous groups) [Huerta-Cepas *et al*. 2019; Jensen *et al*. 2008], or OrthoMCL [Fischer *et al*. 2011; Li *et al*. 2003].

BLAST is based on a heuristic approximation of the Smith-Waterman algorithm [Smith and Waterman 1981] to expedite the generation of local sequence alignments. Consequently, algorithms based on the calculation of all-vs-all pairwise sequence alignments with BLAST [Altschul *et al*. 1990] and the subsequent filtering for rBBH are some of the most straightforward and efficient methods to detect orthologs [Dalquen and Dessimoz 2013; Wolf and Koonin 2012]. A sequence pair is considered a reciprocal hit if there is an existing alignment between gene 1 and gene 2 that is also found in the other direction (gene 2 and gene 1). As orthologous sequences are related by species divergence, they tend to be each other’s nearest neighbors in reciprocal comparison of two genomes [Enright *et al*. 2002]. Thus, filtering for rBBH results in a list of sequence pairs related through local sequence identities that are most likely to be orthologs.

As another pairwise method, Roundup was designed to efficiently detect orthologous sequences in large datasets with relatively low resource demand [DeLuca *et al*. 2012]. The reciprocal smallest distance employed by the algorithm combines rBBH with global alignments and maximum likelihood evolutionary distances between sequences. This method may represent an effective algorithm for large sequence datasets, but the search is limited to the genomes from the UniProt database, which is included in the Roundup database spanning eukaryotes, bacteria, archaea, viruses, and viroids, but represents only a fraction of available sequence data.

Processes to improve upon the pairwise method exist. The National Center for Biotechnology Information (NCBI) employs an algorithm that uses not only bidirectional pairwise gene comparisons to detect orthologs for the reconstruction of their COGs but detects triangles of best BLAST hits after removing paralogous sequences [Galperin *et al*. 2019; Tatusov *et al*. 2000], which represents a more stringent approach for the filtering of orthologous sequences. The OMA algorithm [Dessimoz *et al*. 2005] employs a strategy of filtering for rBBH as an initial step while removing out-paralogs and leaving in-paralogs and thus generating a database of pairwise orthologs. Subsequently, hierarchical orthologous groups (HOGs) and OMA groups (maximum weight cliques) are reconstructed by building a network of sequences connected by their pairwise identity and clustering the proteins into families [Zahn-Zabal *et al*. 2020]. In contrast to other algorithms, the full Smith-Waterman algorithm [Altenhoff *et al*. 2018; Roth *et al*. 2008] is applied generating more accurate rBBH compared to alignments based on BLAST searches. However, this requires considerably more resources which makes it impractical for larger datasets.

A number of tree-based orthology prediction algorithms use tree reconciliation of the gene trees with a species tree and are therefore considered more accurate in predicting paralogous sequences but also require more computational resources compared to graph-based methods [Lechner *et al*. 2014]. However, phylogenetic inference has several practical issues such as LGT, false identification of homologous sequences, and variability of evolutionary rate [Brocchieri 2001] which limit the accuracy of the species tree employed for tree reconciliation. Examples of tree-based prediction algorithms that employ tree reconciliation techniques are PhylomeDB [Huerta-Cepas *et al*. 2008], TreeFam [Li *et al*. 2006; Schreiber *et al*. 2014], and Ensembl Compara [Vilella *et al*. 2009].

To circumvent the issue of applying species trees for tree reconciliation, the algorithm of LOFT (levels of orthology from trees) [van der Heijden *et al.* 2007] uses a species overlap rule allowing the detection of duplication events without information about the evolutionary history of the involved species. This method also employs a species tree but does not focus on the tree topology. However, tree-based orthology prediction algorithms are generally not suitable for large-scale protein family reconstruction among prokaryotic genomes, such as those performed in the course of this work, as there are currently no accepted species trees available that would serve as a standard for comparison for all of the thousands of genomes currently available. Generating such phylogenies was one of the aims of this thesis.

Comparing the different algorithms for orthology inference, the standard bidirectional best BLAST hit approach is best suited for a large dataset such as applied in this work. The filtering for rBBH is the most flexible and easily adapted for bigger datasets. Other algorithms are either higher in resource demand or are limited by the genomes included in their databases. Furthermore, the filtering for rBBH often outperforms or performs similarly as well as the more complex algorithms of orthology projects [Altenhoff and Dessimoz 2009; Altenhoff *et al.* 2016]. The relatively low resource demand, easy and flexible application paired with the generally good results in benchmarking studies determined the decision to employ rBBHs for the analyses included in the present work.

1.2.2 Influence of stringency on protein family properties

One of the critical parameters of protein family reconstruction is the stringency threshold of the pairwise global sequence alignments. The stringency is expressed as the amino acid sequence identity of the employed matrix. Sequence similarity of a high stringency such as 50-60% amino acid sequence identity gives rise to small clusters of homologous protein sequences (protein families) while creating a larger number of orthologous groups. A low stringency, for example in the range of 20% up to 30%, generates a smaller number of more inclusive clusters [Ku *et al.* 2015]. The present work was conducted with the purpose of examining aspects of prokaryotic and of eukaryotic genome evolution, especially the intersection of both.

Therefore, a low stringency threshold was to be chosen as evolutionary events between eukaryotes and prokaryotes date back as far as 1.6 billion years ago [Betts *et al*. 2018] and lower stringency during protein family reconstruction facilitates the study of early molecular evolution [Cantarel *et al*. 2006; Landan and Graur 2009].

A problem arises at low stringencies, however. In a sufficiently large database, such as the one used in this thesis, even unrelated sequences can produce global alignments with more than 20% amino acid identity [Jaroszewski *et al*. 2002]. Because there are 20 amino acids, one might think that random sequence similarity would be in the range of 5% sequence identity, but this is only the case if all amino acids are equally frequent in proteins, which is not the case [Athey *et al*. 2017; Dayhoff and Eck 1968; Dessimoz *et al*. 2006]. Due to the unequal distribution of the amino acids, unrelated sequences can have higher amino acid identity — especially in large sequence databases. This is why the range from 20% to 30% sequence identity is generally named the twilight zone [Doolittle 1986; Rost 1999; Jeffroy *et al*. 2006]. Studies of sequence alignments with less than 25% amino acid identity show that only a fraction exhibits true sequence homology [Rost 1999].

In light of this, it is not surprising that previous studies have shown that a pairwise sequence identity below 25% generates very inaccurate alignments and therefore phylogenetic trees reconstructed from these alignments yield unreliable topologies [Landan and Graur 2009; Rost 1999], whereby unreliable means that alignments and trees can contain random or misleading information. Previous work has also shown that the protein families with 25% sequence identity tend to recover the roughly 30 protein families that are generally recognized as being universal to all genomes [Ku *et al*. 2015]. This is symptomatic for a general problem encountered when studying early evolution. Lower stringency detects deeper similarities, but the clusters become very large and some proteins that are related at the level of three-dimensional structures have effectively no sequence similarity. There are limits to what, and how far back in time, sequence similarity can probe. The present work focused on the level of sequence similarity that is suitable for constructing trees for the purpose of phylogeny which is equal or greater than 25%. This threshold excludes some sequences from analysis but should produce results that are representative for the data as a whole, unless there is some inherent bias such that sequences sharing ≤25% identity evolve in a fundamentally different manner than those with ≥25% sequence identity. There is currently no evidence to suggest such a bias, notwithstanding the well-known poor sequence conservation of membrane spanning domains in proteins [Sojo *et al*. 2016].

Another parameter used for the identification of reliable sequence alignments is the expectation (E) value. It displays a means to identify significant hits in a dataset of nucleotide or protein sequences. The lower the E value, the less chance of incidental or random alignment hits. Calculation of the E value is dependent on the length of the sequences that are compared and both the size and the amino acid frequency distribution of the database that is used to obtain the pairwise sequence identities. This parameter needs to be prespecified before starting the pairwise alignment program. In clustering practice, an E value threshold of $\leq 10^{-10}$ is typically used for the reconstruction of protein families paired with the 25% amino acid sequence identity threshold, though the 25% criterion is more stringent.

1.2.3 Clustering of large protein sequence datasets

At the onset of this investigation in 2017, few groups in the world were using clustering methods to generate trees for all genes for significant genome samples. The majority employed only a few organisms for their studies [Cotton and McInerny 2010; Esser *et al*. 2004; Pisani *et al*. 2007; Rochette *et al*. 2014; Thiergart *et al*. 2012]. With the conclusion of this work, this approach is now more widespread in the evolutionary genomics community. Examples include the estimation of the roots of the ‘tree of life’ [Weiss *et al*., 2016] and the roots within its archaeal [Williams *et al*. 2017] and bacterial [Coleman *et al*. 2020] subtrees. With the number of fully sequenced genomes growing, so does the need for methods to extract evolutionary information from the increasing data using phylogenomics, as well as methods for clustering in order to study the relationship between genes among all sequenced genomes as the trees of all data might present a picture that differs from the tree produced by the ca. 30 genes that are universally distributed across genomes. The present work is intended to deliver a contribution towards that goal.

As mentioned in section 1.2.1, protein sequence clustering is necessary to identify homologous genes from a number of organisms, so that the phylogenetic relationship between the genes and the organisms can be analyzed. The approach employed for the present work was developed for generating all possible gene trees for all curated prokaryotic organisms from the Reference Sequence database (RefSeq) [O'Leary *et al.* 2016] available at the beginning of this thesis as well as a large sample of curated eukaryotic genomes displaying a wide range of representatives from the six eukaryotic supergroups (GenBank [Benson *et al.* 2014], JGI [Nordberg *et al.* 2014], Ensembl Protists [Kersey *et al.* 2018], NCBI [O'Leary *et al.* 2016]). This selection was comprised of 5,655 prokaryotic organisms — 5,443 bacterial and 212 archaeal strains — with a total number of 19,050,992 protein sequences and 150 eukaryotic genomes with 3,420,731 protein sequences. This data was used to generate both eukaryote and prokaryote specific protein families and combined clusters containing both eukaryotic and prokaryotic genes. These clusters were then used for phylogenomic analyses to study the evolutionary transition from prokaryotic to eukaryotic cells. There are different methods for protein family reconstruction that can be applied, and different applications for the resulting gene clusters — for example functional characterization, phylogenetic inference, and structural annotation [Altenhoff and Dessimoz 2009; Roth *et al.* 2008]. Handling the extensive amount of sequence information applied in this work required, during the course of this investigation, the optimization of the applied algorithms as well as improving running time and memory requirements for each step during the reconstruction of protein clusters.

As highlighted in section 1.2.1, the detection of orthologous sequences in a dataset of this extent is best performed by calculating all-vs-all BLAST searches and subsequent filtering for rBBH according to the stringency thresholds discussed in section 1.2.2. Although the BLAST algorithm uses an approximation method to expedite the calculation of local sequence alignments [Altschul *et al.* 1990], it is generally sufficient for detecting homologous sequences. For the purpose of clustering however, the local pairwise alignments produced by BLAST are less reliable than generating global alignments using the Needleman-Wunsch algorithm [Needleman and Wunsch 1970]. Therefore, the Needleman-Wunsch algorithm was applied to the rBBH of the initial BLAST search in order to calculate global amino acid sequence pairs accordingly. The calculation of global alignments could not be performed with all 22 million (19,050,992 prokaryotic and 3,420,731 eukaryotic) protein sequences in the present study, as the Needleman-Wunsch algorithm would take considerably more time and computational resources. The resulting list of sequence pairs was thus filtered for global sequence identity of ≥25% and subsequently applied to the Markov Clustering algorithm (MCL) to generate the protein families [Enright *et al*. 2002; van Dongen 2000], as has been done in previous studies involving the reconstruction of protein families on a larger scale [Ku *et al*. 2015; Ku and Martin 2016; Nelson-Sathi *et al*. 2012; Weiss *et al*. 2016].

While the protein clusters generated by MCL are useful for phylogenomic analysis, they are imperfect. One of the sensitive parameters in clustering is the generation of the all-by-all matrix of pairwise alignments in the first BLAST step. With the present data, this involved computing 5×10^{14} single gene comparisons for the 19,050,992 prokaryotic and 3,420,731 eukaryotic protein sequences, a serious computational challenge. Recently, a new sequence alignment program was developed — DIAMOND [Buchfink *et al*. 2014] — which is significantly faster than BLAST. Initially the algorithm was designed for the processing of short raw reads from next generation sequence data, comparing DNA sequences to protein databases such as employed by BLASTX. Although the DIAMOND algorithm improves upon BLAST in terms of speed, at the time the pairwise sequence identity matrix at the foundation of the present work was generated there was no functioning implementation to compare protein sequences to protein databases, hence BLASTP was used. However, the main computational limitation in clustering is not the generation of the pairwise identity matrix, it is the size of the resulting matrix that has to be filtered and ultimately read by the MCL algorithm.

The present work was substantially enabled by the central computing facilities at the University of Düsseldorf, the Zentrum für Informations- und Medientechnologie (ZIM). The ZIM provided a computing environment that allowed the initiation of algorithms requiring several terabytes of random-access memory (RAM) at a time on a single high performance computing cluster, an environment specifically designed for the kind of work performed in this thesis. This large amount of RAM greatly facilitated this work by enabling the calculation of large matrix analyses for and in the clustering procedure. Specifically, the rBBH step mentioned above involves the filtering of reciprocal best BLAST hits with the effect of reducing the frequency of paralogs among clusters [Wolf and Koonin 2012]. However, this simple procedure becomes complicated if the number of BLAST hits generates a file size that exceeds the available RAM because for the filtering of the best reciprocal sequence pairs all best BLAST hits need to be retained in internal storage. Without a very large RAM and restructuring the previously employed filtering algorithms, this step becomes a bottleneck in the phylogenomic and clustering pipeline. This is one reason why comparatively few bioinformatic groups cluster large data sets. The local computing environment was important for the success of the present work.

After filtering for significant sequence pairs, the next step in phylogenomic inference involves the sorting of sequences from the different genomes into collections of homologous genes — clustering of genes into protein families. These families will then be applied for the multiple sequence alignment in order to construct phylogenetic gene trees for evolutionary analyses. The clustering of sequences into protein families can be performed with different algorithms that employ graph-based (MCL [Enright *et al*. 2002; van Dongen 2000], ProClust [Pipenbacher *et al*. 2002]), sequence-based (CD-HIT [Fu *et al*. 2012], UCLUST [Edgar 2010], InParanoid [Remm *et al*. 2001], LinClust [Steinegger and Söding 2018]) or tree-based algorithms (TreeCluster [Balaban *et al*. 2019]). These algorithms often use prefiltering algorithms — especially those that are mostly automated. This is necessary to reduce the computing time of cluster reconstruction but reduces the accuracy of the resulting protein clusters. However, the pipeline employed in this work was designed to retain as much phylogenetic information as possible for the generation of the protein families. The MCL algorithm creates a network of the applied sequence-IDs (nodes) that are connected by edges that are weighted according to a predetermined value — in this case the global pairwise sequence identity. Flow simulation modeling is applied in order to remove weak connections and reinforce strong connections between the sequences resulting in clusters of sequences that represent the protein families. This is another step that involves significant computational resources.

In order to calculate phylogenetic trees using the maximum likelihood method, the sequences need to be aligned, that is, the homologous positions in the protein coding genes need to be positioned in the same column of a matrix. There are a number of programs available for multiple alignment of protein sequences, none are optimal [Landan and Graur 2007]. In the present work, the MAFFT algorithm was used [Katoh 2002], which takes information from the initial pairwise sequence alignments into account.

1.2.4 Applications for protein clusters

1.2.4.1 Available databases of protein clusters

There are various openly accessible databases of protein clusters for phylogenetic purposes assembled on the internet. Some of the most frequently used cluster databases for the functional annotation of genes and for comparative genomics are the variations of the clusters of orthologous genes (COGs) from the NCBI [Tatusov *et al*. 1997]. As mentioned in section 1.2.1, COGs are orthologous groups generated by graph-based orthology prediction methods applying multi-species gene comparisons. The linking of sequences of best BLAST searches into triangular sequence pairs employed by the algorithm is very stringent and limits the detection of many protein families. However, this generates highly connected clusters for the reliable functional annotation of sequences. Currently, the database includes 26 functional categories separated into 4,877 COGs that include 3,213,025 proteins from 1,309 prokaryotic organisms (1,187 bacteria and 122 archaea) [Galperin *et al*. 2021]. Variations of the COG database were made in 2003 by introducing eukaryotic genomes to form eukaryotic orthologous groups (KOGs) [Tatusov *et al*. 2003] and the distinction of archaeal COGs (arCOGs) in 2014 [Makarova *et al*. 2015] simultaneously including a new level of classification — superclusters that connect two or more arCOGs to better portray gene family evolution.

The database eggNOG [Jensen *et al*. 2008] presently contains 4,445 bacterial and 168 archaeal genomes, 477 eukaryotic organisms and 2,502 viral proteomes [Huerta-Cepas *et al*. 2019]. The 4.4 million orthologous groups contained in the database are reconstructed by a similar approach employed for the computation of COGs/KOGs and can be applied to predict functional annotation as well as protein domains. The eggNOG database extends the number of orthologous groups from COG/KOG including a more varied number of organisms to enhance phylogenetic resolution.

There are various other graph-based databases of orthologous groups. Some of the more frequently employed are OrthoMCL [Li *et al*. 2003], OrthoDB [Kriventseva *et al*. 2019], InParanoid [Remm *et al*. 2001], and OMA [Dessimoz *et al*. 2005]. The Pfam database [Mistry *et al*. 2021] employs Hidden Markov models to reconstruct a profile of the seed sequence that is used for a search against the integrated sequence database of protein families. This procedure is applied to sort sequences into the protein families included in the database. A similar approach is applied by Prosite, which is comprised of a database of profiles and patterns that are designed to detect specific protein families [Sigrist *et al*. 2002]. A different approach to infer orthology of sequences is to employ phylogenetic trees as a guide. Algorithms that apply guide trees to reconstruct protein families are for example TreeFam [Schreiber *et al*. 2014], Ensembl Compara [Vilella *et al*. 2009], Panther [Mi *et al*. 2013], PhylomeDB [Huerta-Cepas *et al*. 2014], PhyloFacts [Datta *et al*. 2009] and PhIGs [Dehal and Boore 2006].

The main differences between the aforementioned databases are the restriction to specific taxonomic groups (for example: vertebrates — Ensembl, animals — TreeFam, fungi and metazoans — PhIGs, eukaryotes — OrthoMCL) or limiting the applied data to fully sequenced and/or manually curated genomes (COG/KOG, TreeFam). Manual curation of the reconstructed protein families (Pfam, COG/KOG), the use of additional information such as existing phylogenetic trees (PhyloFacts) or protein domain information (Pfam, Prosite) and functional sites (Prosite) are additional distinctions between these databases. Furthermore, the tree-based algorithms can be distinguished by the kind of phylogenetic trees employed during the process of assigning proteins into families; automatic tree reconstruction (Panther, PhylomeDB) or manual curation of gene trees (TreeFam, Ensembl). The databases are mostly used for functional annotation of genes, but some also contain expression information and can be applied for phylogenetic inference or the characterization of biochemical properties. Due to the possibility to infer gene losses or duplications in orthologous groups reconstructed by tree-based methods, they can be more informative than groups reconstructed by graph-based pipelines [Schreiber *et al*. 2014]. However, graph-based algorithms for protein family reconstruction are generally computationally less demanding and can be more readily applied to very large datasets.

As detailed, the aforementioned databases have in common that they encompass only a small fraction of the available sequenced genomes or are very stringent during the assignment of sequences into protein families, thus creating only very few protein families. Therefore, they only represent a very limited overview of the whole data. In general, the better curated and the more reliable a given sequence database is, the smaller the number of genomes it encompasses and the more slowly it can be updated. Increasing the number of genomes in the databases is difficult due to simple resource demand. Focusing on only a subset of organisms can be necessary for certain analyses but studying early phylogenetic events between prokaryotes and eukaryotes should encompass as much available data as possible. As of March 2021, the number of complete sequenced genomes in RefSeq increased to 1,120 archaea and 210,095 bacteria. Adding eukaryotes and virus genomes, the number increases to 223,187 genomes in total. This is a more than five-fold increase in archaeal genomes and almost 40-fold increase in bacterial genomes compared to the 2016 dataset employed in this work. This extensive biological diversity of eukaryotes and prokaryotes might give additional insights into their deeper phylogenies. Therefore, existing algorithms need to be adapted in order to process the fast-progressing amount of genomic data, if one wants to reconstruct all gene trees from the available data. The methods developed during the course of this work represent a step towards this goal.

1.2.4.2 Eukaryote-prokaryote clusters

As stated previously, there are various resources available for the functional annotation of proteins, structural prediction, and phylogenetic inference that are based on protein clusters. These clusters have their function and warranty. However, protein families available in the public databases have in common that they are specific for a certain group of organisms that span *either* eukaryotic *or* prokaryotic species and often only a subset of either. When studying the evolutionary history of how eukaryotes arose from bacteria and archaea, it is mandatory to use protein families that include homologs from all three domains of life [Brueckner and Martin 2020; Ku *et al*. 2015; Ku and Martin 2016; Nagies *et al*. 2020]. As of the time of writing, no other reports of all phylogenetic trees for protein coding genes that can be detected in a large sample of both eukaryotes and prokaryotes, such as presented in this work, could be identified in the literature databases. However, phylogenetic analyses of the early evolution of eukaryotes are dependent on eukaryote-prokaryote protein families as they represent the most information-rich resource to gain insights into eukaryote origin.

Phylogenetic trees generated to study the transition from prokaryotes to eukaryotes generally focus on a small sample of eukaryotes [Pisani *et al*. 2007; Rochette *et al*. 2014; Thiergart *et al*. 2012] or only one eukaryotic genome [Cotton and McInerny 2010; Esser *et al*. 2004], because assigning prokaryotic sequences to their eukaryotic homolog is extremely complicated with a big dataset of organisms. Proteins of eukaryotes are generally much longer than their prokaryotic counterparts [Brocchieri and Karlin 2005; Liang and Riley 2001; Zhang 2000], which might be in part due to their different gene structure influenced by gene fusion events [Brocchieri and Karlin 2005], ecological habitats [Tekaia *et al*. 2002] as well as enhanced energy availability due to the increased ATP production in eukaryotes by mitochondria [Martin 2017]. Therefore, global alignments between eukaryotic and prokaryotic organisms result in very unspecific sequence alignments and the approach to generate protein families outlined previously could not be maintained. Stated another way, eukaryotic proteins often combine protein domains in ways not observed in prokaryotes, this tends to connect eukaryotic protein families to prokaryotic sequences that, in the absence of eukaryotic homologs, would tend to fall into separate, unrelated clusters. This impairs the clustering procedure and compromises the quality of the resulting clusters.

The eukaryote-prokaryote clustering method employed in this work was developed by Ku *et al*. [2015] and was adapted here to manage the larger dataset. The procedure is simple. In a first step, the sequences from genomes of eukaryotes, archaea, and bacteria are clustered separately to reconstruct protein clusters originating from only one domain each following the approach outlined in section 1.2.4 with stringency thresholds dependent on the evolutionary question analyzed. In a second step, the protein families for each domain were combined with their homologs from the other domains according to local pairwise sequence identities between the eukaryotic and prokaryotic protein sequences. Each prokaryotic domain was first combined with its eukaryotic homolog, to generate either eukaryote-archaea or eukaryote-bacteria clusters. Then the eukaryote-archaea or eukaryote-bacteria cluster pairs were combined into protein families of all three domains, if both had the same corresponding eukaryote homolog. In this way, eukaryotic and prokaryotic sequences could be assigned to a shared protein family and later annotated, although the sequence divergence between eukaryotes and prokaryotes is often relatively high. This procedure delivers clusters in which the eukaryotic sequences are grouped with their best homologs among prokaryotes, and in which all clusters have unique membership, that is, no sequence occurs in more than one cluster.

1.3 Applying protein clusters to evolutionary questions

The clusters in the present work were generated to address specific questions in genome history. The goal was to distill insights into early evolution by querying all genes, rather than querying one, a few or 30 genes as a substitute for the insights that the whole genome can deliver. In the following, the three main questions in the foreground of the present work are introduced.

As the first of these investigations, it was the goal to analyze the evolutionary origin of eukaryotic protein coding genes, by assigning the proteins of eukaryotic genomes to archaeal or bacterial origins. For that, the proportion of eukaryotic clusters among 150 genomes that trace exclusively to bacteria or to archaea was determined in order to provide genome-wide estimates for the evolutionary origin of eukaryotic genes that are shared with prokaryotes. Although it is known that eukaryotes invented many novel proteins that are not present in prokaryotic organisms [Aravind *et al*. 2006], the basic starting material from which the origin of eukaryote specific genes arose was, in the simplest hypothesis, contributed by an archaeal host and a bacterial symbiont [Martin and Müller 1998; Imachi *et al*. 2020]. The relative contribution of prokaryotic genes to the eukaryotic partners was estimated by presence-absence criteria within the clusters using a down-sampling procedure to correct for the unequal bacterial and archaeal sample sizes in the present data. The result of those investigations is summarized in section 3.1, the manuscript Brueckner and Martin [2020].

For the second analysis, the data in the present work was applied in order to better characterize the effects of LGT during prokaryote and eukaryote genome evolution. For over 50 years, microbiologists have known that prokaryotes transfer genes across the species boundary [Lederberg and Tatum 1946]. The first clear evidence for this emerged in hospitals in the form of bacterial pathogens with multiple antibiotic resistance genes [Aminov 2010; Rollo *et al*. 1952]. It was found that resistance to one antibiotic spread as a function of antibiotic usage, not as a function of the evolutionary relationships of the strains that became resistant and that antibiotic resistance genes were spread on plasmids through a process called conjugation [Davies and Davies 2010; Nakaya *et al*. 1960; Popa and Dagan 2011]. It was later discovered that two other mechanisms of gene transfer exist in natural prokaryote populations: transfer via phage (transduction), and DNA transfer via uptake of environmental DNA (transformation) [Ochman *et al*. 2000; Popa and Dagan 2011]. As genome sequence data became available for a large number of prokaryotes, it became apparent that only very few genes were universal [Charlebois and Doolittle 2004; Hansmann and Martin 2000] and that most genes were distributed across genomes in a highly uneven manner [Dagan and Martin 2007]. This provided clear hints that LGT had contributed to a substantial extent to the overall composition of prokaryotic genomes. One of the most common ways to identify LGT events was a comparison of phylogenetic trees [Ochman *et al*. 2000]. If different genes gave different trees for the same set of species this was widely interpreted as evidence for LGT — at least in the early days of genomics. A problem arises however if one wants to estimate the amount of LGT that has occurred in the evolution of a particular lineage or a particular gene, if large genome samples are available. This is because the number of possible trees grows faster than exponentially with each additional leave (sequence) [Cavalli-Sforza and Edwards 1967; Graur 2016] in the tree. For example: if n = 2 there is one possible rooted tree, but with 10 leaves (n = 10) the number already increases to 34,459,425. As a consequence, the comparison of trees with many leaves will always generate discordant phylogenies but it becomes very difficult to determine whether the differences are observed due to LGT or phylogenetic artifacts.

For the present work, a method was required that would permit the estimation of LGT frequency from trees with thousands of leaves but avoided their most serious problem — namely that the deeper branches in trees with more than 500 leaves are not better than random [Semple and Steel 2009]. The approach that was taken for this work focused solely on the most terminal tips of the trees which is the region where phylogeny works best. For example, the neighbor-joining methods start by inferring the topology at the tips of the tree and works towards the deeper branches of the tree from there [Saitou and Nei 1987], and all modern maximum-likelihood methods for phylogenetic tree reconstruction start from such a neighbor-joining topology [Chor and Tuller 2005]. Rather than asking which branches from two trees are discordant, the work in this thesis asked how many prokaryotic phyla in a given tree are monophyletic. This procedure provided an estimate for the verticality of phyla in a given tree and summed across all trees and all phyla provided estimates for the relative amount of recent LGT that occurred among phyla and across trees. This was designated as verticality (*V*) and comprises the main topic of section 3.2, the manuscript by Nagies *et al*. [2020].

The third question required the extraction of more information from each tree as it concerned the role of gene duplication, primarily in eukaryotic genome evolution. Gene duplications [Ohno 1970] were among the earliest discoveries of molecular phylogenetic investigation because they represent the underlying theme of globin gene evolution. Globins were one of the first protein families for which extensive sequence information existed for many species. The evolution of globins was investigated long before the advent of genome sequencing technologies. Already during the 1960's, many globin sequences were determined using protein sequencing methods, as the globins were easy to isolate from red blood cells (hemoglobins) and muscles (myoglobins) [Goodman *et al*. 1987]. They helped to form the paradigm for eukaryote gene evolution that novelty (novel sequences) arises through gene duplications [Ohno 1970], a paradigm that was further extended during the last 20 years with the recognition that whole genome duplications are very common in eukaryote genome evolution [Crow and Wagner 2006]. In whole genome duplication, all genes in the genome undergo duplications — making the quantification of genome duplications challenging for large datasets. The approach to this problem employed in the present work harnessed the properties of the minimum ancestor deviation method (MAD) [Tria *et al*. 2017] to identify duplication events. The results of that work are summarized in section 3.3, the manuscript by Tria *et al*. [2021].

2 Aim of this thesis

There are several hypotheses regarding the events that led to the development of eukaryotic cells. Most studies focus on a very limited number of organisms to infer early evolutionary events. This approach has its merits in specific cases, but in order to get a more comprehensive picture of the mechanisms regarding the transition of prokaryotic to eukaryotic life all available data should be assessed. The primary goal of the present work was to reconstruct protein families for 5,655 prokaryotic and 150 eukaryotic organisms that contain orthologous genes of both eukaryotes and prokaryotes. These eukaryote-prokaryote clusters and the subsequently generated phylogenetic trees could then be linked to phylogenetic and functional properties in order to elucidate questions regarding early evolutionary events.

To better understand the transition period from simple unicellular life forms to eukaryote complexity, the work in this dissertation aims to illuminate mechanisms regarding the gene acquisition from prokaryotes to the eukaryotic genome. Specific goals of the present thesis and the included publications were:

(a) Investigate the proportion of archaeal and bacterial genes regarding eukaryotic homologs.

(b) Characterize the effects of LGT during prokaryote and eukaryote genome evolution.

(c) Analyze the role of gene duplications in eukaryotic genome evolution.

3 Publications

3.1 Bacterial genes outnumber archaeal genes in eukaryotic genomes

Year:	2020
Authors:	Julia Brueckner, William F. Martin
Published in:	Genome Biol Evol
Contribution:	First author Major: Collection and analysis of the data, study design, figure design and illustration. With the last author: Manuscript writing and editing.

Bacterial Genes Outnumber Archaeal Genes in Eukaryotic Genomes

Julia Brueckner and William F. Martin*

Institute for Molecular Evolution, Heinrich Heine University Düsseldorf, Germany

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

Accepted: March 2, 2020

Data deposition: Accession numbers of the sequence data are supplied in supplementary material on our website: http://www.molevol.de/resources/index.html.

Downloaded from https://academic.oup.com/gbe/article/12/4/282/5788535 by guest on 11 February 2021

Abstract

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

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

Introduction

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

In the traditional three domain tree of life, eukaryotes are seen as a sister group to archaea (Woese et al. 1990; Da Cunha et al. 2017, 2018) (fig. 1*a*). In newer two-domain

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

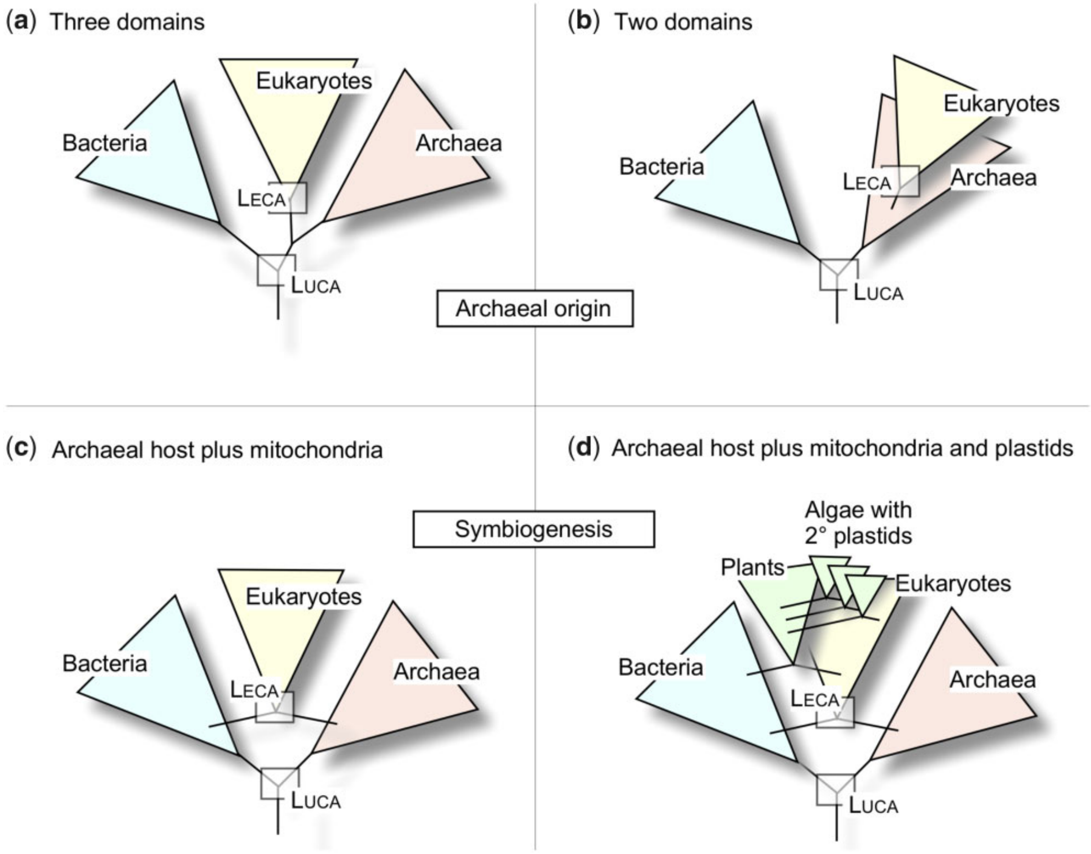


Fig. 1.—Differing views on the relationships of eukaryotes to prokaryotes. (*a*) The three domain tree. (*b*) The two-domain tree with an archaeal origin of eukaryotes. (*c*) Symbiogenesis at the origin of eukaryotes. (*d*) Symbiogenesis at the origin of eukaryotes plus plastids at the origin of the plant kingdom and secondary symbiotic events among algae (see Embley and Martin 2006; Gould et al. 2008; McInerney et al. 2014; Martin 2017).

trees, eukaryotes are viewed as branching from within the archaea (Cox et al. 2008; Williams et al. 2013) (fig. 1*b*). In both the two domain and the three domain hypotheses, this is often seen as evidence for "an archaeal origin" of eukaryotes (Cox et al. 2008; Williams et al. 2013) (fig. 1*a*, *b*). Germane to an archaeal origin is the view that eukaryotes are archaea that became more complex by gradualist evolutionary processes, such as point mutation and gene duplication (Field et al. 2011; Schlacht et al. 2014). Countering that view are two sets of observations relating to symbiogenesis (origin through symbiosis) for eukaryotes (fig. 1*c*, *d*). First, the archaea that branch closest to eukaryotes in the most recent phylogenies are very small in size (0.5 μm), they lack any semblance of eukaryote-like cellular complexity, and they live in obligate association with bacteria (Imachi et al. 2020), clearly implicating symbiosis (Imachi et al. 2020) rather than point mutation as the driving force at the origin of the eukaryotic clade (fig. 1*c*). Second, and with a longer history in the literature, are the findings that mitochondria trace to the LECA (Embley and Hirt 1998; van der Giezen 2009; McInerney et al. 2014) and that many genes in eukaryote genomes trace to gene transfers from endosymbiotic organelles (Martin and Herrmann 1998; Timmis et al. 2004; Ku et al. 2015). A symbiogenic origin of eukaryotes would run counter to one of the key goals of phylogenetics, namely to place eukaryotes in a natural system of phylogenetic classification where all groups are named according to their position in a bifurcating tree. If eukaryotes arose via symbiosis of an archaeon (the host) and a bacterium (the mitochondrion), then eukaryotes would reside simultaneously on both the archaeal and the bacterial branches in phylogenetic schemes (Brunk and Martin 2019; Newman et al. 2019), whereby plants and algae that stem from secondary symbioses (Gould et al. 2008) would reside on recurrently anastomosing branches as in figure 1*d*.

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

Downloaded from https://academic.oup.com/gbe/article/12/4/282/5788535 by guest on 11 February 2021

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

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

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

Materials and Methods

Sequence Clustering

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

Downloaded from https://academic.oup.com/gbe/article/12/4/282/5788535 by guest on 11 February 2021

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

Assignment of Bacterial or Archaeal Origin

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

Cluster Annotation

Protein annotation information according to the BRITE (Biomolecular Reaction pathways for Information Transfer and Expression) hierarchy was downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG v. September 2017) website (Kanehisa et al. 2016), including protein sequences and their assigned function according to the KO numbers (Suppl. Material 8a, b). The sequences of each protein family from the 2,587 EPCs were locally aligned with "blastp" to the KEGG database to identify the annotation for each protein. In order to assign each protein to a KEGG function, only the best BLAST hit of the given protein with an *E* value ≤ 1e−10 and alignment coverage of 80% was selected. After assigning a function based on the KO numbers of KEGG for each protein in the EPCs, the majority rule was applied to identify the function for each cluster. The occurrence of the function of each protein was added and the most prevalent function was assigned for each cluster (supplementary table 4, Supplementary Material online). Poorly characterized sequences or sequences with no assigned function were ignored, resulting in 1,836 clusters with annotations.

Presence and Absence of EPCs across Genomes

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

Results

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

The proportion of bacterial and archaeal genes for each eukaryote is shown in figure 2*b*. Overall, 44% of eukaryotic sequences are archaeal in origin and 56% are bacterial. Across 150 genomes, eukaryotes possess 12% more bacterial genes than archaeal genes. There are evident group specific

Downloaded from https://academic.oup.com/gbe/article/12/4/282/5788535 by guest on 11 February 2021

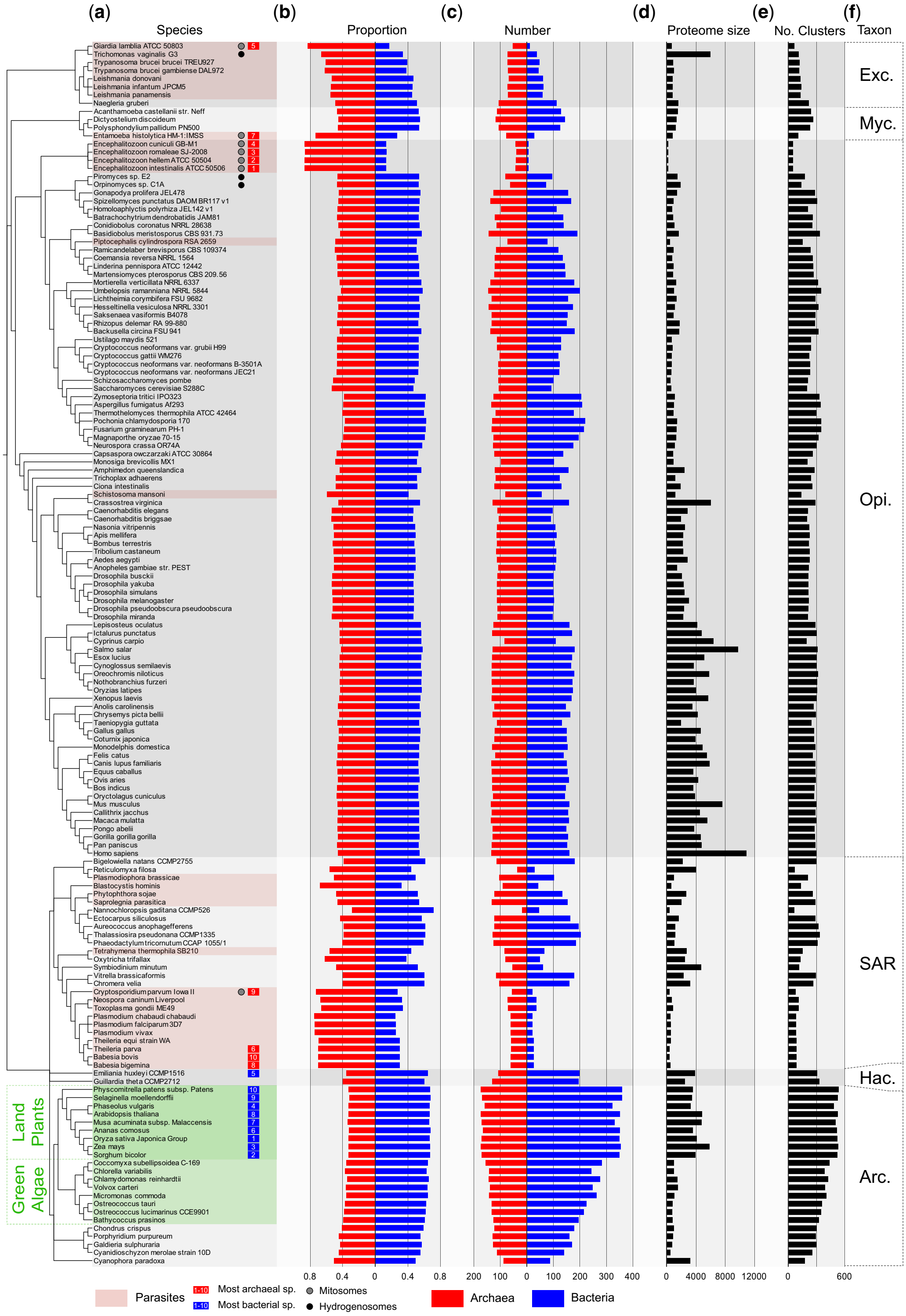


Fig. 2.—Bacterial and archaeal genes in eukaryotic genomes. Protein sequences from 150 eukaryotic genomes and 5,655 prokaryotic genomes (5,433 bacteria and 212 archaea) were clustered into eukaryote–prokaryote clusters (EPC) using the MCL algorithm (Enright et al. 2002) as described (Ku et al. 2015).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

Downloaded from https://academic.oup.com/gbe/article/12/4/282/5788535 by guest on 11 February 2021

Table 1
Proportion of Bacterial and Archaeal Derived Genes in Eukaryotic Genomes

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

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

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

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

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

The proportions for different eukaryotic groups are shown in table 1. Land plants have the highest proportion of bacterial derived genes at 67%, or a 2:1 ratio of bacterial genes relative to archaeal. The eukaryote with the highest proportion of bacterial genes in our sample is rice, with 67.1% bacterial and 32.9% archaeal genes. The higher proportion of bacterial genes in plastid containing eukaryotes relative to other groups corresponds with the origin of the plastid and gene transfers to the nucleus (Ku et al. 2015). The eukaryote with the highest proportion of archaeal genes in our sample are the human parasite *Encephalitozoon intestinalis* and the rabbit parasite *Encephalitozoon cuniculi*, with 86% archaeal and 14% bacterial derived genes. Parasitic eukaryotes have the largest proportions of archaeal genes, but not by novel acquisitions, rather by having lost large numbers of bacterial genes as a result of reductive evolution in adaptation to nutrient rich environments. This is evident in figure 2c, where the numbers of archaeal and bacterial genes per genome are shown. Parasites, with their reduced genomes, such as *Giardia lamblia*, *Trichomonas vaginalis*, or *Encephalitozoon* species, appear more archaeal. The number of archaeal, or bacterial genes in an organism does not correlate with genome size (supplementary fig. 1, Supplementary Material online, Pearson correlation coefficient: archaeal $r^2 = 0.38$, bacterial $r^2 = 0.33$).

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

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

Fig. 2.—Continued
eukaryotic sequences in the eukaryote–prokaryote clusters. (*f*) Taxonomic groups are labeled on the far right panel (Arc.—Archaeplastida, Exc.—Excavata, Hac.—Hacrobia, Myc.—Mycetozoa, Opi.—Opisthokonts). Highlighted in green is the branch with the taxa of plants and green algae, parasites are highlighted in red. The black dots indicate organisms with hydrogenosomes, the gray dot indicates organisms with mitosomes.

Downloaded from https://academic.oup.com/gbe/article/12/4/282/5788535 by guest on 11 February 2021

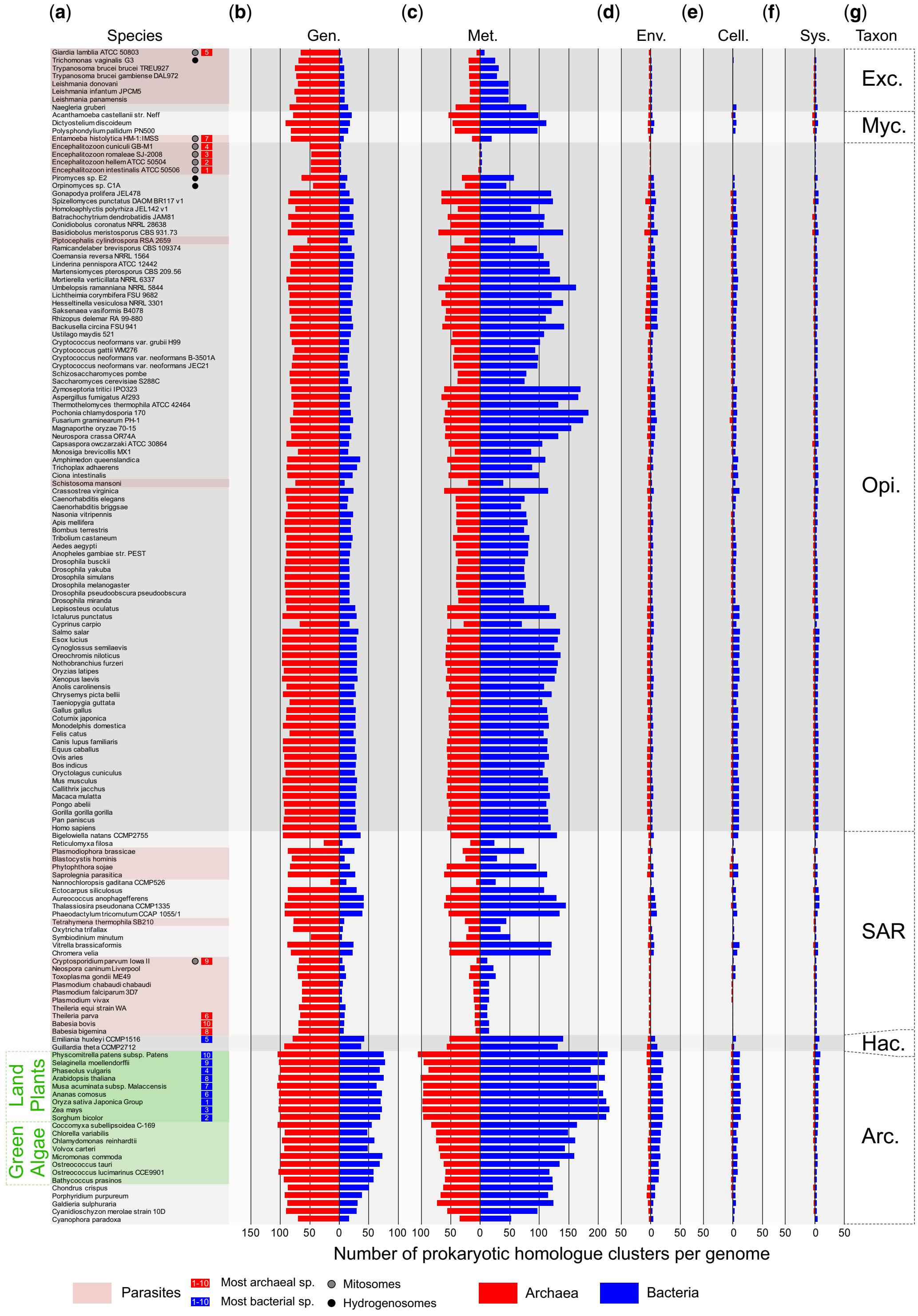


FIG. 3.—Functional categories. Protein sequences from 150 eukaryotic genomes and 5,655 prokaryotic genomes were clustered into 2,587 eukaryote–prokaryote clusters (EPC) (Ku et al. 2015). Sorted according to a reference tree for eukaryotic lineages generated from the literature and taxonomic groups are labeled. The red bars indicate eukaryotic gene families that are archaeal in origin, blue indicates a bacterial origin of the gene family. Functional annotations according to the KEGG BRITE hierarchy on the level A was assigned for each EPC, identifying the function for each sequence in the protein cluster by performing a protein BLAST against the KEGG database and then applying the most prevalent function per protein family. Only the categories (*b*) "Genetic Information Processing" (Gen.), (*c*) "Metabolism" (Met.), (*d*) "Environmental Information Processing" (Env.), (*e*) "Cellular Processes" (Cell.), and (*f*) "Organismal Systems" (Sys.) are depicted, as the label "Human Diseases" was hardly represented. Species names are indicated in column (*a*) and taxonomic groups (*f*) are labeled on the far right panel (Arc.—Archaeplastida, Exc.—Excavata, Hac.—Hacrobia, Myc.—Mycetozoa, Opi.—Opisthokonts). Highlighted in green is the branch uniting land plants and green algae; the black and gray dots indicate organisms with hydrogenosomes or mitosomes, respectively.

Downloaded from https://academic.oup.com/gbe/article/12/4/282/5788535 by guest on 11 February 2021

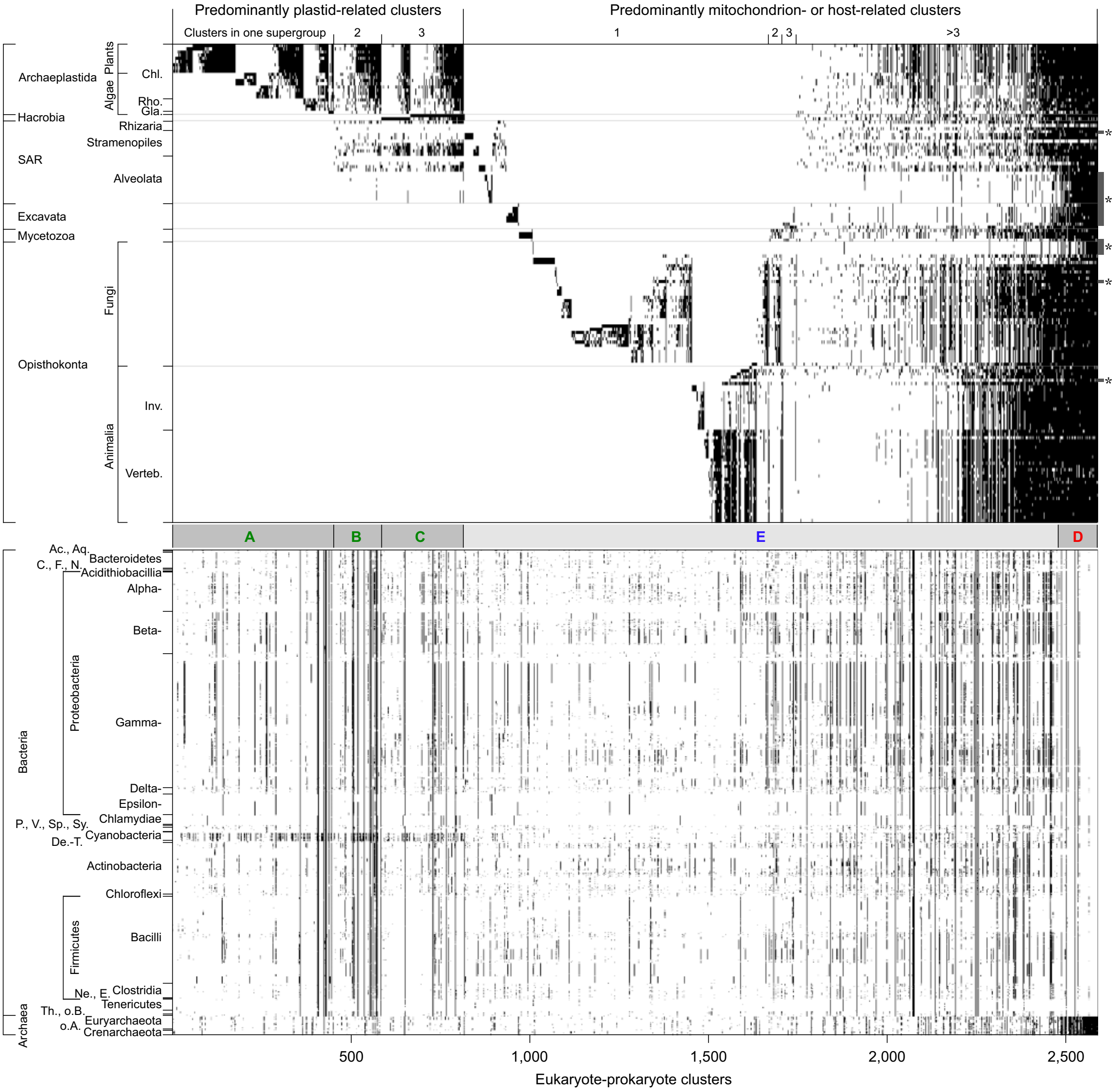


Fig. 4.—Gene sharing matrix. Each black tick represents the presence of a gene in the respective taxon. First, the 2,587 EPCs (*x* axis) were sorted according to their distribution across the six eukaryotic supergroups with the photosynthetic lineages on the left (block *A*–*C*). Host- or mitochondrion-related genes distributed across the six supergroups are depicted in block *E*. Clusters with mostly archaeal homologs are indicated in block *D* (Chl.—Chloroplastida, Rho.—Rhodophyta, Gla.—Glaucophyta, Inv.—Invertebrates, Verteb.—Vertebrates; Ac.—Acidobacteria, Aq.—Aquificiae, C.—Chlorobi, F.—Fusobacteria, N.—Nitrospirae, P.—Planctomycetes, V.—Verrucomicrobia, Sp.—Spirochaetes, Sy.—Synergistetes, De.-T.—Deinococcus-Thermus, Ne.—Negativicutes, E.—Erysipelotrichia, Th.—Thermotogae, o. B.—other Bacteria, o. A.—other Archaea).

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

The distribution of the genes in the 2,587 EPCs across genomes for six supergroups is depicted in figure 4. The order of eukaryotic and prokaryotic organisms (rows) can be found in supplementary table 5, Supplementary Material online. Block A represents only Archaeplastida, block B depicts genes found in Archaeplastida and SAR, block C encompasses all genes that are distributed across the three taxa that contain plastids; Archaeplastida, SAR, and Hacrobia. The lower part of the figure shows the prokaryotic homologous genes. Cyanobacterial genes are especially densely distributed across blocks A–C. Genes that are predominantly mitochondrion- or host-related are indicated in blocks D and E. Eukaryotic genes that are universally distributed across the six supergroups are mainly archaeal in origin (block D). Especially organisms with reduced genomes, such as parasites (marked with asterisks on the right), have lost genes associated with metabolism,

Downloaded from https://academic.oup.com/gbe/article/12/4/282/5788535 by guest on 11 February 2021

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

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

Discussion

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

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

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

As improved culturing and sequencing of complete archaeal genomes progresses, new lineages are being characterized at the level of scanning electron microscopy that

Downloaded from https://academic.oup.com/gbe/article/12/4/282/5788535 by guest on 11 February 2021

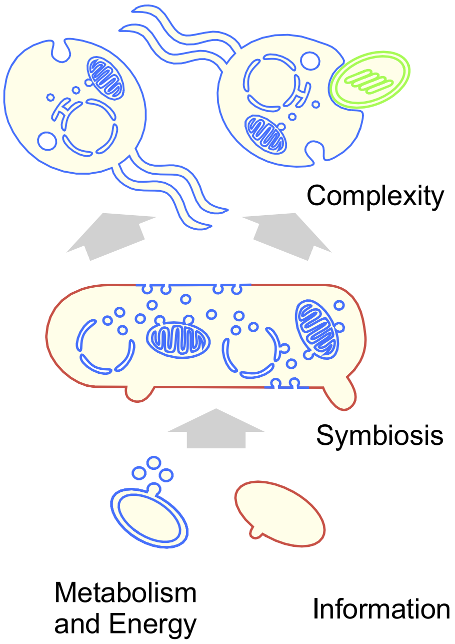


Fig. 5.—Bacterial and archaeal contributions to eukaryotes. Schematic representation of eukaryote origin involving an archaeal host and a mitochondrial symbiont that transforms the host via gene transfer from the endosymbiont (Martin and Müller 1998; Imachi et al. 2020). The model combines elements of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

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

Supplementary Material

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

Acknowledgments

This study was supported by the European Research Council (grant 666053); and the Volkswagen Foundation (grant 93 046). We thank Nils Kapust, Michael Knopp, Sriram Garg, Josip Skejo, Verena Zimorski, and Sven Gould for helpful discussions.

Author Contributions

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

Literature Cited

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Downloaded from https://academic.oup.com/gbe/article/12/4/282/5788535 by guest on 11 February 2021

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Associate editor: Davide Pisani

Downloaded from https://academic.oup.com/gbe/article/12/4/282/5788535 by guest on 11 February 2021

3.2 A spectrum of verticality across genes

Year: 2020

Authors: Falk S. P. Nagies*, Julia Brueckner*, Fernando D. K. Tria, William F. Martin

Published in: PLoS Genetics

Contribution: Shared first author

Major: Collection and analysis of data, study design, figure design and illustration. With the other first author and last author: Manuscript writing and editing.

*These authors contributed equally to this work

RESEARCH ARTICLE

A spectrum of verticality across genes

Falk S. P. Nagies*☯, Julia Brueckner☯, Fernando D. K. Tria, William F. Martin

Institute for Molecular Evolution, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

☯ These authors contributed equally to this work.
* Falk.Nagies@hhu.de

Check for updates

OPEN ACCESS

Citation: Nagies FSP, Brueckner J, Tria FDK, Martin WF (2020) A spectrum of verticality across genes. PLoS Genet 16(11): e1009200. https://doi.org/10.1371/journal.pgen.1009200

Editor: Takashi Gojobori, National Institute of Genetics, JAPAN

Received: July 31, 2020

Accepted: October 16, 2020

Published: November 2, 2020

Copyright: © 2020 Nagies et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Files for S9 Table are available on the repository of our university: http://dx.doi.org/10.25838/d5p-12. The other supplementary files are provided with the manuscript and Supporting Information.

Funding: This study was supported by the European Research Council (666053), the Volkswagen Foundation (93 046), and the Moore-Simons Project on the Origin of the Eukaryotic Cell (9743) which were awarded to WFM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Lateral gene transfer (LGT) has impacted prokaryotic genome evolution, yet the extent to which LGT compromises vertical evolution across individual genes and individual phyla is unknown, as are the factors that govern LGT frequency across genes. Estimating LGT frequency from tree comparisons is problematic when thousands of genomes are compared, because LGT becomes difficult to distinguish from phylogenetic artefacts. Here we report quantitative estimates for verticality across all genes and genomes, leveraging a well-known property of phylogenetic inference: phylogeny works best at the tips of trees. From terminal (tip) phylum level relationships, we calculate the verticality for 19,050,992 genes from 101,422 clusters in 5,655 prokaryotic genomes and rank them by their verticality. Among functional classes, translation, followed by nucleotide and cofactor biosynthesis, and DNA replication and repair are the most vertical. The most vertically evolving lineages are those rich in ecological specialists such as Acidithiobacilli, Chlamydiae, Chlorobi and Methanococcales. Lineages most affected by LGT are the α-, β-, γ-, and δ- classes of Proteobacteria and the Firmicutes. The 2,587 eukaryotic clusters in our sample having prokaryotic homologues fail to reject eukaryotic monophyly using the likelihood ratio test. The low verticality of α-proteobacterial and cyanobacterial genomes requires only three partners—an archaeal host, a mitochondrial symbiont, and a plastid ancestor—each with mosaic chromosomes, to directly account for the prokaryotic origin of eukaryotic genes. In terms of phylogeny, the 100 most vertically evolving prokaryotic genes are neither representative nor predictive for the remaining 97% of an average genome. In search of factors that govern LGT frequency, we find a simple but natural principle: Verticality correlates strongly with gene distribution density, LGT being least likely for intruding genes that must replace a preexisting homologue in recipient chromosomes. LGT is most likely for novel genetic material, intruding genes that encounter no competing copy.

Author summary

Because multicellular life is a latecomer in Earth history, most of evolutionary history is microbial evolution. Scientists investigate microbial evolution by studying the evolution of genes. One of the main surprises of the genomic era is the amount of lateral gene transfer that has gone on in prokaryote genome evolution. Gene transfer clouds evolutionary history, but by how much: How lateral and how vertical is the microbial evolutionary

Competing interests: The authors have declared that no competing interests exist.

process across genes, genomes and lineages? We introduce measures of verticality in genome evolution that permit a ranking of genes and lineages according to their degree of verticality. We show that genes already present in genomes are less likely to be replaced by a newly introduced copy than genes that offer new evolutionary opportunities for the recipient, providing a simple and natural mechanism that limits and promotes lateral gene transfer frequency. Only a very small minority of prokaryotic genes evolve vertically. While the 100 genes that are most widely used to describe the phylogenetic relationships of microbes are indeed the most vertical, they are not at all representative for the evolution of other genes. These findings have broad implications for how we understand the evolutionary process as inferred from gene trees.

Introduction

Prokaryotes undergo recombination that is facilitated by the mechanisms of lateral gene transfer (LGT) [1,2]—transformation, conjugation, transduction, and gene transfer agents [3]. These mechanisms introduce DNA into the cell for recombination and do not obey taxonomic boundaries, species or otherwise. Over time they generate pangenomes [4,5] that are superimposed upon vertical evolution of a conserved core. About 30 genes are present in all genomes [6–9], a few more are nearly universal [10], many are found only in strains of one species [5], but the vast majority of genes are distributed between those extremes according to a power law [11]. Previous work has shown that LGT is subject to natural barriers [12,13], that LGT affects core metabolism less than it affects peripheral metabolism [14] and that LGT is affected by regulatory interaction networks [15]. LGT generates collections of genes in each genome that are of different evolutionary age [16], transferred genes are non-randomly associated [17,18], and major events of gene flux have occurred during evolution [9,19]. In principle, each gene should be transferable, because the mechanisms that introduce DNA into the cell are not selective with regard to the nature of sequences introduced, notwithstanding the CRISPR activity associated with phage defense [20]. If all genes are transferrable, what determines verticality?

At the level of strains or species, gene distributions within rapidly evolving pangenomes have been well-studied [21–25]. Less well understood are the factors that govern the distribution of genes across prokaryotic genomes at higher taxonomic levels. These reflect processes that occurred in deep evolutionary time and, in some cases, underpin the physiological identity of major prokaryotic clades. Though LGT impacts prokaryotic evolution, it does not obscure lineage identity, because despite the abundance of LGT, biologists 100 years ago were able to recognize the identity of many higher level taxa, for example Cyanobacteria and Spirochaetes [26], that we still recognize today. Hence there must exist a spectrum of verticality in prokaryote lineage evolution. It follows that a natural spectrum of verticality across prokaryotic genes should exist as well. Here we rank 101,422 gene families from 5,655 prokaryotic genomes according to conservative estimates of verticality and report how this attribute affects phylogenetic inference in microbial evolution in general and as it impacts inference of eukaryote origin in particular.

Results

The verticality of genes

The two main parameters influencing reconstruction of gene evolution across prokaryotes are sequence conservation and phylogenetic distribution, both of which are easy to estimate from clustering methods based on pairwise sequence comparisons. The degree of congruence among trees for overlapping leaf sets is, by contrast, determined by two unknowns: the accuracy of phylogenetic inference relative to the true gene trees, and the relative amount of LGT that has, or has not, occurred in the evolution of each gene (verticality ***V***). There are many methods of tree comparison, but not for measures of gene verticality. If a gene occurs in many lineages, one invariably observes discordance between the branching pattern generated by the gene and that generated by some standard such as rRNA, yet whether such discordance is due to LGT or to technical issues involving alignment and phylogeny [27] is virtually impossible to determine, because knowledge of the amino acid substitution process underlying sequence divergence in real alignments is irretrievable from real data [28]. That problem is exacerbated in trees having thousands of leaves, where random phylogenetic differences are inevitable. For example, there are $3 \cdot 10^{80}$ possible topologies for a tree with 52 leaves, and there are about 10^{80} protons in the universe [29]. A comparison of two trees, each with 52 (or 520, or 5,200) leaves for an alignment of 400 amino acid sites, evaluates many branches that are not better than random.

Earlier surveys of lateral gene transfer across 116 prokaryotic genomes using nucleotide frequency comparisons were reported over a decade ago [30]. In the era of computers that can calculate all trees for all genes, we sought a measure of verticality that is based on phylogenetic principles but independent of the problems inherent to topological comparisons of large trees. To obtain such an estimate, we leveraged two simple but robust assumptions. First, we assume that the higher order taxa of prokaryotes (referred to here as phyla) that microbiologists have traditionally recognized based on morphological, physiological and rRNA sequence criteria are real and constitute monophyletic groups. On that premise, the null hypothesis for phylogenetic behavior of a given gene in a given prokaryotic phylum is vertical evolution (phylum monophyly). Our second assumption for estimating verticality is that molecular phylogeny works most reliably at the tips of trees, the terminal branches. This assumption is the basis of Neighbor Joining [31], almost all alignment programs [32], and maximum likelihood methods, which typically start the topology search from an NJ tree [33]. By reading the trees only at the tips, we disregard phyletic patterns in deeper branches, where pairwise sequence similarity fades and the processes underlying sequence differences, alignments, and branching pattern differences become more numerous, more poorly constrained and more prone to inference errors.

To estimate ***V***, we read the information contained in each tree solely with regard to the branching patterns of phyla by posing the following recursive set of questions: 1) For each phylum that exists in our data, do sequences from the phylum occur in the tree? 2) If so, do they form a monophyletic group (a clade) or are they singletons? 3) How many clades do they form in that tree? 4) For each clade for tree *i* and phylum *j*, what is the phylogenetic composition of the sister group? That set of questions is repeated for all phyla in tree *i*, the results are tabulated, and the procedure repeated for the next tree. The resulting data contains information both about the verticality of all genes (how often phyla appeared monophyletic for each gene) and about the verticality of genome evolution in all phyla (how often phyla were monophyletic across all genes in the phylum). In a world without LGT and perfect data that reconstructs the true tree from the alignment, all phyla would be monophyletic, all genes from the same phylum would have the same sister phylum and each gene would appear to be inherited vertically. In real data, LGT exists and the data are not perfect, but by looking only at the tips we can estimate verticality without random effects among deeper branches. Note that the true branching order of phyla relative to one another has no bearing upon our estimate of ***V***, nor does the relative branching of lower order taxa within each phylum. For a given gene, we calculate ***V*** as follows. For each tree, phyla that are not monophyletic are given a score of zero, the number of genomes present in the tree for each monophyletic phylum is divided by the number of genomes from that phylum among the 5,655 genomes in the data; that proportion is summed across all monophyletic phyla in the tree, that sum is $\boldsymbol{V}$ for that tree or cluster. For n phyla, $\boldsymbol{V}$ obtains a value between 0 and n.

This measure scores the verticality of a gene across all phyla in which it occurs and gives a higher rank to genes that recover phylum monophyly in a tree sampling many phyla than to those with a more narrow distribution, where the opportunity to observe LGT in tree tips is reduced. Note that an accurate taxonomic assignment for each gene is important for estimating $\boldsymbol{V}$, for which reason we do not include metagenomic data, where binning can lead to assemblies of genes from different lineages. Clustering all 19,050,992 genes yielded 448,821 clusters with genes spanning at least two sequenced genomes, with 261,058 clusters spanning at least three genomes for tree reconstruction with an average of 66.4 genomes and 68.7 sequences each. Removing trees that contained sequences from only one phylum left 101,422 trees containing on average 138.8 genomes and 146.7 sequences (median 18 for both).

The first question we asked was whether gene duplications are frequent, as they might emulate LGT and thus mask verticality. For smaller data sets it is known that gene duplications in prokaryotes are generally rare as compared to eukaryotes [34] and that genome sizes constrain the number of duplicates (or transfers) that a genome can accommodate [11]. Estimating ancient duplications for this data set is not possible as duplications and transfers would be indistinguishable, but recent duplications can be quantified. We found 32,277 cases in which the sister of a terminal leaf (gene) occurred within the same prokaryotic genome. For 5,655 prokaryotic genomes this yields 5.7 genome specific duplications per genome. For comparison, 150 eukaryote genomes [35] harbor 109,056 genome specific duplications corresponding to 727 genome specific duplications per genome. Thus, based upon the values for recent duplications in the present sample, gene duplications per genome are 134-fold less frequent in prokaryotes than in eukaryotes. We also plotted the fraction of terminal duplicates normalized for genome size and compared the distribution in eukaryotes versus prokaryotes using all genomes. The cumulative distribution function (**S1A Fig**) shows that a eukaryotic genome has, on average, 4% recent duplications while prokaryotes have 0.2%. This is not an effect of unequal sample size, because the average 20:1 ratio is robust for 100 random samples of 150 prokaryotic genomes (**S1B Fig**). That duplications are 20–134 fold less frequent in prokaryotes than in eukaryotes in this sample of 5,655 genomes corresponds well with the earlier estimate from six groups of closely related bacteria that ~98% of gene families in prokaryotes result from LGT, not duplication [34]. It suggests that in prokaryotic genomes, duplication (paralogy) does not impact estimates of $\boldsymbol{V}$ in prokaryotic genomes to an appreciable extent, a caveat for methods that allow and infer roughly equal probabilities of LGT and duplication, both for prokaryotes and for eukaryotes [36].

The values of $\boldsymbol{V}$ obtained for all genes allows us to rank them by their relative degree of verticality or LGT, as one prefers. What governs LGT? Few factors have been suggested to govern the rate of LGT that genes undergo. It has been suggested that LGT is limited by the number of intermolecular interactions in which a molecule in involved [37]. Although many genes with high values of $\boldsymbol{V}$ encode ribosomal proteins, which have many interactions, many ribosomal proteins have modest values of $\boldsymbol{V}$. We found that the majority of highly vertical genes are soluble proteins as opposed to being components of macromolecular complexes, and that verticality $\boldsymbol{V}$ strongly correlates with the gene's distribution frequency across genomes, as shown in **Fig 1**, where the value of $\boldsymbol{V}$ estimated for each gene is plotted against the number of genomes in which it occurs. **Fig 1A** shows the verticality and distribution of all 101,422 clusters that generate trees. **Fig 1B** displays the verticality the 8,547 clusters that contain more conserved sequences, that is, those that have an average branch length ≤ 0.1 substitutions per site. The

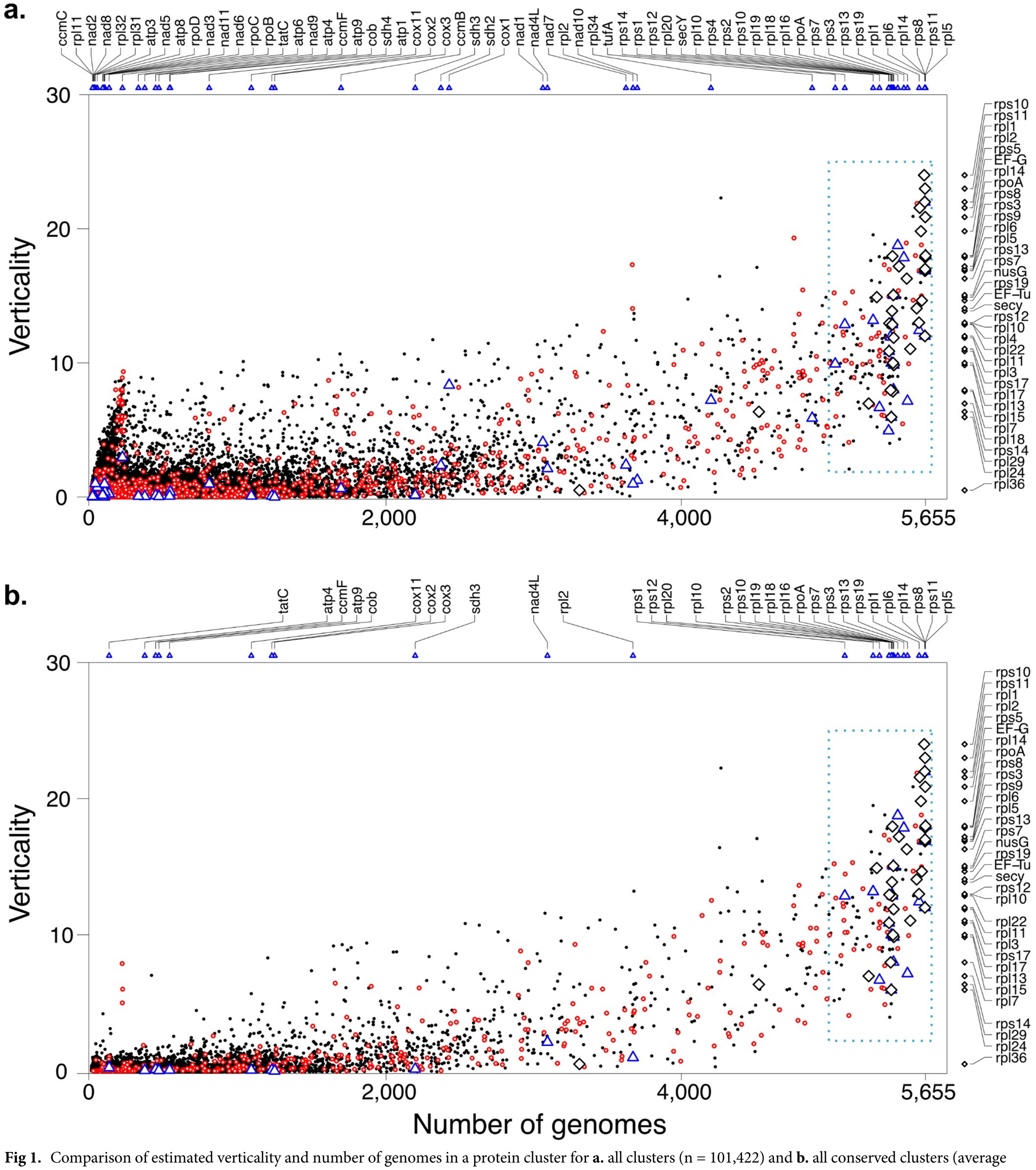


Fig 1. Comparison of estimated verticality and number of genomes in a protein cluster for **a.** all clusters (n = 101,422) and **b.** all conserved clusters (average branch length ≥ 0.1; n = 8,547). Unrooted trees were analyzed if at least two taxonomic groups were present. Verticality was calculated as the sum of monophyletic taxonomic groups in a cluster adjusted by the fraction of a taxonomic group represented in the cluster. The procedure for determining verticality on the basis of an example is shown in S3 Fig. This value correlates with the number of genomes, an approximation of universality, which is even more apparent when clusters of high evolutionary rate were filtered out (a.: $p < 10^{-300}$, Pearson´s $R^2 = 0.726$; b.: $p < 10^{-300}$, $R^2 = 0.829$). In both plots clusters of special interest were marked: The eukaryotic-prokaryotic clusters (EPCs) are highlighted in red and the clusters that correspond to a gene from the mitochondrial genome of *Reclinomonas americana* [45] are displayed in blue triangles along the abscissa of the plot and in the graph. For the latter, the gene

identifier was noted above each plot. Ribosomal proteins are indicated by the black diamond on the right of each plot and in the graph [6]. Notably, the ribosomal protein clusters show a steep gradient of verticality among conserved clusters with similarly wide distribution.

https://doi.org/10.1371/journal.pgen.1009200.g001

spike of sequences at the left of **Fig 1A** represents sequences that tend to be vertically inherited within closely related lineages but whose clusters span only a few genomes because they are not well conserved, for which reason the spike, which encompasses 836 clusters (0.8%; see **S1 Table**), is not present in **Fig 1B**.

The value of V as calculated has desirable properties because it takes distribution into account. In order to see whether verticality is correlated with distribution, we also calculated values of verticality that are independent of distribution, using the number of monophyletic phyla per tree multiplied by the average root-to-tip distance [38] (weighted verticality, V_w; **S10 Table**) instead of dividing by the number of phyla in which the gene is present. The correlation between gene distribution frequency and weighted verticality V_w as inferred independent of distribution frequency was significant at $p < 10^{-300}$ (**S2 Fig**, **S2 Table**). From that one obtains a very general observation about verticality and gene distribution: The most densely distributed genes tend to have the highest verticality, that is, the lowest frequency of recent LGT as determined by phylogenetic criteria.

Why should the most densely distributed genes tend to be most resistant to LGT? We suggest that the reason is simple: If a well-regulated, codon-bias adapted [2] resident copy of a gene already exists in the genome, it would have to be displaced by the intruding copy. Selection in prokaryotes can be intense, as evidenced by codon bias itself: synonymous substitutions that impair codon bias for highly expressed genes are tenaciously counter selected in nature [2]. The existence of a preexisting copy of a gene in the genome reduces the probability of LGT in a highly significant manner ($R^2 = 0.726$; **Fig 1B**). This is all the more noteworthy because the genes that most frequently enter a recipient cell via LGT in nature will be those that are themselves the most widespread genes in nature—that is, the most common genes will be introduced into recipients with the highest frequency. Prokaryotic genes thus have a home field advantage relative to intruders.

The mechanisms of LGT (transduction, transformation, conjugation, gene transfer agents) operate constantly across all prokaryotic genomes in the wild. All things being equal, new coding sequences enter the prokaryotic genome as a random sample of genes available in the environment [39,24], producing natural variation in gene content upon which selection and drift [40] can act to prolong or curtail the gene's lifespan, or residence time, in the descendant clonal lineage. Genes that interfere with the workings of the cell [13] are eliminated quickly from the accessory genome and therefore have a short residence time. Neutral genes that merely constitute functionless ballast can persist in the pangenome longer before loss, while genes that are of value under circumstances encountered by the recipient can become fixed [23,24], in which case they start to shift from the accessory genome to the core genome, thereby defining new genomic lineages of vertical core descent.

The gene families that we observe to be the most vertical (**Fig 1**, **S1 Fig**) are those that are most widely distributed among genomes and hence the most prevalent in nature. This would be puzzling were it not for an inhibitory effect that presence of a preexisting copy exerts on the success rate of LGT. Transposases constitute a special case. They are likely the most common genes in nature [41], but there are different classes of transposases [41], hence they do not fall into one cluster. The fate of transposases is not governed by selection and drift, as they self-amplify within genomes, increasing their copy number by virtue of their ability to do so [42], not by virtue of selection and drift.

The verticality of genes has practical importance for prokaryotic phylogeny, because modern approaches to prokaryotic systematics typically aim to increase the amount of information per lineage beyond that provided by ribosomal RNA. Since 1997, phylogenetic studies of prokaryotic genomes have typically concatenated dozens of sequences into longer alignments [6,43,44]. However, it is not enough to just combine sequences into longer alignments, the sequences ideally need to share the same evolutionary history. ***V*** provides a measure for how vertically a gene tends to evolve over evolutionary time spans. Ranking all genes by their verticality (**Fig 1**; **S1 Table**) provides criteria for inclusion of genes for phylogenetic studies. For orientation, in **Fig 1** we have labelled along the ordinate the genes in current use for phylogenetic studies of archaeal lineages and their relationship to the host that acquired the mitochondrion at eukaryote origin [45]. They differ in their degree of verticality. A number of sequences that are not widely used for phylogeny exhibit higher verticality; these are shown in **Fig 2** and listed in **S6 Table**. Similarly, genes encoded in mitochondrial DNA are typically used to investigate the relationship of mitochondria to bacterial lineages [46]. Those genes are a subset of the genes found in *Reclinomonas americana* mitochondrial DNA [47], which are indicated along the abscissa in **Fig 1**.

From the standpoint of phylogenetics, the main message of **Fig 1** is twofold. First, the genes most commonly used as markers in broad scale prokaryotic phylogenetic studies are, in terms of their distribution and their verticality, not representative for the genome as a whole. Worse, without the comparative information from **Fig 1** they could even be positively misleading, because without measures to compare verticality across genes, one might assume that the tendency of the most widely distributed genes to be vertically inherited is representative for the phylogenetic behavior of all genes. But that is not the case. Widely distributed genes tend to be vertically inherited but they are not a representative sample for the phylogenetic behavior of the genome as a whole. The vast majority of prokaryotic genes are not inherited vertically, hence the small vertically inherited sample is not a good proxy for the phylogenetic behavior of prokaryotic genes. Vertically inherited genes in prokaryotes are not a random sample, they are a biased sample. This is also known as the tree of 1% [9] and is most clearly seen in **Fig 1B**, where the more conservatively evolving, hence phylogenetically more useful genes are shown. The vast majority of genes that occur in two or more phyla in prokaryotes fail to recover phylum monophyly to any appreciable extent, also for estimates of ***V*** that are independent of distribution (**S2 Fig**), and most of them are present in only very few phyla to begin with. The mean and median values of ***V*** in **Fig 1A** are 0.27 and 0.04, in **Fig 1B** 0.70 and 0.06, respectively. The second main message of **Fig 1** concerns the relationship of eukaryotic clusters to prokaryotic clusters. We mapped these prokaryotic clusters to eukaryotic clusters (see **Methods**) as indicated by red circles in **Fig 1**. Their significance will be discussed in a later section.

The most vertical and lateral genes and categories

Table 1 lists the 20 most vertically and 20 least vertically inherited genes in sequenced prokaryotic genomes, both for the complete sample and for the conserved fraction of genes. Among the most vertical are the ribosomal proteins, ribosomal protein S10 currently being the most vertical protein in genomes, followed by other proteins involved in information processing. The least vertically inherited genes by our conservative tip-based approach, comprise various categories (**Table 1**), the complete lists of genes ranked by verticality is given in **S1 Table**.

Although we have no estimate of ***V*** for rRNA, as its sequence in part defines phyla, the tendency we see for widely distributed protein coding genes to resist LGT would also explain why rRNA is itself so refractory to transfer [13,48], the rRNA genes that are present in a recipient genome are difficult to improve upon or match in functional efficiency, and the rRNA gene product can comprise up to 20% of the cell's dry weight [49]. Genes for rRNA thereby carry great inertia against LGT and are therefore difficult to displace by intruding copies. The rank

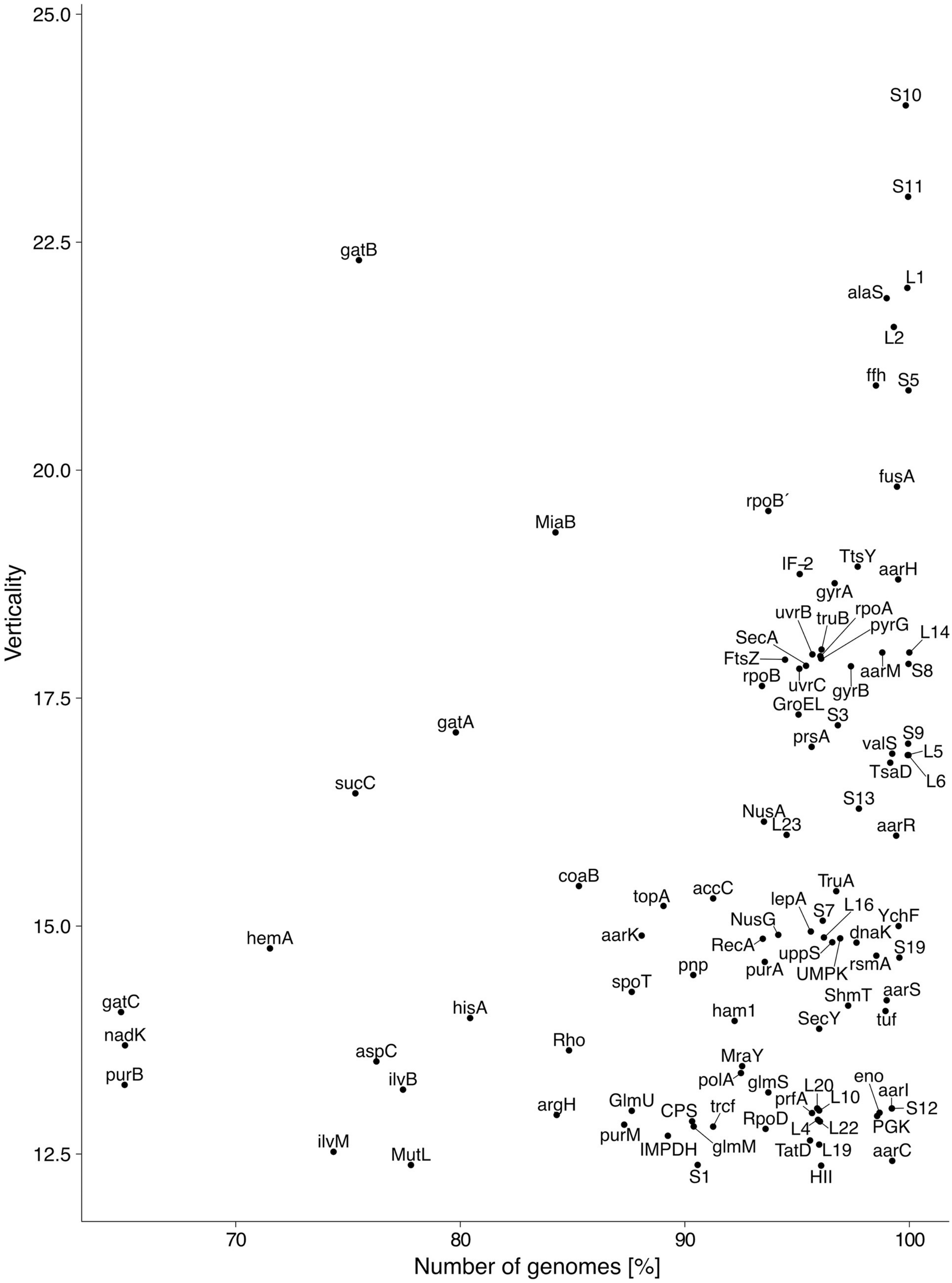


Fig 2. Comparison of estimated verticality and number of genomes [%] for the 100 most vertical clusters. Identity and Annotation of clusters can be found in S6 Table. This is a representation of some of the clusters shown in the blue rectangle of Fig 1A.

https://doi.org/10.1371/journal.pgen.1009200.g002

Table 1. Maximum likelihood trees from 19,050,992 protein sequences from 5,433 bacterial and 212 archaeal species were calculated for clusters obtained by MCL, yielding 101,422 trees with at least four sequences and two taxonomic groups present. Each of the 101,422 trees were assigned a protein label according to the NCBI sequence header that was represented the most. On the left panel all trees were annotated and sorted according to their verticality score for the genes (V_g). The number of organisms in the respective cluster is stated as N_{spec}. On the right panel the same values are stated only for conserved protein families–determined by average branch length ≤ 0.1.

	All 101,422 protein families			The 8,547 most conserved protein families		
	V_g	Protein family	N_{spec}	V	Protein family	N_{spec}
Most vertical						
	24.00	30S ribosomal protein S10	5,646	24.00	30S ribosomal protein S10	5,646
	23.00	30S ribosomal protein S11	5,652	23.00	30S ribosomal protein S11	5,652
	22.30	Asp/glu–tRNA amidotransferase subunit B	4,269	22.30	Asp/glu–tRNA amidotransferase subunit B	4,269
	22.00	50S ribosomal protein L1	5,650	22.00	50S ribosomal protein L1	5,650
	21.89	Alanine–tRNA ligase	5,598	21.89	Alanine–tRNA ligase	5,598
	21.57	50S ribosomal protein L2	5,616	21.57	50S ribosomal protein L2	5,616
	20.93	Sec family type I SRP[a] protein	5,571	20.93	Sec family type I SRP[a] protein	5,571
	20.88	30S ribosomal protein S5	5,653	20.88	30S ribosomal protein S5	5,653
	19.82	Translation elongation factor G	5,624	19.82	Translation elongation factor G	5,624
	19.55	DNA-directed RNA polymerase subunit beta	5,300	19.55	DNA-directed RNA polymerase subunit beta	5,300
	19.32	tRNA methylthiotransferase MiaB	4,764	18.86	Translation initiation factor IF-2	5,379
	18.94	Signal recognition particle-docking protein FtsY	5,525	18.80	Histidine–tRNA ligase	5,627
	18.86	Translation initiation factor IF-2	5,379	18.76	DNA gyrase subunit A	5,467
	18.80	Histidine–tRNA ligase	5,627	18.00	50S ribosomal protein L14	5,655
	18.76	DNA gyrase subunit A	5,467	18.00	Methionine–tRNA ligase	5,587
	18.03	tRNA pseudouridine synthase B	5,434	17.98	Excinuclease ABC subunit B	5,411
	18.00	50S ribosomal protein L14	5,655	17.96	DNA-directed RNA polymerase subunit alpha	5,431
	18.00	Methionine–tRNA ligase	5,587	17.93	CTP synthetase	5,433
	17.98	Excinuclease ABC subunit B	5,411	17.88	30S ribosomal protein S8	5,653
	17.96	DNA-directed RNA polymerase subunit alpha	5,431	17.85	Preprotein translocase subunit SecA	5,395
Most lateral						
	0	Heavy metal-responsive transcriptional regulator	2,392	0	SDH cyt b556 large subunit	2,344
	0	SDH cyt b556 large subunit	2,344	0	RnfH family protein	2,004
	0	Anaerobic ribo.-triP[b] reductase activating protein	2,078	0	Hypothetical protein	1,964
	0	Thiol:disulfide interchange protein DsbC	1,952	0	Amino acid ABC transporter permease	1,666
	0	RnfH family protein	2,004	0	Succinate dehydrogenase, HM[c] anchor protein	1,800
	0	Disulfide bond formation protein B 1	1,808	0	LysR family transcriptional regulator	1,267
	0	Hypothetical protein	1,964	0	Hypothetical protein	1,688
	0	Amino acid ABC transporter permease	1,666	0	Maleylacetoacetate isomerase	1,430
	0	LysR family transcriptional regulator	1,431	0	Sigma-E factor regulatory protein RseB	1,599
	0	Succinate dehydrogenase, HM[c] anchor protein	1,800	0	tRNA synthase TrmP	1,567
	0	LysR family transcriptional regulator	1,267	0	tRNA 5-methoxyuridine(34) synthase CmoB	1,525
	0	Hypothetical protein	1,688	0	Chemotaxis phosphatase CheZ family protein	1,483
	0	Maleylacetoacetate isomerase	1,430	0	Hypothetical protein	1,505
	0	Sigma-E factor regulatory protein RseB	1,599	0	Hypothetical protein	1,345
	0	tRNA synthase TrmP	1,567	0	Outer membrane protein assembly protein	1,301
	0	tRNA 5-methoxyuridine(34) synthase CmoB	1,525	0	Deoxyribonuclease I	1,269
	0	Chemotaxis phosphatase CheZ family protein	1,483	0	Formate dehydrogenase accessory protein FdhE	1,241
	0	Hypothetical protein	1,505	0	Flagellar export protein FliJ	1,208
	0	Hypothetical protein	1,345	0	Hypothetical protein	1,200

(*Continued*)

Table 1. (Continued)

	All 101,422 protein families			The 8,547 most conserved protein families		
	V_g	**Protein family**	N_{spec}	V	**Protein family**	N_{spec}
	0	Hypothetical protein	1,325	0	Hypothetical protein	1,179

Notes

[a] SRP protein–general secretory pathway protein signal recognition particle protein

[b] ribo.-triP–ribonucleoside-triphosphate

[c] HM–hydrophobic membrane

https://doi.org/10.1371/journal.pgen.1009200.t001

of functional categories (Table 2) with respect to verticality reveals that the clusters functionally associated with translation rank highest, followed by nucleotide metabolism (many proteins without intermolecular interactions), replication, folding and vitamin synthesis. Genes for vitamin synthesis are not highly expressed but are widely distributed and are highly vertical. The least vertical categories comprise drug resistance and community interactions. Cognoscenti might surmise that there are no real surprises in the ranking of functional categories

Table 2. Assignment of KEGG level B functional annotations. On the left panel all prokaryotic maximum likelihood trees were annotated and sorted according to their average verticality score (V_{avg}). The number of clusters employed for this analysis are indicated (N_{clust}). The same procedure was performed on the right panel only for conserved protein families–determined by average branch length ≤ 0.1.

All 101,422 protein families			The 8,547 most conserved protein families		
Function	V_{avg}	N_{clust}	**Function**	V_{avg}	N_{clust}
Translation	5.31	2,428	Translation	14.82	284
Metabolism of cofactors and vitamins	4.86	2,443	Nucleotide metabolism	10.21	160
Nucleotide metabolism	4.28	1,419	Metabolism of cofactors and vitamins	7.95	199
Amino acid metabolism	3.83	3,777	Carbohydrate metabolism	7.23	534
Carbohydrate metabolism	3.63	4,836	Replication and repair	7.11	187
Biosynthesis of other secondary metabolites	3.62	507	Energy metabolism	7.07	208
Glycan biosynthesis and metabolism	3.42	3,349	Amino acid metabolism	7.06	438
Metabolism	3.31	4,260	Folding, sorting and degradation	6.77	118
Energy metabolism	3.28	2,705	Metabolism of other amino acids	5.87	81
Xenobiotics biodegradation and metabolism	3.26	1,606	Metabolism	5.67	337
Replication and repair	3.14	3,502	Enzyme families	5.53	164
Transport and catabolism	3.02	2,843	Biosynthesis of other secondary metabolites	5.50	25
Metabolism of terpenoids and polyketides	2.97	1,473	Xenobiotics biodegradation and metabolism	5.36	103
Metabolism of other amino acids	2.95	745	Glycan biosynthesis and metabolism	5.33	158
Transcription	2.84	7,245	Signal transduction	5.10	240
Folding, sorting and degradation	2.79	1,873	Membrane transport	4.69	1,431
Lipid metabolism	2.65	2,864	Cell motility	4.37	124
Enzyme families	2.59	3,735	Metabolism of terpenoids and polyketides	4.31	85
Cellular processes and signaling	2.49	3,905	Transport and catabolism	4.31	143
Signal transduction	2.48	6,712	Lipid metabolism	4.20	215
Membrane transport	2.46	19,992	Transcription	4.12	409
Genetic information processing	2.31	4,838	Cellular processes and signaling	3.75	257
Cellular community prokaryotes	2.21	3,986	Cellular community prokaryotes	3.55	172
Drug resistance	2.15	1,754	Genetic information processing	3.23	269
Cell motility	1.94	3,620	Drug resistance	3.10	88
Poorly characterized	1.41	178,665	Poorly characterized	1.68	2,970

https://doi.org/10.1371/journal.pgen.1009200.t002

with respect to $\boldsymbol{V}$, an indication that our measure of $\boldsymbol{V}$ is recovering meaningful information about gene evolution.

The verticality of phyla

By averaging the verticality of all genes that occur in a given phylum, we can also estimate the verticality of phyla and rank them accordingly. This is shown in **Table 3**, for bacteria and archaea separately, where P_{mono} indicates the proportion of trees in which the given phylum was monophyletic. No phyla were always monophyletic, with values of P_{mono} topping out at about 0.8, meaning that the phylum was monophyletic in 80% of the trees in which its sequences occurred. At the level of phyla, for all genes and for the conserved genes, Acidithiobacilli emerge as the most vertically evolving bacteria, while the Thermococcales and Methanococcales emerge as the most vertically evolving archaea. The most laterally evolving bacteria are the Erysipelotrichia, a group of firmicutes related to Clostridia, and the Clostridia themselves for all genes, while for the conserved genes, the Gammaproteobacteria finish last when it comes to avoiding LGT. The archaea most riddled by LGT are the halophiles, which are methanogens that acquired their respiratory chain and aerobic lifestyle from bacteria [19]. Though not strict, there is a clear tendency for bacteria with a specialist lifestyle to resist LGT, and a tendency for generalists like the divisions of the proteobacteria to harbor less vertically evolving chromosomes, that, is to undergo LGT.

The Gammaproteobacteria were the worst offenders when it came to LGT among the 8,547 conserved gene trees, showing gammaproteobacterial monophyly in less than 20% of trees that contained the phylum. Of course, it is possible that verticality is violated by recurrent exchanges among specific pairs of taxa or by phylogenetic artefact involving true neighbors, which for Gammaproteobacteria would be the Betaproteobacteria in traditional schemes. In order to check for such effects, each time we scored a tip-resident clade in our trees, we also scored the phylogenetic membership within its sister group. A sister group can either itself be monophyletic, containing sequences from only one phylum, or it can be mixed, containing sequences from two or more different phyla. The summary is shown in **Fig 3**, where the frequencies of phyla in the sister group are shown. Note that a phylum can appear as its own sister group when its monophyletic clade is broken by recent LGT to a member of a different phylum: the gene tree does not change, but the taxon label of the donated gene does, leaving sequences of the donor phylum that branch below the recent export in the sister group. This is illustrated in **S3d Fig**. While methanogens and halophilic archaea tend to interleave, as do archaea as a whole, the dominant signal in the sister group plot is that Gammaproteobacteria tend to be the sister of virtually every phylum, meaning that they are the recipient of genes from all phyla in our sample. The tendency to undergo recent LGT—recent because we are only scoring terminal branches—is also clearly manifest in Bacilli, Betaproteobacteria, Alphaproteobacteria and Actinobacteria, all of which harbor lineages with large genomes, large pangenomes, and diverse generalist lifestyles.

The verticality of individual genomes

Averaging the value of verticality across all genes in a genome gives an estimate for the verticality of the genome, $\boldsymbol{V}_g$. The verticality of all genomes investigated here is given in **S4 Table**. The most vertical genomes are those with the highest proportion of genes involved in translation. This is because the process of reductive genome evolution [50] always hones in on the ribosome, translation and information processing, as these functions are prerequisite to gene expression. The widely distributed genes involved in information processing are the most vertical (**Table 1**), such that the gammaproteobacterial endosymbiont *Carsonella ruddii* [51]

Table 3. Verticality of prokaryotic taxa across protein families with at least two taxonomic groups. The list of bacterial (top) and archaeal (bottom) taxa occurring in all trees (right) and only trees that were filtered for conservation (average branch length in the tree ≤ 0.1) (left). Archaeal and bacterial phyla with less than 5 representative species in the dataset were excluded. P_{mono} refers the proportion of monophyletic trees. N_{mono} indicates the number of trees in which this taxon is monophyletic whereas N_{trees} shows the number of occurrences of the phyla in the respective dataset.

		All trees– 101,423				**Conserved trees– 8,547**		
	Taxon	**P_{mono}**	**N_{mono}**	**N_{trees}**		**P_{mono}**	**N_{mono}**	**N_{trees}**
Bacteria								
	Acidithiobacillia	0.81	1,677	2,067		0.91	629	688
	Chlamydiae	0.74	1,378	1,867		0.75	482	642
	Tenericutes	0.68	2,770	4,076		0.50	391	776
	Actinobacteria	0.60	30,050	49,958		0.37	1,214	3,293
	Bacilli	0.59	24,365	41,526		0.25	1,017	3,997
	Chlorobi	0.59	1,728	2,946		0.80	494	619
	Thermotogae	0.57	2,252	3,937		0.65	495	764
	Cyanobacteria	0.56	8,655	15,446		0.64	843	1,319
	Deinococcus-Thermus	0.54	3,156	5,858		0.63	705	1,113
	Synergistetes	0.53	1,001	1,872		0.70	484	692
	Epsilonproteobacteria	0.52	3,815	7,270		0.37	513	1,397
	Fusobacteria	0.51	1,805	3,516		0.60	717	1,194
	Spirochaetes	0.50	5,063	10,130		0.44	683	1,564
	Bacteroidetes	0.49	11,677	23,755		0.40	759	1,879
	Gammaproteobacteria	0.48	29,439	61,803		0.18	1,078	5,874
	Negativicutes	0.45	1,892	4,170		0.59	804	1,371
	Nitrospirae	0.43	1,377	3,180		0.47	359	762
	Alphaproteobacteria	0.43	18,086	41,953		0.35	1,312	3,735
	Aquificae	0.43	1,210	2,826		0.43	290	672
	Planctomycetes	0.40	1,755	4,399		0.55	533	961
	Chloroflexi	0.39	2,349	6,003		0.46	521	1,141
	Acidobacteria	0.38	1,789	4,666		0.58	625	1,077
	Betaproteobacteria	0.38	14,203	37,225		0.34	1,601	4,775
	Deltaproteobacteria	0.37	8,512	23,013		0.38	1,005	2,618
	Verrucomicrobia	0.36	1,146	3,152		0.56	601	1,067
	Clostridia	0.32	7,481	23,638		0.34	1,084	3,196
	Erysipelotrichia	0.17	344	2,001		0.43	451	1,058
Archaea								
	Thermococcales	0.73	2,482	3,380		0.79	271	341
	Methanococcales	0.73	1,612	2,220		0.83	236	283
	Methanobacteriales	0.68	1,949	2,857		0.79	282	356
	Sulfolobales	0.66	2,223	3,387		0.75	280	374
	Archaeoglobales	0.62	1,415	2,286		0.79	252	318
	Methanomicrobiales	0.60	1,616	2,693		0.74	301	406
	Methanosarcinales	0.60	3,392	5,654		0.63	318	503
	Thermoproteales	0.55	1,537	2,775		0.61	257	420
	Thermoplasmatales	0.49	662	1,364		0.58	212	366
	Desulfurococcales	0.41	852	2,072		0.44	130	298
	Natrialbales	0.32	1,459	4,503		0.42	246	588
	Haloferacales	0.27	980	3,593		0.40	205	513
	Halobacteriales	0.20	1,024	5,057		0.30	178	591

https://doi.org/10.1371/journal.pgen.1009200.t003

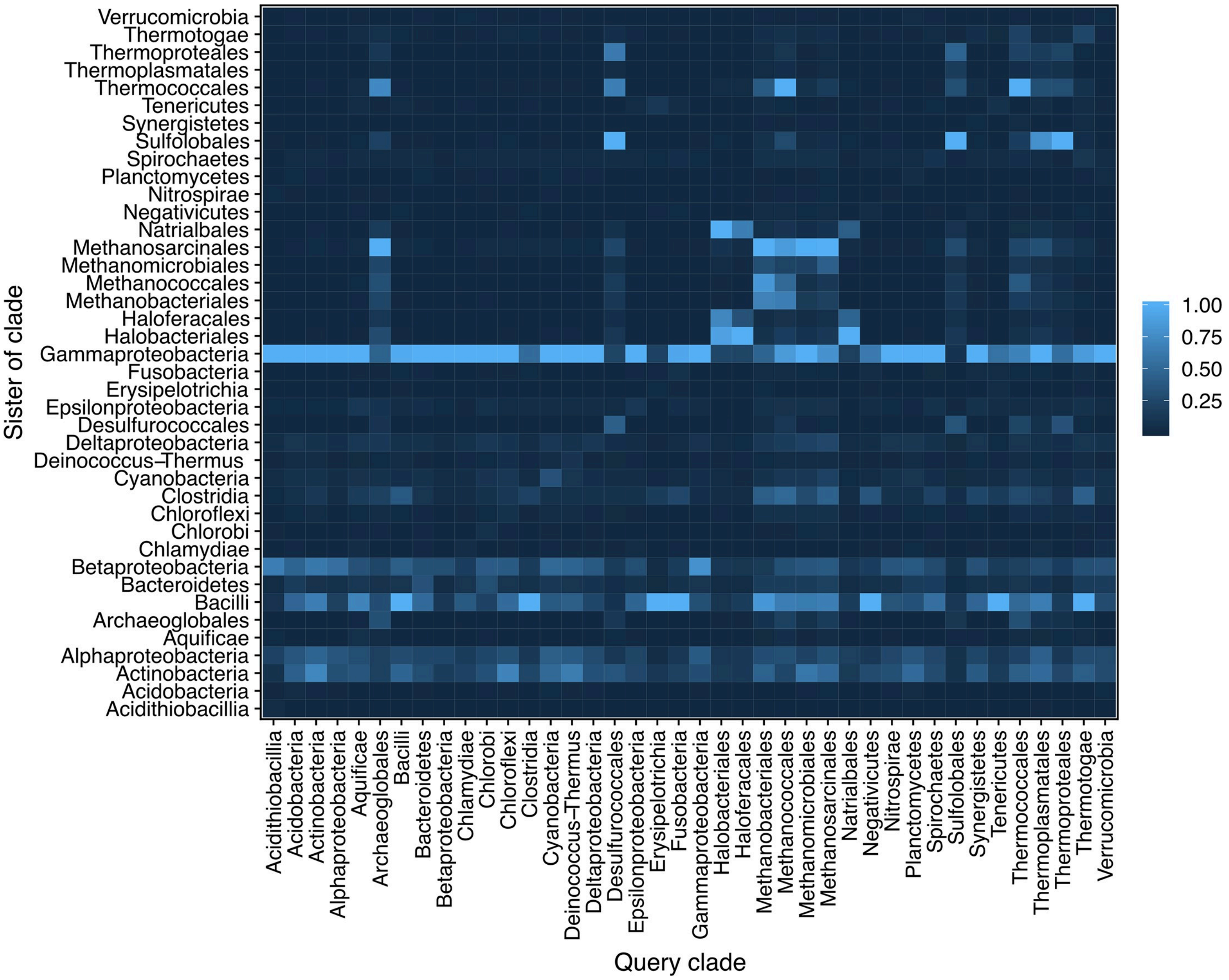


Fig 3. Relative occurrence of a taxonomic group as the sister group of each clade in the unrooted trees. For each taxonomic group in a cluster the sister was determined and counted. Multiple occurrences of different groups in the sister were accounted for by their relative occurrence. If the taxonomic group was paraphyletic, each monophyletic subgroup was determined and the sister of these were noted. The values of these subgroups were added up by multiplying the individual values of the sister by the fraction of the subgroup of the whole taxonomic group. To compare, the final values of each taxonomic group were normalized by dividing by the highest count a possible sister has gotten. It is apparent that Gammaproteobacteria are always overrepresented. It is not clear if the observed effects are due to overrepresentation of certain taxa in the data set or due to relative abundance of LGT.

https://doi.org/10.1371/journal.pgen.1009200.g003

which possesses only 166 protein coding genes, is the most vertical genome in our sample with $\boldsymbol{V}_g = 9.44$. Conversely, the least vertical genomes are the largest, with the actinobacterium *Amycolatopsis mediterranei* ($\boldsymbol{V}_g = 0.84$) having a genome over 10 Mb coming in last. Among the archaea, the most vertical genomes were those of H_2 dependent autotrophs (**S4 Table**). The most vertical genome was that of the highly reduced free living archaeon, *Ignicoccus hospitalis* [52] ($\boldsymbol{V}_g = 4.10$) an extreme specialist that grows only on $H_2 + S^0$, followed by nine H_2 dependent methanogens, starting with the thermophilic methanogen *Methanothermus fervidus* ($\boldsymbol{V}_g = 4.09$), with a genome of 1.2 Mb [53]. The most lateral archaeal genome was that of the halophile *Haloterrigena turkmenica* ($\boldsymbol{V}_g = 1.66$).

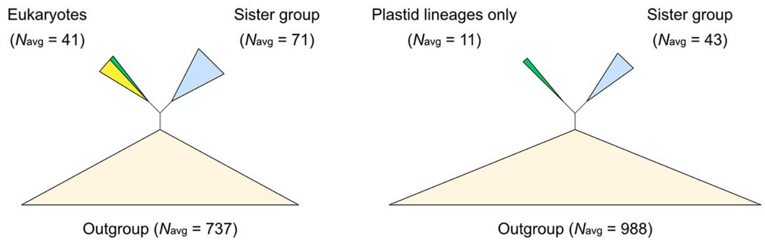
Eukaryotes

In an ideal world of vertically inherited genes and infallible phylogeny, all genes would produce the same tree and all eukaryotic genes would trace to the same alphaproteobacterium (the mitochondrion) and the same are archaeon (host), plus the same cyanobacterium in the case of eukaryotes with plastids. But the real data from real genomes reveals that only a small minority of prokaryotic genes, much less than 1%, tend to be inherited vertically. How does the non-verticality of prokaryotic genes, genomes, and phyla impact our ability to infer the origin of eukaryotic genes? For all 3,420,731 protein coding genes from 150 eukaryotic genomes, we constructed clusters, merged them with their cognate prokaryotic clusters to generate eukaryote-prokaryote clusters (EPCs), constructed alignments and ML trees (see **Methods**). The red circles in **Fig 1** mark the prokaryotic clusters that were merged with their unique cognate eukaryotic clusters. The first question concerned eukaryote monophyly. There are many claims in the literature for LGT from prokaryotes to eukaryotes, but few are supported by prokaryotic reference samples that reflect the availability of genome data and fewer still, if any, are supported by systematic tests for eukaryote monophyly. Therefore, we looked closely at the possibility of LGT vs. eukaryote monophyly in our sample.

Among the 2,575 maximum-likelihood (ML) trees reconstructed from the merged eukaryote-prokaryote clusters, only 475 of the best trees found (18.4%) failed to recover eukaryotes as monophyletic. Does that finding represent evidence for LGT to eukaryotes in 18% of these trees, that is, is the best tree identified significantly better than the case of eukaryote monophyly? To test whether the lack of eukaryote monophyly in those 475 trees is due to reconstruction errors or due to prokaryote-eukaryote LGT, we compared the ML trees against trees with constrained eukaryote monophyly using likelihood tests. We employed the Kishino-Hasegawa test (KH), the approximately-unbiased test (AU) and the Shimodaira-Hasegawa test (SH) (see **Methods** for details). At the 5% significance level (p-value ≤ 0.05), the KH test rejected eukaryote monophyly for 6% of the trees (30 out of 475), that is, the null hypothesis (eukaryote monophyly) was rejected at a rate very close to that expected by chance. The AU test rejected eukaryote monophyly for 3 trees while the SH test did not reject eukaryote monophyly for any tree at the p-value of ≤ 0.05 (**S4 Fig** and **S5 Table**). Thus, the absence of a pure eukaryotic clade in some of the best trees found by ML trees results from challenges in distinguishing alternative trees that are statistically identical to the true tree, or to trees recovering eukaryote monophyly, in terms of their likelihood values, a problem that becomes more acute for phylogenetic inference using large samples because the tree space for the ML method to search grows exponentially. In terms of traits, eukaryotes are the strongest monophylum in nature, a status corroborated by the lack of any evidence that would support a case for the non-monophyly (LGT) of eukaryotic genes.

What do trees say about the origin of eukaryotic genes? In the following, to avoid the effects of trees for poorly conserved genes (**Fig 1A**), we consider only those 685 trees in which the eukaryotic cluster mapped to one of the conserved prokaryotic clusters in **Fig 1B**. For each tree, we determined the prokaryotic sister group to the eukaryotic clade, and scored whether it was a pure sister containing sequences from only one prokaryotic phylum or a mixed sister group containing a mixed sister group from two or more phyla. The results are summarized in **Fig 4B**.

By the measure of phylogenetic inference, every prokaryotic phylum sampled in our study appears as a donor of genes to the eukaryote common ancestor, either by presence in a mixed sister group or as a pure sister (**Fig 4B**). This is true not only for bacteria, which would be expected to trace mitochondrial ancestry, but also for archaea, which since their discovery have been linked to the origin of the host. Can we naïvely interpret such observations at face a.



b.

	All taxa, except photosynthetic only Filtered for conservation – 456 trees						Photosynthetic taxa only Filtered for conservation – 229 trees						
Taxon	P_{mono}	S_{non}	S_{mix}	S_{pure}	$S_{p,avg}$	N_{trees}	P_{mono}	S_{non}	S_{mix}	S_{pure}	$S_{p,avg}$	N_{trees}	N_{gen}
Bacteria													
Acidithiobacillia	0.96	59	8	1	6.0	68	1	46	2	0	-	48	5
Chlamydiae	0.77	52	4	4	30.5	60	0.76	36	3	3	38.0	42	113
Chlorobi	0.75	55	5	0	-	60	0.90	48	1	0	-	49	12
Synergistetes	0.68	52	10	0	-	62	0.88	41	0	0	-	41	5
Deinococcus-Thermus	0.66	97	13	1	1.0	111	0.66	53	2	1	4.0	56	26
Thermotogae	0.66	55	12	0	-	67	0.69	40	2	0	-	42	32
Cyanobacteria	0.53	99	23	4	55.3	126	0.52	47	11	41	48.0	99	94
Tenericutes	0.52	51	11	2	1.0	64	0.24	32	0	1	16.0	33	141
Fusobacteria	0.51	65	9	1	20.0	75	0.69	42	0	0	-	42	20
Epsilonproteobacteria	0.51	78	12	1	8.0	91	0.74	48	1	1	1.0	50	253
Planctomycetes	0.47	91	22	3	1.0	116	0.57	48	7	3	1.0	58	8
Nitrospirae	0.46	55	8	0	-	63	0.24	49	2	0	-	51	8
Negativicutes	0.46	82	10	2	1.0	94	0.56	48	0	0	-	48	12
Verrucomicrobia	0.45	77	19	2	1.5	98	0.47	54	3	2	2.5	59	10
Acidobacteria	0.43	92	20	2	1.0	114	0.63	59	3	1	1.0	63	8
Chloroflexi	0.32	94	18	2	1.0	114	0.30	56	4	3	1.0	63	26
Aquificae	0.30	53	10	0	-	63	0.24	48	1	1	1.0	50	16
Spirochaetes	0.29	109	28	3	5.7	140	0.30	60	2	2	24.5	64	66
Bacilli	0.27	145	22	16	7.9	183	0.27	82	4	4	1.0	90	1,012
Actinobacteria	0.26	165	41	29	8.4	235	0.33	94	7	9	1.4	110	593
Erysipelotrichia	0.24	53	9	0	-	62	0.26	42	0	0	-	42	8
Bacteroidetes	0.22	137	31	12	3.8	180	0.28	70	8	4	6.0	82	185
Deltaproteobacteria	0.21	128	52	13	2.0	193	0.23	78	15	5	5.2	98	74
Alphaproteobacteria	0.21	144	41	35	59.0	220	0.26	77	11	16	5.1	104	460
Clostridia	0.21	116	24	9	1.7	149	0.22	71	2	0	-	73	169
Betaproteobacteria	0.21	144	51	13	8.6	208	0.35	93	13	3	1.0	109	493
Gammaproteobacteria	0.19	170	45	45	25.1	260	0.16	81	19	24	46.7	124	1,565
Other bacteria	-	-	-	-	-	143	-	-	-	-	-	77	29
Archaea													
Methanococcales	0.88	60	3	1	1.0	64	0.84	27	3	1	1.0	31	16
Archaeoglobales	0.85	64	6	3	6.3	73	0.81	25	3	4	4.5	32	8
Sulfolobales	0.84	77	4	1	29.0	82	0.86	27	0	1	5.0	28	29
Methanobacteriales	0.83	73	2	2	8.0	77	0.81	23	3	1	4.0	27	17
Thermococcales	0.82	73	8	3	22.3	84	0.78	21	1	1	14.0	23	27
Methanomicrobiales	0.80	68	7	6	3.3	81	0.66	24	5	3	6.0	32	9
Thermoproteales	0.46	73	15	3	4.3	91	0.42	20	4	2	1.5	26	10
Methanosarcinales	0.43	78	6	11	5.4	95	0.59	35	2	7	12.9	44	32
Thermoplasmatales	0.29	70	11	4	3.3	85	0.38	21	7	1	1.0	29	5
Haloferacales	0.28	91	26	3	2.7	120	0.10	32	6	1	1.0	39	7
Desulfurococcales	0.26	64	14	2	1.5	80	0.38	21	4	4	1.3	29	11
Natrialbales	0.26	91	26	3	2.7	120	0.33	35	5	3	1.7	43	9
Halobacteriales	0.17	112	9	4	1.0	125	0.11	39	3	2	1.5	44	16
Other archaea	-	-	-	-	-	131	-	-	-	-	-	51	16

Fig 4. Identification of the prokaryotic sister group to the eukaryotes in 2,575 eukaryotic-prokaryotic unrooted gene trees (EPC). a. shows the average clade sizes for eukaryotes, the sister group to eukaryotes and the outgroup in the analyzed trees for (right) the 229 trees with only plastid derived lineages and (left) for the 456 EPCs containing all taxa except photosynthetic lineages. **b.** details the list of bacterial (top) and archaeal (bottom) phyla occurring in the trees with only plant lineages (right) and all other trees (left) that were filtered for conservation (average branch length of the tree ≤ 0.1). Archaeal and bacterial phyla with less than 5 representative species in the dataset were collapsed into 'other archaea' and 'other bacteria' groups. P_{mono} refers to the proportion of trees with a branch (split) separating the species of the respective phylum from all the others in the tree; S_{non} refers to the number of occurrence of the phylum only in the outgroup clade; S_{mix} refers to the number of occurrences of the phylum as a mixed sister (more than one phylum in the clade); S_{pure} refers to the number of occurrences of the phylum as pure sister (as the single phylum); $S_{p,avg}$ shows the average size of the sister clade when the phylum occurs as a pure sister clade. N_{trees} show the number of occurrences of the phyla across the trees and N_{gen} indicates the number of species in each taxon included in the complete dataset.

https://doi.org/10.1371/journal.pgen.1009200.g004

value? Is it reasonable to believe that every phylum sampled here donated a gene, or several, to eukaryotes at their origin? If we break the data down to families, genera, or species, the number of donors grows accordingly (all prokaryotic organisms employed in this study were in the sister group to eukaryotes at least once), such that each gene in eukaryotes would correspond to an individual donation, as some would argue [54]. But that logic leads straight to the erroneous conclusion that ancestral plastid and mitochondrial genomes were assembled by acquisition *one gene at a time* [55] the converse of what they are in plain sight, namely reduced genomes of single bacterial endosymbionts [50] that underwent reductive evolution by transferring genes to the nucleus. Worse yet, the same problem ensues at the origin of plastids (**Fig 4B**, right column), because for photosynthetic eukaryotes again all phyla, including the archaea, appear as donors. Many genes that are germane to photosynthesis in eukaryotes trace to the plant common ancestor (plants being monophyletic) but only a minority of them trace specifically to Cyanobacteria, and those that do, do not trace to the same cyanobacterium [56,57].

If we only consider pure sister groups to eukaryotes, the most common apparent gene donor was Gammaproteobacteria, followed by Alphaproteobacteria, Actinobacteria and Bacilli. There is at least one theory in the literature invoking the participation of those groups at eukaryote origin [58]. However, a similar pattern recurs for plastids, which have the strongest pure sister signal from Cyanobacteria followed again by Gammaproteobacteria (for which there is no plastid origin theory) and Alphaproteobacteria. The problem of inferring symbionts from gene trees becomes more evident when we consider apparent archaeal contributions to the origins of plastids (**Fig 4B**), because there are no archaea that synthesize chlorophyll. We are confronted with a conflict. Blind inference of symbionts from trees cannot account for the origin of organelle genomes, the strongest form of evidence for the origin of organelles in the first place. The 'one tree one donor' logic carries a weighty premise that is never spelled out by its proponents, namely that the donated genes never underwent LGT among free living prokaryotes in the 1.5 billion years since organelle origin. If we approach the problem from the standpoint of theory testing in the presence of prior knowledge about the underlying process, namely one symbiont 1.5 billion years ago (as evidenced by the single origin of plastids and mitochondria respectively), what would look like many donors if we were to assume that prokaryotic gene evolution is vertical, is clearly the result of LGT among free-living prokaryotes, where, in real data, gene evolution is lateral.

For example, were the gammaproteobacterial signal in heterotrophic eukaryotes a result of gene acquisitions from donors with gammaproteobacterial rRNA, then that same signal would reflect a gammaproteobacterial origin of plastids (**Fig 4B**), which seems unlikely and is not covered by any theory. If on the other hand it were due to the low verticality of Gammaproteobacteria as a phylum, then Gammaproteobacteria should appear as the sister to many different groups of prokaryotes, which is precisely the observation (**Fig 3**). We asked whether there is a non-random signal across all genes that singles out Cyanobacteria (plastids) and Alphaproteobacteria (mitochondria) specifically as donors. This is shown in **Fig 5**, where we have plotted the distribution of trees that identify Alphaproteobacteria, Cyanobacteria or Gammaproteobacteria as pure sisters to (donors of) eukaryotic genes. Though Gammaproteobacteria appear as the pure sister in many trees (**Fig 4B**), the genes that do so are primarily of low verticality. Only the Alphaproteobacteria have a significant enrichment of vertical genes as sisters relative to the sample (**Fig 5A**), but the significance is marginal ($p < 0.01$). The Cyanobacteria are not significantly enriched in high verticality sisters, because of a large number of low verticality cases (**Fig 5C and 5D**). The majority of the gammaproteobacterial sister cases are low verticality genes (**Fig 5E and 5F**).

Throughout this discussion, we recall that the ancestor of mitochondria was not a phylum of proteobacteria, it was a single proteobacterium that engaged in a singular symbiosis. The

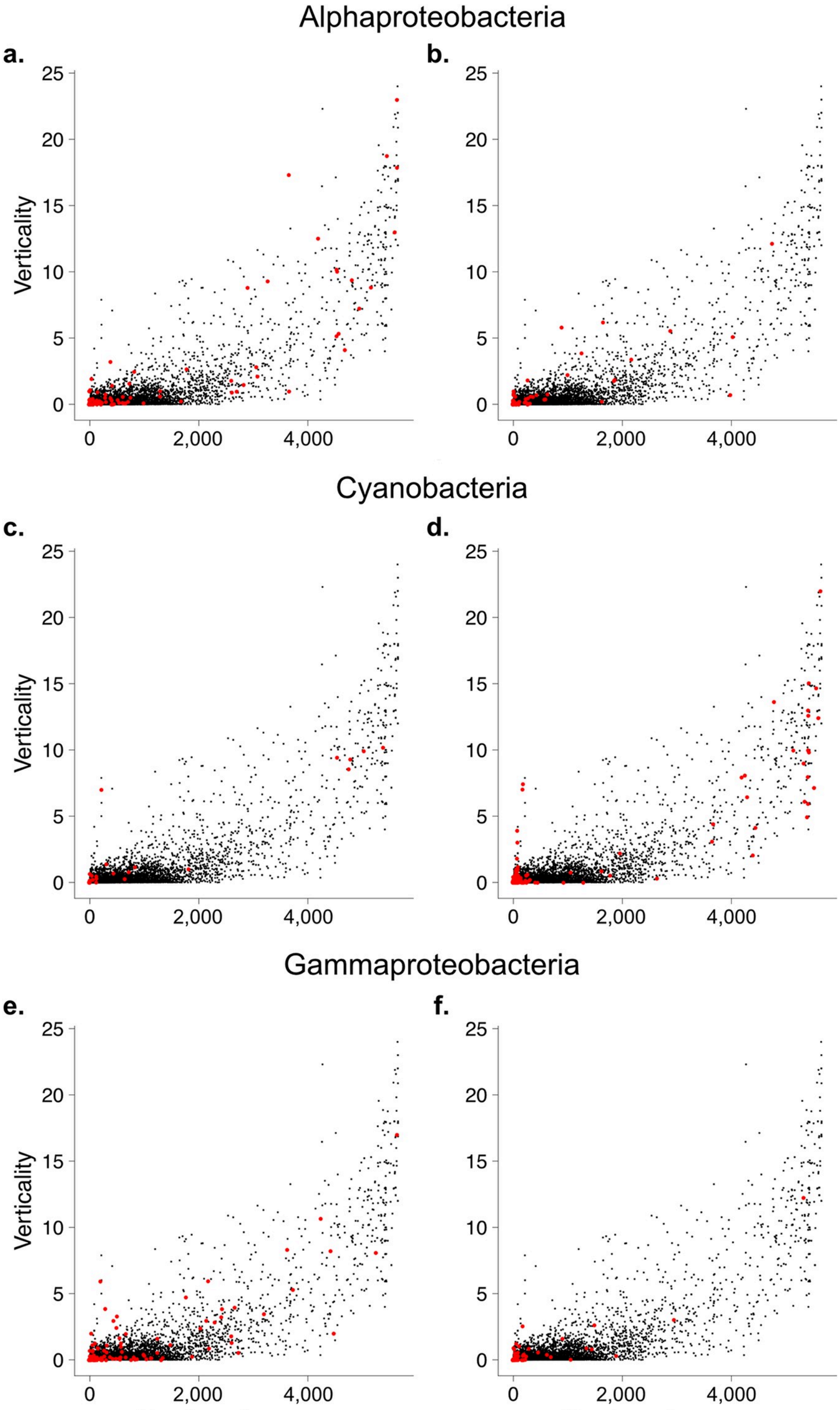


Fig 5. Mapping of EPCs to prokaryotic clusters. The EPCs were separated according to the pure sister group of eukaryotes in the unrooted trees: **a.** and **b.** Alphaproteobacteria, **c.** and **d.** Cyanobacteria, **e.** and **f.** Gammaproteobacteria. The left panel shows EPCs that may include all eukaryotic supergroups but no groups that include only photosynthetic lineages, the right panel shows only EPCs that only include photosynthetic eukaryotes (lineages from SAR, Hacrobia and Archaeplastida). Meaning the latter are indicative of plastid endosymbiosis. Plots for all taxa see S5 Fig.

https://doi.org/10.1371/journal.pgen.1009200.g005

same is true for plastids, whose origin was not the result of a symbiosis with the cyanobacterial phylum, it was a symbiosis with a single cyanobacterium. The genes that trace to those organelle origin events are, however, like almost all prokaryotic genes, of low verticality within prokaryotes.

A critic might ask whether eukaryotes, if their genes are of monophyletic origin relative to prokaryotes, score higher than all prokaryotes in terms of a comparable measure of verticality (supergroups instead of phyla). The problem there is a different one, namely paralogy. The underlying theme of eukaryotic genome evolution is recurrent gene and genome duplication [59,60], massive paralogy impairs eukaryote gene monophyly although gene duplications carry phylogenetic information in their own right [35]. The genes that have remained in plastid and mitochondrial genomes encode proteins involved in the electron transport chain of the bioenergetic membrane supporting photosynthesis and respiration, respectively, and the ribosomal proteins [61] involved in synthesizing those proteins in the organelle [62]. Why do those ribosomal proteins reflect an alphaproteobacterial [46] and cyanobacterial [56] origin of the organelle more clearly than non-ribosomal genes? It is not because non-ribosomal genes were acquired from different biological donors. Rather, it is because the prokaryotic reference set of ribosomal proteins is inherited in a vertical manner among free living prokaryotes; all other prokaryotic genes are inherited more laterally (**Fig 1**), evoking the illusion of many different donors to eukaryotes in phylogenetic analyses (**Fig 4B**). Yet that illusion rests upon the tacit assumption that prokaryotes inherit their genes vertically, which is however untrue [2,34,63,64,65].

Discussion

Even though gene evolution in prokaryotes has substantial lateral components, rRNA-based investigations and some protein phylogenetic studies tend to recover groups that microbiologists recognized long before molecular systematics. Hence the groups are in some cases real and there must be a vertical component to prokaryote evolution. The vertical component has, however, been difficult to quantify across lineages. Equally elusive have been estimates for verticality itself, yet suitable methods to quantify that component have been obscure, as have means to quantify verticality across prokaryotic genes. Quantification of discordance in tree comparisons represents one approach [66] to estimate LGT or lack thereof, but its utility is limited when large genome samples are involved, because the number of possible trees exceeds the number that a computer can examine by hundreds of orders of magnitude for trees containing 60 leaves or more. By exploiting the common wisdom that phylogeny works better at the tips of trees than at their deeper branches, we have obtained robust estimates of verticality.

Though many genes that are currently used in molecular systematic studies based on their widespread occurrence have low verticality, across all genes $\boldsymbol{V}$ does increase with distribution density. We suggest that this is so because the displacement of a well-regulated preexisting copy is less likely than the transient and rarely permanent, in some cases lineage founding [67], acquisition of novel traits. That most genes in prokaryotes have both restricted distribution and low verticality underscores the need to identify genes that are inherited vertically across large data sets for the purpose of higher-level broad scale phylogenetic analyses. We found no genes among the 101,422 total clusters and 8,547 conserved clusters that recovered monophyly of all 40 phyla. At the same time all phyla were disguised as gene donors to eukaryotes both at the origin of mitochondria and at the origin of plastids because of LGT among the prokaryotic reference set.

The spectrum of verticality across genes observed here precludes the need to propose, based on trees that implicate non-alphaproteobacterial or non-cyanobacterial gene donors, genetic contributors at the origin of eukaryotes beyond the host, the mitochondrion and, later, the cyanobacterial antecedent of plastids, because LGT among prokaryotes can account for the same gene-tree based observations, more directly and with fewer corollary assumptions, while simultaneously accounting for a larger set of observations among the prokaryotic reference set. The criterion of verticality can furthermore be of practical use in the selection of genes for molecular systematic studies.

Methods

Prokaryotic dataset

Protein sequences for 5,655 prokaryotic genomes were downloaded from NCBI [68] (version September 2016; see **S3 Table** for detailed species composition). We performed all-vs-all BLAST [69] searches (BlastP version 2.5.0 with default parameters) and selected all reciprocal best hits with e-value $\leq 10^{-10}$. The protein pairs were aligned with the Needleman-Wunsch algorithm [70] (EMBOSS needle) and the pairs with global identity values $<$ 25% were discarded. The retained global identity pairs were used for clustering using Markov clustering algorithm [71] (MCL) version 12–068, changing default parameters for pruning (-P 180000, -S 19800, -R 25200). Clusters distributed in at least 4 genomes spanning 2 prokaryotic phyla were retained, resulting in 101,422 used clusters in total. Sequence alignments for each cluster were generated using MAFFT [72], with the iterative refinement method that incorporates local pairwise alignment information (L-INS-i; version 7.130). The resulting alignments were used to reconstruct maximum-likelihood trees with RAxML version 8.2.8 [73] (parameters: -m PROTCATWAG -p 12345) (**S9 Table**). The trees were rooted with the Minimal Ancestor Deviation method (MAD) [74].

Eukaryotic dataset

Protein sequences for 150 eukaryotic genomes were downloaded from NCBI, Ensembl Protists and JGI (see **S7 Table** for detailed species composition). To construct gene families, we performed an all-vs-all BLAST [66] of the eukaryotic proteins (BlastP version 2.5.0 with default parameters) and selected the reciprocal best BLAST hits with e-value $\leq 10^{-10}$. The protein pairs were aligned with the Needleman-Wunsch algorithm (EMBOSS needle) [70] and the pairs with global identity values $<$ 25% were discarded. The retained global identity pairs were used to construct gene families with the MCL algorithm [71] (version 12–068) with default parameters. We considered only the gene families with multiple gene copies in at least two eukaryotic genomes. Protein-sequence alignments for the multi-copy gene families were generated using MAFFT [72], with the iterative refinement method that incorporates local pairwise alignment information (L-INS-i, version 7.130). The alignments were used to reconstruct maximum likelihood trees with IQ-tree [75], applying the parameters '-bb 1000' and '-alrt 1000' (version 1.6.5), with subsequent rooting with MAD [74].

Eukaryotic-prokaryotic dataset

To assemble a dataset of conserved genes for phylogenies linking prokaryotes and eukaryotes, eukaryotic, archaeal and bacterial protein sequences were first clustered separately before homologous clusters between eukaryotes and prokaryotes were identified. Eukaryotic protein sequences from 150 genomes (**S7 Table**) were clustered with MCL [71] using global identities from best reciprocal BLAST hits for protein pairs with e-value $\leq 10^{-10}$ and global identity $\geq$ 40%. The clusters with genes distributed in at least two eukaryotic genomes were retained. Similarly, prokaryotic protein sequences from 5,655 genomes were clustered using the best reciprocal BLAST for protein pairs with e-value $\leq 10^{-10}$ and global identity $\geq$ 25% (for archaea and bacteria, separately). The resulting clusters with gene copies in at least 5 prokaryotic genomes were retained. Eukaryotic and prokaryotic clusters were merged using the reciprocal best cluster procedure [57]. We merged a eukaryotic cluster with a prokaryotic cluster if $\geq$ 50% of the eukaryotic sequences in the cluster have their best reciprocal BLAST hit in the same prokaryotic cluster and vice-versa (cut-offs: e-value $\leq 10^{-10}$ and local identity $\geq$ 30%) yielding 2,587 eukaryotic-prokaryotic clusters (EPCs). EPCs with ambiguous cluster assignment were discarded. Protein-sequence alignments for 2,575 EPCs were generated using MAFFT (L-INS-i, version 7.130); for twelve clusters, the alignment did not compute as sequence quality was low. The alignments were used to reconstruct maximum-likelihood trees with IQ-tree (version 1.6.5) employing the parameters '-bb 1000' and '-alrt 1000' (**S5 Table**).

Verticality

The verticality measure for each gene was defined as the sum of monophyly scores for all monophyletic taxa present in the unrooted trees. Only for the calculation of the average root-to-tip measurements (**S2 Fig**) rooted trees were necessary. This supplementary analysis was then performed with MAD rooted trees. Our species set contains 42 taxa corresponding mostly to phyla level, except for Proteobacteria, Firmicutes and Achaea (see **S8 Table**). For a given tree, the monophyly score S_a for taxon a was defined as:

$$S_a = {}^{n_a}/_{N_a},\ \text{if } a \text{ is monophyletic in tree}$$

$$S_a = 0,\ \text{otherwise}$$

where n_a is the number of species in the tree affiliated to a and N_a is the total number of species from a among the 5,655 genomes in our set. The verticality measure V_g for a gene was then defined as:

$$V_g = \sum S_a,\ \text{for all taxa } a \text{ present in tree}$$

The analyses were conducted with custom scripts using NewickUtilities [76] and ETE [77]. Taxon and genome verticality were defined as the average gene verticality across all gene trees where the taxon (or genome) were present. In addition, weighted taxon verticality for each taxon was defined as the weighted average across all gene trees where the phylum appears, weighted meaning here that values of monophyletic clusters were summed up while values of paraphyletic clusters were subtracted.

Functional annotation

Two annotation strategies were performed for each protein cluster. First, protein annotation information according to the BRITE (Biomolecular Reaction pathways for Information Transfer and Expression) hierarchy was downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG v. September 2017) website [78], including protein sequences and their assigned function according to the KO numbers. The protein sequences of the 5,655 organisms were mapped to the KEGG database using local alignments with 'blastp'. Only the best BLAST hit of the given protein with an e-value $\leq 10^{-10}$ and alignment coverage of 80% was selected. After assigning a function based on the KO numbers of KEGG for each protein in the clusters, the majority rule was applied to identify the function for each cluster. The occurrence of the function of each protein in the cluster was added and the most prevalent function was assigned for each cluster.

The second annotation used the NCBI headers. For this, the appearance of a word among all sequence headers of a cluster was counted. Then, each header was given a score based on the sum of how often its words appeared among all headers. The header with the highest score was then chosen as the cluster annotation.

Tests for eukaryote monophyly

For 475 gene trees where eukaryotes were not recovered as monophyletic, we conducted the Kishino-Hasegawa (KH) test [79], the Shimodaira-Hasegawa (SH) test [80] and the approximately-unbiased (AU) test [81] to assess whether the observed non-monophyly was statistically significant. We reconstructed trees constraining eukaryotic sequences to be monophyletic, but not imposing any other topological constraint, using FastTree [82] (version 2.1.10 SSE3) and recording all trees explored during the tree search with the '-log' parameter. The sample of monophyletic trees were used as input in IQ-tree (version 2.0.3; parameter: '-zb 100000 -au') to perform the KH, SH and AU tests against the unconstrained tree (non-monophyletic). If the best constrained tree did not show significant difference relative to the unconstrained tree (p-value ≤ 0.05), then we considered that eukaryotic monophyly cannot be rejected.

Sister analyses

Prokaryotes. The sister for each prokaryotic taxon was defined as the clade with the smallest branch to the query clade. Two cases had to be differentiated: Mono- or paraphyletic taxonomic groups in a tree. Monophyly was tested as described above with NewickUtilities. For these taxonomic groups, the sister groups could also be directly obtained by using NewickUtilities (nw_clade -s). Finally, all different taxonomic groups in the sister groups were given a score equal to their proportion in the sister group. For paraphyly of a taxonomic group (main group), the monophyletic subgroups were determined with the python package ETE 3 [77]. Each of these subgroups was handled as an individual group in the cluster and the sister clades were determined. Again, if several taxonomic groups were present in a sister group, then these were given a score equal to their proportion in the sister. To get from the scores of each subgroup to the total score of the main group, each subgroup´s scores was multiplied by the proportion of genomes the subgroup has of the main group. Subsequently, the score of a potential sister group to the main group was calculated by summing up its adjusted score over all subgroups. For each taxonomic group, sister scores were normalized by dividing each score through the highest sister score and then plotted as a heatmap.

Eukaryotes. To infer the prokaryotic sisters to eukaryotes we used 2,575 EPC trees. The majority of the EPC trees (2,100) support eukaryotic monophyly. For 475 trees for which eukaryotes did not branch together we recalculated trees constraining eukaryotic monophyly because the Shimodaira-Hasegawa tests failed to reject eukaryotic monophyly for all the 475 trees (see **Methods** section 'tests for eukaryote monophyly' and main text). Note that in unrooted trees for which eukaryotes are monophyletic, the prokaryotic side of the tree is bisected by one internal node into two prokaryotic subclades, each subclade being the potential sister to eukaryotes (**Fig 4A**). We considered the prokaryotic subclade with the smallest number of leaves for our inferences of sister-relations.

Terminal gene duplications

Terminal gene duplications were inferred using the rooted gene trees as pairs of genes sampled from the same genome that appeared as reciprocal sisters in the tree. Gene trees with ambiguous MAD roots were discarded.

Statistical tests

To test the correlations of variables, the Pearson´s correlation test was used [83]. The test results of various combinations for example Number of genomes and number of phyla, that are not mentioned in the text are given in **S2 Table**.

Supporting information

S1 Table. All relevant information about all 101,422 clusters employed in this study.
(XLSX)

S2 Table. Calculated correlations for Fig 1 and S1 Fig.
(TIF)

S3 Table. List of all prokaryotic organisms.
(TXT)

S4 Table. Average verticality per genome and per taxonomic group (phylum).
(XLSX)

S5 Table. List of all 2,575 EPC trees with information if likelihood ratio test was performed.
(XLSX)

S6 Table. Identity and Annotation of the 100 most vertical clusters.
(XLSX)

S7 Table. List of all eukaryotic organisms.
(TXT)

S8 Table. List of all 42 taxonomic groups with labels.
(TXT)

S9 Table. List of all 101,422 RAxML-MAD rooted prokaryote-only trees employed in this analysis.
(DOCX)

S10 Table. Underlying data for S2 Fig.
(XLSX)

S1 Fig. Cumulative distribution function of the fraction of terminal duplicates normalized for genome size compared to the distributions in eukaryotes versus prokaryotes using all genes. a. Shows the cumulative frequency of the proportion of duplications of all 5,655 prokaryotic organisms (red) compared to the 150 eukaryotes (blue) in our dataset. **b.** Shows the cumulative frequency of 100 random sample sets of 150 prokaryotic organisms each (red) versus the 150 eukaryotic organisms (blue) in the dataset.
(TIF)

S2 Fig. Relationship of Verticality, calculated from average root-leave distance in MAD rooted trees, and number of genomes in cluster. Comparison of verticality, normalized by multiplying raw monophyly count by their average root to leave distance of each tree, and number of genomes in a protein cluster for **a.** all clusters (n = 101,422) and **b.** all conserved clusters (average branch length $\geq$ 0.1; n = 8,547). The plot is created analogous to Fig 1 in the main text and this alternative verticality calculation also correlates to number of genomes (A: $p < 10–300$, Pearson´s R2 = 0.571; B: $p < 10–300$, R2 = 0.686). The correlation is more consistent when comparing verticality to number of phyla represented in a cluster (a: $p < 10–300$,

Pearson´s R2 = 0.754; b: p < 10–300, R2 = 0.767, see S2 Table for more details). The eukaryotic-prokaryotic clusters (EPCs) are highlighted in red and the clusters that correspond to a gene from the mitochondrial genome of *Reclinomonas americana* [45] are displayed in blue triangles along the abscissa of the plot and in the graph. For the latter, the gene identifier was noted above each plot. Ribosomal proteins are indicated by the black diamond on the right of each plot and in the graph [6]. Notably, these clusters show a steep decline in clusters with lower verticality among the conserved clusters.
(TIF)

S3 Fig. Schematic representation of the calculation for the verticality of a gene (Vg) on the base of one tree with 30 genomes spanning four phyla. Each phylum is indicated by one color as depicted in the legend of the table. If the phylum is monophyletic in the tree, the number of genomes in the tree are divided by the number of genomes of this phylum present in the dataset of 5,655 organisms–phyla **e** and **f** in the panels **a.** and **b.** of the figure. If the phylum is paraphyletic, the verticality is set to '0'–phyla **g** and **h** in panels **c.** and **d.** of the figure. This number represents the verticality for each phylum. The sum of all verticality scores for the phyla in the tree is then the verticality for the tree and conversely, for the gene.
(TIF)

S4 Fig. Likelihood tests of eukaryote monophyly. The Kishino-Hasegawa (KH) test, Shimodaira-Hasegawa (SH) test and the Approximately-Unbiased (AU) test were performed for 475 prokaryote-eukaryote genes for which eukaryotes were not recovered monophyletic in the ML trees. The histogram shows the distribution of p-values (horizontal axis) for the tests of the unconstrained ML trees against ML trees with constrained eukaryote monophyly. A test was considered significant (eukaryote monophyly was rejected) if p-value ≤ 0.05.
(TIF)

S5 Fig. EPCs with pure sister taxon mapped to conserved clusters. Mapping of EPCs to prokaryotic clusters. The EPCs were separated according to the pure sister group of eukaryotes in the trees and plotted in the same way as in Fig 4 of the main text. The left panel shows EPCs that may include all eukaryotic supergroups, the right panel shows only EPCs that include archaeplastidal eukaryotes. Meaning the latter are indicative of plastid endosymbiosis. For a better overview a headline is included in each plot that lists the taxonomic group represented, if it shows EPCs linked to the mitochondrial ('P and O', left panel) or to the plastidal endosymbiosis event ('Plant only', right panel), and the number of EPCs that are shown as red dots.
(GZ)

Acknowledgments

We thank the central computing unit, ZIM, at the University of Düsseldorf for providing the computational platform for these analyses.

Author Contributions

Conceptualization: Falk S. P. Nagies, Julia Brueckner, William F. Martin.

Data curation: Falk S. P. Nagies, Julia Brueckner.

Formal analysis: Falk S. P. Nagies, Julia Brueckner, Fernando D. K. Tria, William F. Martin.

Funding acquisition: William F. Martin.

Investigation: Falk S. P. Nagies, Julia Brueckner, Fernando D. K. Tria, William F. Martin.

Methodology: Falk S. P. Nagies, Julia Brueckner, Fernando D. K. Tria, William F. Martin.

Project administration: William F. Martin.

Resources: Falk S. P. Nagies, Julia Brueckner.

Supervision: William F. Martin.

Validation: Falk S. P. Nagies, Julia Brueckner, Fernando D. K. Tria, William F. Martin.

Visualization: Falk S. P. Nagies, Julia Brueckner.

Writing – original draft: William F. Martin.

Writing – review & editing: Falk S. P. Nagies, Julia Brueckner, Fernando D. K. Tria, William F. Martin.

References

1. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. High frequency of horizontal gene transfer in the oceans. Science 2010; 330(6000):50. https://doi.org/10.1126/science.1192243 PMID: 20929803

2. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000; 405(6784):299–304. https://doi.org/10.1038/35012500 PMID: 10830951

3. Popa O, Dagan T. Trends and barriers to lateral gene transfer in prokaryotes. Curr Opin Microbiol 2011; 14(5):615–623. https://doi.org/10.1016/j.mib.2011.07.027 PMID: 21856213

4. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, et al. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J Bacteriol 2008; 190(20):6881–6893. https://doi.org/10.1128/JB.00619-08 PMID: 18676672

5. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol 2010; 60(4):708–720. https://doi.org/10.1007/s00248-010-9717-3 PMID: 20623278

6. Hansmann S, Martin W. 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 2000; 50 Pt 4:1655–1663 https://doi.org/10.1099/00207713-50-4-1655 PMID: 10939673

7. Charlebois RL, Doolittle WF. Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res 2004; 14(12):2469–2477. https://doi.org/10.1101/gr.3024704 PMID: 15574825

8. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science 2006; 311(5765):1283–1287. https://doi.org/10.1126/science.1123061 PMID: 16513982

9. Dagan T, Martin W. The tree of one percent. Genome Biol 2006; 7(10):118. https://doi.org/10.1186/gb-2006-7-10-118 PMID: 17081279

10. Koonin EV, Wolf YI, Puigbò P. The phylogenetic forest and the quest for the elusive tree of life. Cold Spring Harb Symp Quant Biol 2009; 74:205–213. https://doi.org/10.1101/sqb.2009.74.006 PMID: 19687142

11. Dagan T, Artzy-Randrup Y, Martin W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A 2008; 105(29):10039–10044. https://doi.org/10.1073/pnas.0800679105 PMID: 18632554

12. Ku C, Martin WF. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70% rule. BMC Biol 2016; 14(1):89. https://doi.org/10.1186/s12915-016-0315-9 PMID: 27751184

13. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. Genome-wide experimental determination of barriers to horizontal gene transfer. Science 2007; 318(5855):1449–1452. https://doi.org/10.1126/science.1147112 PMID: 17947550

14. Pál C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 2005; 37(12):1372–1375. https://doi.org/10.1038/ng1686 PMID: 16311593

15. Lercher MJ, Pál C. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol 2008; 25(3):559–567. https://doi.org/10.1093/molbev/msm283 PMID: 18158322

16. Chen W-H, Trachana K, Lercher MJ, Bork P. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Mol Biol Evol 2012; 29(7):1703–1706. https://doi.org/10.1093/molbev/mss014 PMID: 22319151

17. Dilthey A, Lercher MJ. Horizontally transferred genes cluster spatially and metabolically. Biol Direct 2015; 10:72. https://doi.org/10.1186/s13062-015-0102-5 PMID: 26690249

18. Grassi L, Caselle M, Lercher MJ, Lagomarsino MC. Horizontal gene transfers as metagenomic gene duplications. Mol Biosyst 2012; 8(3):790–795. https://doi.org/10.1039/c2mb05330f PMID: 22218456

19. Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, et al. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc Natl Acad Sci U S A 2012; 109(50):20537–20542. https://doi.org/10.1073/pnas.1209119109 PMID: 23184964

20. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science 2007; 315(5819):1709–1712. https://doi.org/10.1126/science.1138140 PMID: 17379808

21. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. Proc Natl Acad Sci U S A 2015; 112(27):E3574–E3581. https://doi.org/10.1073/pnas.1501049112 PMID: 26100894

22. Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. The ecology and evolution of pangenomes. Curr Biol 2019; 29(20):R1094–R1103. https://doi.org/10.1016/j.cub.2019.08.012 PMID: 31639358

23. Croll D, McDonald BA. The accessory genome as a cradle for adaptive evolution in pathogens. PLoS Pathog 2012; 8(4):e1002608. https://doi.org/10.1371/journal.ppat.1002608 PMID: 22570606

24. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol 2017; 2:17040. https://doi.org/10.1038/nmicrobiol.2017.40 PMID: 28350002

25. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol 2015; 23:148–154. https://doi.org/10.1016/j.mib.2014.11.016 PMID: 25483351

26. Chatton E. *Pansporella perplexa*. Amoebien a spores protégées parasite des daphnies. Réflexions sur la biologie et la phylogénie des protozoaires. Ann Sci Nat Zool 1925; 8:5–85.

27. Creevey CJ, Fitzpatrick DA, Philip GK, Kinsella RJ, O'Connell MJ, Pentony MM, et al. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proc Biol Sci 2004; 271(1557):2551–2558. https://doi.org/10.1098/rspb.2004.2864 PMID: 15615680

28. Semple C, Steel MA. Phylogenetics. Reprinted. Oxford: Oxford Univ. Press; 2009. (Oxford lecture series in mathematics and its applications; vol 24).

29. McPherson RA. The Numbers Universe: An outline of the dirac/eddington numbers as scaling factors for fractal, black hole universes. Electronic Journal of Theoretical Physics 2008; 5(18).

30. Nakamura Y, Itoh T, Matsuda H, Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet 2004; 36(7):760–766. https://doi.org/10.1038/ng1381 PMID: 15208628

31. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987; 4(4):406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454 PMID: 3447015

32. Landan G, Graur D. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol 2007; 24(6):1380–1383. https://doi.org/10.1093/molbev/msm060 PMID: 17387100

33. Criscuolo A. morePhyML: improving the phylogenetic tree space exploration with PhyML 3. Mol Phylogenet Evol 2011; 61(3):944–948. https://doi.org/10.1016/j.ympev.2011.08.029 PMID: 21925283

34. Treangen TJ, Rocha EPC. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet 2011; 7(1):e1001284. https://doi.org/10.1371/journal.pgen.1001284 PMID: 21298028

35. Tria FDK, Brückner J, Skejo J, Xavier JC, Zimorski V, Gould SB, et al. Gene duplications trace mitochondria to the onset of eukaryote complexity; 2019. (vol 176) bioRxiv. https://doi.org/10.1101/781211

36. Szöllősi GJ, Davín AA, Tannier E, Daubin V, Boussau B. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. Philos Trans R Soc Lond B, Biol Sci 2015; 370(1678):20140335. https://doi.org/10.1098/rstb.2014.0335 PMID: 26323765

37. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 1999; 96(7):3801–3806. https://doi.org/10.1073/pnas.96.7.3801 PMID: 10097118

38. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2016; 2(1):vew007. https://doi.org/10.1093/ve/vew007 PMID: 27774300

39. Niehus R, Mitri S, Fletcher AG, Foster KR. Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nat Commun 2015; 6:8924. https://doi.org/10.1038/ncomms9924 PMID: 26592443

40. Nei M. Molecular evolutionary genetics. New York: Columbia University Press; 1987.

41. Aziz RK, Breitbart M, Edwards RA. Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Res 2010; 38(13):4207–4217. https://doi.org/10.1093/nar/gkq140 PMID: 20215432

42. Nevers P, Saedler H. Transposable genetic elements as agents of gene instability and chromosomal rearrangements. Nature 1977; 268(5616):109–115. https://doi.org/10.1038/268109a0 PMID: 339095

43. 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. Pl Syst Evol 1997; 206(1–4):337–351.

44. 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(6681):162–165. https://doi.org/10.1038/30234 PMID: 11560168

45. Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, et al. Isolation of an archaeon at the prokaryote-eukaryote interface. Nature 2020; 577(7791):519–525. https://doi.org/10.1038/s41586-019-1916-6 PMID: 31942073

46. Fan L, Wu D, Goremykin V, Xiao J, Xu Y, Garg S, et al. Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within alphaproteobacteria. Nat Ecol Evol 2020. https://doi.org/10.1038/s41559-020-1239-x PMID: 32661403

47. 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. https://doi.org/10.1038/387493a0 PMID: 9168110

48. Tian R-M, Cai L, Zhang W-P, Cao H-L, Qian P-Y. Rare Events of Intragenus and Intraspecies Horizontal Transfer of the 16S rRNA Gene. Genome Biol Evol 2015; 7(8):2310–2320. https://doi.org/10.1093/gbe/evv143 PMID: 26220935

49. Schönheit P, Buckel W, Martin WF. On the origin of heterotrophy. Trends Microbiol 2016; 24(1):12–25. https://doi.org/10.1016/j.tim.2015.10.003 PMID: 26578093

50. Husnik F, Keeling PJ. The fate of obligate endosymbionts: reduction, integration, or extinction. Curr Opin Genet Dev 2019; 58–59:1–8. https://doi.org/10.1016/j.gde.2019.07.014 PMID: 31470232

51. Tamames J, Gil R, Latorre A, Peretó J, Silva FJ, Moya A. The frontier between cell and organelle: genome analysis of *Candidatus Carsonella ruddii*. BMC Evol Biol 2007; 7:181. https://doi.org/10.1186/1471-2148-7-181 PMID: 17908294

52. Podar M, Anderson I, Makarova KS, Elkins JG, Ivanova N, Wall MA, et al. A genomic analysis of the archaeal system *Ignicoccus hospitalis-Nanoarchaeum equitans*. Genome Biol 2008; 9(11):R158. https://doi.org/10.1186/gb-2008-9-11-r158 PMID: 19000309

53. Anderson I, Djao ODN, Misra M, Chertkov O, Nolan M, Lucas S, et al. Complete genome sequence of *Methanothermus fervidus* type strain (V24S). Stand Genomic Sci 2010; 3(3):315–324. https://doi.org/10.4056/sigs.1283367 PMID: 21304736

54. Gabaldón T. Relative timing of mitochondrial endosymbiosis and the "pre-mitochondrial symbioses" hypothesis. IUBMB Life 2018; 70(12):1188–1196. https://doi.org/10.1002/iub.1950 PMID: 30358047

55. Kapust N, Nelson-Sathi S, Schönfeld B, Hazkani-Covo E, Bryant D, Lockhart PJ, et al. Failure to recover major events of gene flux in real biological data due to method misapplication. Genome Biol Evol 2018; 10(5):1198–1209. https://doi.org/10.1093/gbe/evy080 PMID: 29718211

56. 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 U S A 2002; 99(19):12246–12251. https://doi.org/10.1073/pnas.182432999 PMID: 12218172

57. 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 U S A 2015; 112(33):10139–10146. https://doi.org/10.1073/pnas.1421385112 PMID: 25733873

58. Martin WF, Garg S, Zimorski V. Endosymbiotic theories for eukaryote origin. Philos Trans R Soc Lond B, Biol Sci 2015; 370(1678):20140330. https://doi.org/10.1098/rstb.2014.0330 PMID: 26323761

59. Hittinger CT, Carroll SB. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 2007; 449(7163):677–681. https://doi.org/10.1038/nature06151 PMID: 17928853

60. van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet 2009; 10(10):725–732. https://doi.org/10.1038/nrg2600 PMID: 19652647

61. Maier U-G, Zauner S, Woehle C, Bolte K, Hempel F, Allen JF, et al. Massively convergent evolution for ribosomal protein gene content in plastid and mitochondrial genomes. Genome Biol Evol 2013; 5 (12):2318–2329. https://doi.org/10.1093/gbe/evt181 PMID: 24259312

62. Allen JF, Martin WF. Why have organelles retained genomes? Cell Syst 2016; 2(2):70–72. https://doi.org/10.1016/j.cels.2016.02.007 PMID: 27135161

63. Vos M, Hesselman MC, Te Beek TA, van Passel MWJ, Eyre-Walker A. Rates of lateral gene transfer in prokaryotes: High but why? Trends Microbiol 2015; 23(10):598–605. https://doi.org/10.1016/j.tim.2015.07.006 PMID: 26433693

64. Sela I, Wolf YI, Koonin EV. Theory of prokaryotic genome evolution. Proc Natl Acad Sci U S A 2016; 113(41):11399–11407. https://doi.org/10.1073/pnas.1614083113 PMID: 27702904

65. Martin W. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. Bioessays 1999; 21(2):99–104. https://doi.org/10.1002/(SICI)1521-1878(199902)21:2<99::AID-BIES3>3.0.CO;2-B PMID: 10193183

66. Puigbò P, Wolf YI, Koonin EV. Genome-wide comparative analysis of phylogenetic trees: The prokaryotic forest of life. Methods Mol Biol 2019; 1910:241–269. https://doi.org/10.1007/978-1-4939-9074-0_8 PMID: 31278667

67. Wright ES, Baum DA. Exclusivity offers a sound yet practical species criterion for bacteria despite abundant gene flow. BMC Genomics 2018; 19(1):724. https://doi.org/10.1186/s12864-018-5099-6 PMID: 30285620

68. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016; 44(D1):D733–D745 https://doi.org/10.1093/nar/gkv1189 PMID: 26553804

69. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology 1990; 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

70. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;(16):276–277. https://doi.org/10.1016/s0168-9525(00)02024-2 PMID: 10827456

71. Enright AJ, van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002; 30(7):1575–1584. https://doi.org/10.1093/nar/30.7.1575 PMID: 11917018

72. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013; 30(4):772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

73. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014; 30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

74. Tria FDK, Landan G, Dagan T. Phylogenetic rooting using minimal ancestor deviation. Nat Ecol Evol 2017; 1:193. https://doi.org/10.1038/s41559-017-0193 PMID: 29388565

75. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 2015; 32(1):268–274. https://doi.org/10.1093/molbev/msu300 PMID: 25371430

76. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 2010; 26(13):1669–1670. https://doi.org/10.1093/bioinformatics/btq243 PMID: 20472542

77. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol 2016; 33(6):1635–1638. https://doi.org/10.1093/molbev/msw046 PMID: 26921390

78. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016; 44(D1):D457–D462 https://doi.org/10.1093/nar/gkv1070 PMID: 26476454

79. Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. Journal of molecular evolution 1989; 29(2):170–9. https://doi.org/10.1007/BF02100115 PMID: 2509717

80. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 1999; 16(8):1114–1116 https://doi.org/10.1093/oxfordjournals.molbev.a026201

81. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Systematic biology 2002; 51(3):492–508. https://doi.org/10.1080/10635150290069913 PMID: 12079646

82. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE 2010; 5(3):e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

83. Havlicek LL, Peterson NL. Robustness of the pearson correlation against violations of assumptions. Percept Mot Skills 1976; 43(3_suppl):1319–1334 https://doi.org/10.2466/pms.1976.43.3f.1319

3.3 Gene duplications trace mitochondria to the onset of eukaryote complexity

Year:	2021
Authors:	Fernando D. K. Tria*, Julia Brueckner*, Josip Skejo, Joana C. Xavier, Nils Kapust, Michael Knopp, Jessica L. E. Wimmer, Falk S. P. Nagies, Verena Zimorski, Sven B. Gould, Sriram G. Garg, William F. Martin
Published in:	Genome Biol Evol
Contribution:	Shared first author Major: Collection and analysis of data, figure design and illustration. With the other first author and last author: Manuscript writing and editing.

*These authors contributed equally to this work

Gene Duplications Trace Mitochondria to the Onset of Eukaryote Complexity

Fernando D.K. Tria,*,†,1 Julia Brueckner,†,1 Josip Skejo,1,2 Joana C. Xavier,1 Nils Kapust,1 Michael Knopp,1 Jessica L.E. Wimmer,1 Falk S.P. Nagies,1 Verena Zimorski,1 Sven B. Gould,1 Sriram G. Garg,1 and William F. Martin1

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

[2]Faculty of Science, University of Zagreb, Croatia

†These authors contributed equally to this work.

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

Accepted: 14 March 2021

Abstract

The last eukaryote common ancestor (LECA) possessed mitochondria and all key traits that make eukaryotic cells more complex than their prokaryotic ancestors, yet the timing of mitochondrial acquisition and the role of mitochondria in the origin of eukaryote complexity remain debated. Here, we report evidence from gene duplications in LECA indicating an early origin of mitochondria. Among 163,545 duplications in 24,571 gene trees spanning 150 sequenced eukaryotic genomes, we identify 713 gene duplication events that occurred in LECA. LECA's bacterial-derived genes include numerous mitochondrial functions and were duplicated significantly more often than archaeal-derived and eukaryote-specific genes. The surplus of bacterial-derived duplications in LECA most likely reflects the serial copying of genes from the mitochondrial endosymbiont to the archaeal host's chromosomes. Clustering, phylogenies and likelihood ratio tests for 22.4 million genes from 5,655 prokaryotic and 150 eukaryotic genomes reveal no evidence for lineage-specific gene acquisitions in eukaryotes, except from the plastid in the plant lineage. That finding, and the functions of bacterial genes duplicated in LECA, suggests that the bacterial genes in eukaryotes are acquisitions from the mitochondrion, followed by vertical gene evolution and differential loss across eukaryotic lineages, flanked by concomitant lateral gene transfer among prokaryotes. Overall, the data indicate that recurrent gene transfer via the copying of genes from a resident mitochondrial endosymbiont to archaeal host chromosomes preceded the onset of eukaryotic cellular complexity, favoring mitochondria-early over mitochondria-late hypotheses for eukaryote origin.

Key words: evolution, paralogy, gene transfer, endosymbiosis, gene duplication, eukaryote origin.

Significance

The origin of eukaryotes is one of evolution's classic unresolved issues. At the center of debate is the relative timing of two canonical eukaryotic traits: cellular complexity and mitochondria. Gene duplications fostered the evolution of novel eukaryotic traits and serve as a rich phylogenetic resource to address the question. By investigating gene duplications that trace to the last eukaryotic common ancestor we found evidence for mitochondria preceding cellular complexity in eukaryote evolution. Our results demonstrate that gene duplications were already rampant in the last eukaryote common ancestor, and we propose that the vast majority of duplications resulted from cumulative rounds of gene transfers from the mitochondrial ancestor to the genome of the archaeal host cell.

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

Genome Biol. Evol. 13(5) doi:10.1093/gbe/evab055 Advance Access publication 19 March 2021

Introduction

The last eukaryote common ancestor (LECA) lived about 1.6 Ba (Betts et al. 2018; Javaux and Lepot 2018). It possessed bacterial lipids, nuclei, sex, an endomembrane system, mitochondria, and all other key traits that make eukaryotic cells more complex than their prokaryotic ancestors (Speijer et al. 2015; Gould et al. 2016; Zachar and Szathmáry 2017; Barlow et al. 2018; Betts et al. 2018). The closest known relatives of the host lineage that acquired the mitochondrion are, however, small obligately symbiotic archaea from enrichment cultures that lack any semblance of eukaryotic cell complexity (Imachi et al. 2020). This steep evolutionary grade separating prokaryotes from eukaryotes increasingly implicates mitochondrial symbiosis at eukaryote origin (Gould et al. 2016; Imachi et al. 2020). Yet despite the availability of thousands of genome sequences, and five decades to ponder Margulis (Margulis et al. 2006) resurrection of endosymbiotic theory (Mereschkowsky 1910; Wallin 1925), the timing, and evolutionary significance of mitochondrial origin remains a polarized debate. Gradualist theories contend that eukaryotes arose from archaea by slow accumulation of eukaryotic traits (Cavalier-Smith 2002; Booth and Doolittle 2015; Hampl et al. 2019) with mitochondria arriving late (Pittis and Gabaldón 2016), whereas symbiotic theories have it that mitochondria initiated the onset of eukaryote complexity in a nonnucleated archaeal host (Imachi et al. 2020) by gene transfers from the organelle (Martin and Müller 1998; Lane and Martin 2010; Gould et al. 2016; Martin et al. 2017).

Information from gene duplications can help to resolve this debate. Gene and genome duplications are a genomic proxy for biological complexity and are the hallmark of eukaryotic genome evolution (Ohno 1970). Gene families that were duplicated during the transition from the first eukaryote common ancestor (FECA) to LECA could potentially shed light on the relative timing of mitochondrial acquisition and eukaryote complexity if they could be inferred in a quantitative rather than piecemeal manner. Duplications of individual gene families (Hittinger and Carroll 2007) and whole genomes (Scannell et al. 2006; Van De Peer et al. 2009) have occurred throughout eukaryote evolution. This is in stark contrast to the situation in prokaryotes, where gene duplications are rare at best (Treangen and Rocha 2011) and whole-genome duplications of the kind found in eukaryotes are altogether unknown. In an earlier study, Makarova et al. (2005) used a liberal criterion and attributed any gene present in two major eukaryotic lineages as present in LECA. Their approach overlooks eukaryotic lineage phylogeny, leading to the inference of 4,137 families that might have been duplicated in LECA. More recently, Vosseberg et al. (2021) examined nodes in trees derived from protein domains that could be scored as duplications among the 7,447–21,840 genes that they estimated to have been present in LECA and used branch lengths to estimate the timing of duplication events. However, they did not report integer numbers for duplications because of their approach based on the analyses of very large protein-domain trees instead of discrete protein-coding gene trees. Here, we addressed the problem of which, what kind of, and how many genes were duplicated in LECA and discuss the implications of our findings for the mitochondria-early versus mitochondria-late debate.

Results and Discussion

To ascertain when the process of gene duplication in eukaryote genome evolution commenced and whether mitochondria might have been involved in that process, we inferred all gene duplications among the 1,848,936 protein-coding genes present in 150 sequenced eukaryotic genomes. For this, we first clustered all eukaryotic proteins using a low stringency clustering threshold of 25% global amino acid identity (see Materials and Methods) in order to recover the full spectrum of eukaryotic gene duplications in both highly conserved and poorly conserved gene families. We emphasize that we employed a clustering threshold of 25% amino acid identity because our procedure was designed to allow for the construction of alignments and phylogenetic trees for each cluster. The 25% threshold keeps the alignments and trees out of the "twilight zone" of sequence identity (Jeffroy et al. 2006), where alignment and phylogeny artifacts based on comparisons of nonhomologous amino acid positions arise.

We then identified all genes that were duplicated across 150 sequenced eukaryotic genomes. In principle, genes present only in one copy in any genome could have also undergone duplication, with losses leading to single-copy status. Quantifying duplications in such cases are extremely topology-dependent. We therefore focused our attention on genes for which topology-independent evidence for duplications existed, that is, genes that were present in more than one copy in at least one genome. Eukaryotic gene duplications were found in all six supergroups: Archaeplastida, Opisthokonta, Mycetozoa, Hacrobia, SAR, and Excavata (Adl et al. 2012), whereby 941,268 of all eukaryotic protein-coding genes, or nearly half the total, exist as multiple copies in at least one genome. These are distributed across 239,012 gene families, which we designate as multicopy gene families. However, 89.7% of these gene families harbor only recent gene duplications, restricted to a single eukaryotic genome (inparalogs). The remaining 24,571 families (10.3%) harbor multiple copies in at least two eukaryotic genomes, with variable distribution across the supergroups (fig. 1). Opisthokonts (animals and fungi) together harbor a total of 22,410 multicopy gene families present in at least two genomes. The animal lineage harbors 19,530 multicopy gene families, the largest number of any lineage sampled, followed by the plant lineage (Archaeplastida) with 6,495 multicopy gene families. Of particular importance for the present study, among the 24,571 multicopy gene families, we

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

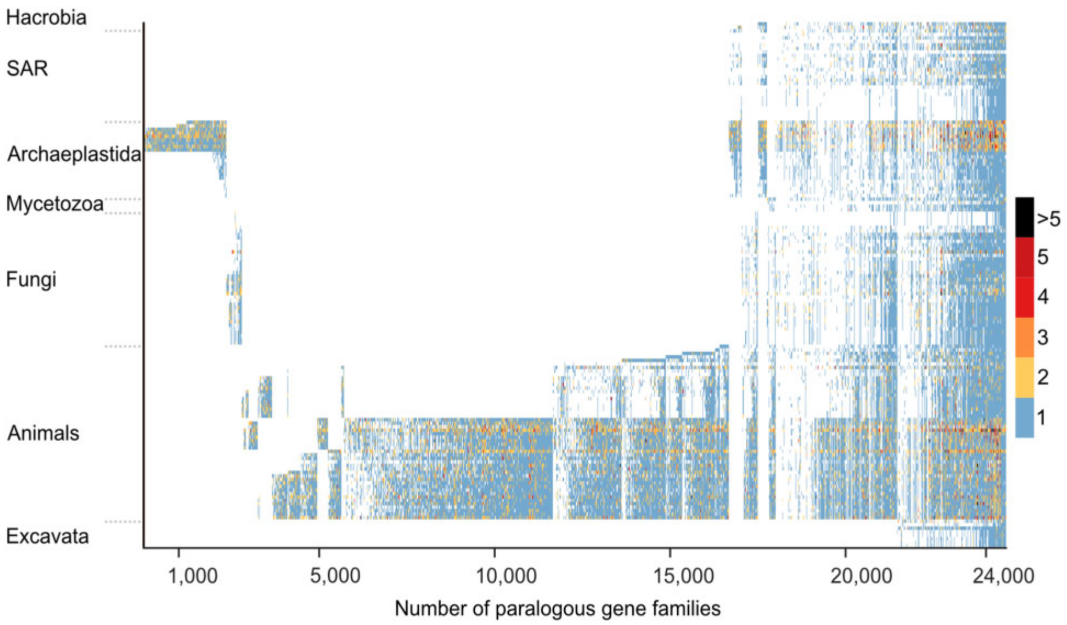


Fig. 1.—Distribution of multicopy genes across 150 eukaryotic genomes. All eukaryotic protein-coding genes were clustered, aligned, and used for phylogenetic inferences. The resulting gene families present as multiple copies in more than one genome are plotted (see Materials and Methods). The figure displays the 24,571 multicopy gene families (horizontal axis) and the colored scale indicates the number of gene copies in each eukaryotic genome (vertical axis). The genomes were sorted according to a reference species tree (supplementary data 7) and taxonomic classifications were taken from NCBI. Animals and fungi together form the opisthokont supergroup.

identified 1,823 that are present as multiple copies in at least one genome from all six supergroups and are thus potential candidates of gene duplications tracing to LECA. In order to distinguish between the possibility of 1) duplications within supergroups after diversification from LECA and 2) duplications giving rise to multiple copies in the genome of LECA, we used phylogenetic trees.

To infer the relative phylogenetic timing of eukaryotic gene duplication events, we focused our attention on the individual protein alignments and maximum-likelihood trees for all 24,571 gene families with paralogs in at least two eukaryotic genomes. We then assigned gene duplications in each tree to the most recent internal node possible, allowing for multiple gene duplication events and losses as needed (see Materials and Methods) and permitting any branching order of supergroups. This approach minimized the number of inferred duplication events and identified a total of 163,545 gene duplications, 160,676 of which generated paralogs within a single supergroup (inparalogs at the supergroup-level). An additional 2,869 gene duplication events trace to the common ancestor of at least two supergroups (fig. 2*a* and supplementary table 1). The most notable result however was the identification of 713 gene duplication events distributed in 475 gene trees that generated paralogs in the genome of LECA before eukaryotic supergroups diverged. For these 475 gene trees, the resulting LECA paralogs are retained in at least one genome from all six supergroups, as indicated in red in figure 2*a*. The sample of 475 genes provides a conservative estimate of genes that duplicated in LECA. Among the 1,823 gene families having multiple copies in members of all six supergroups, note that only in 475 families (26%) do the duplications actually trace to LECA in the trees. These results indicate that most duplications in eukaryotes are lineage specific (figs. 1 and 2), and furthermore raise caveats regarding earlier estimates of duplications in LECA (Makarova et al. 2005; Vosseberg et al. 2021) based on more permissive criteria.

LECA's Duplications Constrain the Position of the Eukaryotic Root

The six supergroups plus LECA at the root represent a seven-taxon tree with the terminal edges bearing 97% of gene duplication events (fig. 2). Gene duplications that map to internal branches of the rooted supergroup tree can result from duplications in LECA followed by vertical inheritance and differential loss in some supergroups, or they result from more recent duplications following the divergence from LECA. Branches that explain the most duplications are likely to reflect the natural supergroup phylogeny, because support for conflicting branches is generated by random nonphylogenetic patterns of independent gene losses (Van De Peer et al. 2009). There is a strong phylogenetic signal contained within the eukaryotic gene duplication data (fig. 2). Among all possible internal branches, those supported by the most frequent

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

duplications are compatible with the tree in figure 2*b*, which places the eukaryotic root on the branch separating Excavates from other supergroups, as implicated in previous studies of concatenated protein sequences (Hampl et al. 2009; He et al. 2014). However, massive gene loss in specific supergroups (in excavates, e.g., see fig. 1) could impair identification of the eukaryotic root (Zmasek and Godzik 2011; Ku et al. 2015; Albalat and Cañestro 2016). Indeed, the high frequency of duplications that trace to LECA readily explains why resolution of deep eukaryotic phylogeny or the position of the eukaryotic root with traditional phylogenomic approaches (Ren et al. 2016) is so difficult (see also supplementary table 2): LECA was replete with duplications and paralogy. Paralogy imposes conflicting signals onto phylogenetic systematics, but gene duplications harbor novel phylogenetic information in their own right (fig. 2), as shared gene duplications discriminate between alternative eukaryote supergroup relationships.

Eukaryotic Duplications Are Not Transferred across Supergroups

Like the nucleus, mitochondria, and other eukaryotic traits (Speijer et al. 2015; Gould et al. 2016; Zachar and Szathmáry 2017; Barlow et al. 2018; Betts et al. 2018; Imachi et al. 2020), the lineage-specific accrual of gene and genome duplications distinguish eukaryotes from prokaryotes (Ohno 1917; Scannell 2006; Hittinger and Carroll 2007; Van De Peer et al. 2009; Treangen and Rocha 2011). Nonetheless, one might argue that the distribution of duplications observed here does not reflect lineage-dependent processes at all, but lateral gene transfers (LGTs) among eukaryotes instead

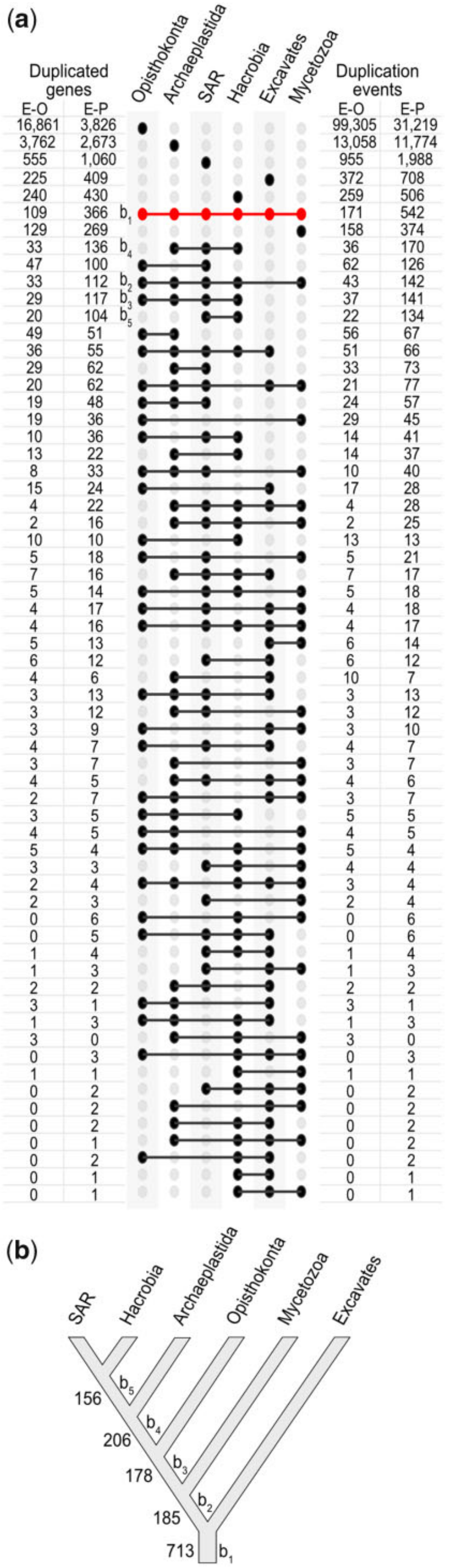


Fig. 2.—Distribution of paralogs descending from gene duplications across six eukaryotic supergroups. (*a*) The figure shows the distribution of paralogs resulting from gene duplications in eukaryotic-specific genes (E-O) and eukaryotic genes with prokaryotic homologs (E-P) (see Materials and Methods for details). Duplicated genes refer to the numbers of gene trees with at least one duplication event with descendant paralogs across the supergroups (filled circles in the center). Number of duplication events refers to the total number of gene duplications. The red row circles indicate gene duplications with descendant paralogs in species from all six supergroups and, thus, tracing to LECA regardless of the eukaryotic phylogeny. An early study assigned 4,137 duplicated gene families to LECA but attributed all copies present in any two major eukaryotic groups to LECA (Makarova et al. 2005). In the present sample, we find 2,869 gene duplication events that trace to the common ancestor of at least two supergroups. Our stringent criterion requiring paralog presence in all six supergroups leaves 713 duplications in 475 gene families in LECA. (*b*) Rooted phylogeny of eukaryotic supergroups that maximizes compatibility with gene duplications. Gene duplications mapping to five edges are shown (b_1, b_2, . . . , b_5). The tree represents almost exactly all edges containing the most duplications, the exception is the branch joining Hacrobia and SAR because the alternative branch joining SAR and Opisthokonta is better supported. However, the resulting subtree ((Opisthokonta, SAR),(Archaeplastida, Hacrobia)) accounts for 249 duplications, fewer than the (Opisthokonta,(Archaeplastida,(SAR, Hacrobia))) subtree shown (262 duplications). The position of the root identifies additional gene duplications tracing to LECA (table 1 and supplementary table 4).

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

(Andersson et al. 2003; Keeling and Palmer 2008; Leger et al. 2018). That is, a duplication could, in theory, originate in one supergroup and one or more gene copies could subsequently be distributed among other supergroups via eukaryote-to-eukaryote LGT. However, were that theoretical possibility true then neither duplications, nor any trait, nor any gene could be traced to LECA because all traits and genes in eukaryotes could, in the extreme, simply reflect 1.6 Byr of lineage-specific invention within one supergroup followed by lateral gene traffic among eukaryotes rather than descent with modification (Andersson et al. 2003; Keeling and Palmer 2008; Leger et al. 2018).

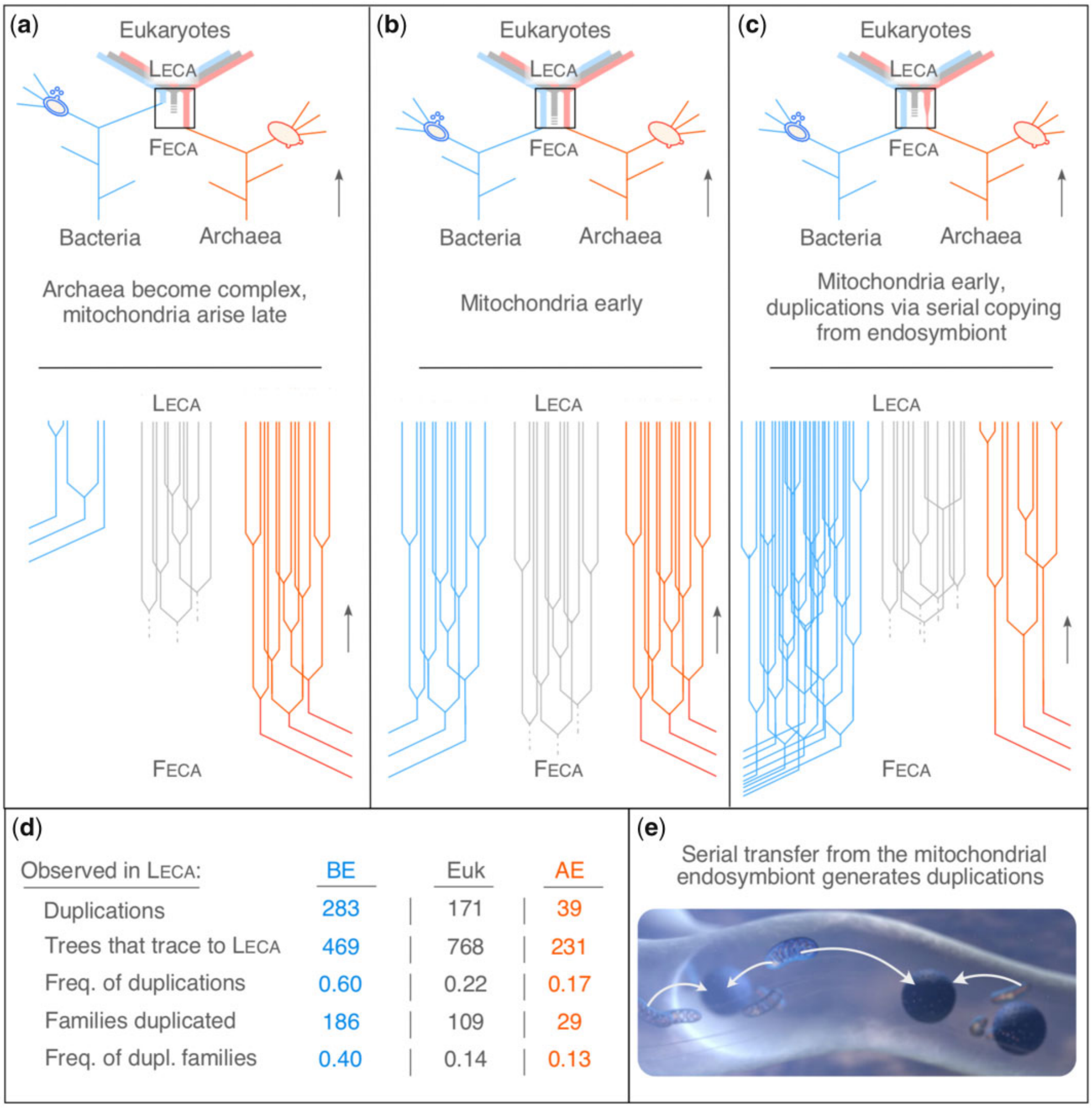
However, the present data themselves exclude the deeply improbable eukaryote-to-eukaryote lateral duplication transfer theory in a subtle but strikingly clear manner. How so? Figures 1 and 2*a* show that 30,439 gene lineages bearing duplications (93% of the total) are restricted in their distribution to "only one supergroup," whereas only 2,245 (7% of the total) are shared among two to five supergroups. That is, only 7% of the duplications are shared across supergroups, hence they are the only possible candidates for LGT among supergroups. For the sake of argument, let us entertain the extreme assumption that *all* 2,245 patterns of shared but nonuniversal duplications involved intersupergroup LGT, recalling that there is no intersupergroup LGT in 93% of the genes (fig. 2 and supplementary table 1). With that generous assumption, the intersupergroup LGT frequency would be maximally 7%. That is an extreme upper bound, though, because the observed 93% frequency for duplicates that are supergroup specific and thus have absolutely no observable intersupergroup LGT should apply equally to the 7% of duplications shared across supergroups. Thus, the more realistic maximum estimate is that 0.49% of duplications (7% of 7%) might have been generated by intersupergroup LGT. This estimate is based solely upon the distribution of the duplicates and the premise that eukaryote supergroups are monophyletic. As it concerns the 475 genes with duplications that trace to LECA (fig. 2 and supplementary table 1), this means that 0.49% out of 475, or about 2.3 genes in our data might have been caused by intersupergroup LGT. That is a very low frequency and is consistent with independent genome-wide phylogenetic tests presented previously (Ku et al. 2015) for the paucity of eukaryote-to-eukaryote LGT. If we count duplication events (fig. 2*a*, right panel) rather than gene lineages (fig. 2*a*, left panel), the picture is even more vertical, because 98% of the events are supergroup-specific, hence lacking any patterns that could reflect LGT, meaning that maximally 0.04% (2% of 2%) or 0.19 duplications among 475 (which rounds to zero genes) could be the result of lateral transfer. The supergroup-specific distributions of duplications themselves thus provide very strong evidence that the distribution of duplicated genes in eukaryotes is not the result of eukaryote-to-eukaryote LGT phenomena (Andersson et al. 2003; Keeling and Palmer 2008; Leger et al. 2018) but the result of vertical evolution within supergroups accompanied by gene birth, death (Nei et al. 1997), and differential gene loss (Ku et al. 2015).

LECA's Duplications Support an Early Mitochondrion

Arguably, the timing of mitochondrial origin is the central so far unresolved issue at the heart of eukaryote origin. Several alternative theories for eukaryogenesis have been proposed (reviewed in Martin et al. 2001; Embley and Martin 2006; Poole and Gribaldo 2014; López-García and Moreira 2015; Eme 2017). Symbiogenic theories posit a causal role for mitochondrial endosymbiosis at the origin of cellular eukaryotic complexity (Lane and Martin 2010) with the host being a garden variety archaeon (Martin and Müller 1998). Gradualist theories posit an autogenous origin of eukaryote cell complexity with little or no contribution of the mitochondrion to eukaryogenesis (Cavalier-Smith 2002; Gray 2014). Intermediate theories posit the existence of endosymbioses prior to the origin of mitochondria. These include an endosymbiotic origin of the nucleus (Lake and Rivera 1994), an endosymbiotic origin of peroxisomes (de Duve 2007), an endosymbiotic origin of flagella (Margulis et al. 2000), the lateral acquisition of the cytoskeleton (Doolittle 1998) or, more liberally, additional symbioses preceding the mitochondrion in unconstrained numbers, as long as each symbiosis "explains the origin of any eukaryotic innovation as a response to an endosymbiotic interaction" (Gabaldón 2018). Most current theories posit an origin of the host from archaea (Martin et al. 2015; Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017; Imachi 2020), though theories for eukaryote origins from actinobacteria (Cavalier-Smith 2002), and planctomycetes (Cavalier-Smith and Chao 2020) are discussed. Notwithstanding such diversity of views, the main divide among theories for eukaryote origin remains the relative timing of mitochondrial origin, that is did the mitochondrion initiate or culminate eukaryote origin (Martin et al. 2001; Embley and Martin 2006; Poole and Gribaldo 2014; López-García and Moreira 2015; Eme et al. 2017)? Alternative theories for eukaryote origin generate distinct predictions about the nature of gene duplications in LECA.

Gradualist theories entailing an archaeal host (Cavalier-Smith 2002; Booth and Doolittle 2015; Pittis and Gabaldón 2016; Hampl et al. 2019) predict genes of archaeal origin and eukaryote-specific genes to have undergone numerous duplications during the origin of eukaryote complexity, prior to the acquisition of the mitochondrion. In that case, the mitochondrion arose late, hence bacterial-derived genes would have accumulated fewer duplications in LECA than archaeal-derived or eukaryote-specific genes (fig. 3*a*). Models invoking gradual lateral gene transfers (LGT) from ingested (phagocytosed) food prokaryotes prior to the origin of mitochondria (Doolittle 1998) also predict more duplications in archaeal-derived and eukaryote-specific genes to underpin the origin

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021


(a)
Eukaryotes
LECA
FECA
Bacteria
Archaea
Archaea become complex, mitochondria arise late
LECA
FECA
(b)
Eukaryotes
LECA
FECA
Bacteria
Archaea
Mitochondria early
LECA
FECA
(c)
Eukaryotes
LECA
FECA
Bacteria
Archaea
Mitochondria early, duplications via serial copying from endosymbiont
LECA
FECA
(d)
(e)
Serial transfer from the mitochondrial endosymbiont generates duplications

Observed in LECA:	BE	Euk	AE
Duplications	283	171	39
Trees that trace to LECA	469	768	231
Freq. of duplications	0.60	0.22	0.17
Families duplicated	186	109	29
Freq. of dupl. families	0.40	0.14	0.13

FIG. 3.—Alternative models for eukaryote origin generate different predictions with respect to duplications. In each panel, gene duplications during the FECA to LECA transition (boxed in upper portion) are enlarged in the lower portion of the panel. (*a*) Cellular complexity and genome expansion in an archaeal host predate the origin of mitochondria. (*b*) Mitochondria enter the eukaryotic lineage early, duplications in mitochondrial-derived, host-derived, and eukaryotic-specific genes occur, genome expansion affects all genes equally. (*c*) Gene transfers from a resident endosymbiont generate duplications in genes of bacterial origin in an archaeal host. (*d*) Observed frequencies from gene duplications that trace to LECA (see supplementary table 1). BE refers to eukaryotic genes with bacterial homologs only; AE refers to eukaryotic genes with archaeal homologs only; and Euk refers to eukaryotic genes without prokaryotic homologs. (e) Schematic representation of serial gene transfers from the mitochondrion (white arrows) to the host's chromosomes.

of phagocytotic feeding, but do not predict duplications specifically among acquired genes (whether from bacterial or archaeal food) because each ingestion contributes genes only once.

By contrast, transfers from the endosymbiotic ancestors of organelles continuously generated gene duplications in the host's chromosomes (Timmis et al. 2004; Allen 2015), a process that continues to the present day in eukaryotic genomes (Timmis et al. 2004; Portugez et al. 2018). Symbiogenic theories posit that the host that acquired the mitochondrion was an archaeon of normal prokaryotic complexity (Martin and Müller 1998; Lane and Martin 2010; Gould et al. 2016; Martin et al. 2017; Imachi et al. 2020) and hence lacked duplications underpinning eukaryote complexity. There are examples known in which bacteria grow in intimate association with archaea (Imachi et al. 2020) and in which

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

prokaryotes become endosymbionts within other prokaryotic cells (Martin et al. 2017). However, there are two different ways in which mitochondria could promote the accumulation of duplications. If energetic constraints (Lane and Martin 2010) were the sole factor permitting genome expansion, duplications would accrue in all genes regardless of their origin, such that gene duplications in the wake of mitochondrial origin should be equally common in genes of bacterial, archaeal, or eukaryote-specific origin, respectively (fig. 3*b*). If, on the other hand, the role of mitochondria in gene duplications was mechanistic rather than purely energetic, genes of mitochondrial origin should preferentially undergo duplication. This is because the mechanism of gene transfers from resident organelles involve endosymbiont lysis and the "copying" (Allen 2015) of organelle genomes to the host's chromosomes followed by recombination and mutation (Portugez et al. 2018). Gene transfers from resident endosymbionts specifically generate duplications of endosymbiont genes because new copies of the same genes are recurrently transferred (Timmis et al. 2004; Allen 2015) (fig. 3*c*).

The duplications in LECA reveal a vast excess of duplications in LECA's bacterial-derived genes relative to archaeal-derived and eukaryote-specific genes (fig. 3*d*). Of all gene families tracing to LECA, 26% experienced at least one duplication event during the transition to LECA from FECA. Notably, the excess proportion of duplicates among genes of bacterial origin is significant as judged by the two-tailed binomial test ($P=1.3\times10^{-10}$; proportion of duplicates at 95% CI=[35–44%]; df=1). On the other hand, genes of archaeal origin show significantly fewer duplicates ($P=8.4\times10^{-7}$; proportion of duplicates 95% CI=[8–17%]; df=1) with the proportion of duplicates being similar to eukaryote-specific genes (fig. 3*d*).

Do Bacterial Genes in LECA Stem from the Mitochondrion?

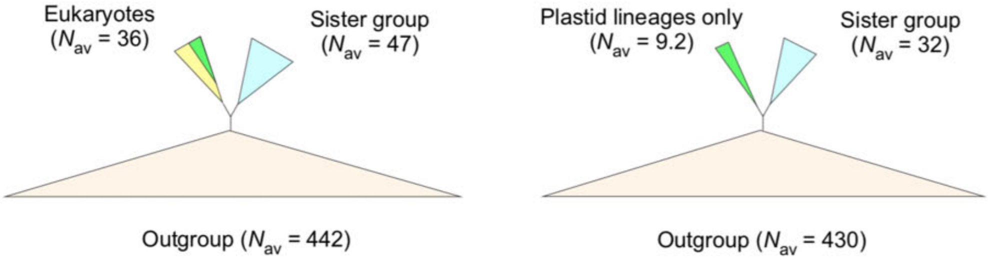
If bacterial genes in LECA stem from the mitochondrion, as opposed to 1) eukaryote-to-eukaryote gene transfers, which were already excluded for >99% of the families with duplications in this data on the basis of their distributions alone, or 2) multiple lineage-specific acquisitions from bacteria via LGT, then the bacterial genes should trace to the eukaryote common ancestor. That is, the eukaryotes should form a monophyletic clade in gene trees that connect prokaryotic and eukaryotic genes. To test this, we generated clusters, alignments, and trees for genes shared by prokaryotes and eukaryotes from 22,471,723 million genes from 5,655 genomes and including 150 eukaryotes (see Materials and Methods). The results from the 2,575 trees that contained at least five prokaryotic and at least two eukaryotic sequences are summarized in figure 4. As with the duplications themselves, eukaryote gene evolution is again vertical. Out of the 2,575 trees only 475 did not recover eukaryotes as monophyletic. However, none of these 475 trees rejected eukaryote monophyly using the Shimodaira–Hasegawa (SH) test (see Materials and Methods) and only 25 trees (1% of the total) rejected eukaryote monophyly using the Kishino–Hasegawa (KH) test. Applying the approximately unbiased (AU) test, only three trees out of 475 rejected eukaryote monophyly. This traces gene origin of ≥1,649 out of the 2,575 genes shared by prokaryotes and eukaryotes to LECA, and the origin of ≤926 genes to the archaeplastidal ancestor because the latter trees contain only photosynthetic eukaryotic lineages (fig. 4*a*).

The 1,649 trees that trace prokaryotic gene origins to LECA fall into two classes with regard to the sister group of the eukaryotic gene: 966 in which the prokaryotic sister group to eukaryotes contained members of only one phylum (a "pure" sister, S_{pure} in fig. 4, 59% of the trees) and those in which the sister to the eukaryotes contained members of more than one phylum (a "mixed" sister, 41% of the trees). The only way to obtain a mixed sister topology of prokaryotic sequences for a eukaryotic gene is via LGT among prokaryotes (Ku and Martin 2016). If we exclude the reality of LGT among prokaryotes, and interpret mixed sister topologies at face value, they would suggest that eukaryotes arose before the diversification of the diverse prokaryotic phyla present in our sample, which would be incompatible with accounts of eukaryote age (Parfrey et al. 2011; Betts et al. 2018), and would furthermore have LECA arising at different times, depending on the membership in the sister group. LGT among the prokaryotic reference sequences in the mixed sister cases (Ku and Martin 2016; Nagies et al. 2020) is clearly the simpler explanation. The pure sister was bacterial in 49% of the trees and archaeal in only 9.5% of the trees. Only in 115 trees (7.0%) was the bacterial pure sister clade alphaproteobacterial. These 115 trees are readily explained because they stem from the mitochondrion, even though the alphaproteobacterial-derived genes in eukaryotes do not all reside in the "same" alphaproteobacterial genome as previously observed (Ku et al. 2015; Nagies et al. 2020), requiring LGT among alphaproteobacteria, at least, to account for the topology. Yet, the crucial and previously underinvestigated issue concerns the remaining 695 pure sister bacterial origin cases (86%) that trace to LECA but reside in a genome that does not carry an alphaproteobacterial taxon label (fig. 4), as recently set forth in a study that examined the phylogeny of only the more conserved fraction of genes shared by prokaryotes and eukaryotes (Nagies et al. 2020).

There are two general ways to explain the 86% of non-alphaproteobacterial genes that trace to LECA. The first is to take one specific aspect of the trees—namely, the taxon label of the sister group—at face value and interpret the data as evidence for independent individual contributions to eukaryotes (via LGT or via multiple resident symbionts) by all of the bacterial phyla in the sample. At the level of the taxa listed in figure 4, that would mean 26 different bacterial donors to LECA in addition to the alphaproteobacterial contribution,

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

(a)



(b)

	1649 Trees also including plastid-bearing lineages						926 Trees only including plastid-bearing lineages						
Taxon	P_{mono}	S_{non}	S_{mix}	S_{pure}	$S_{p,avg}$	N_{trees}	P_{mono}	S_{non}	S_{mix}	S_{pure}	$S_{p,avg}$	N_{trees}	ID_{gen}
Bacteria													
Acidithiobacillia	0.93	125	11	1	6	137	0.96	75	6	1	1	82	5
Chlorobi	0.80	125	12	0	-	137	0.83	93	8	1	1	102	12
Chlamydiae	0.72	129	15	10	16.5	154	0.75	59	11	10	23.8	80	113
Deinococcus-Thermus	0.65	269	31	12	4.3	312	0.64	128	18	5	8.4	151	26
Synergistetes	0.64	129	18	1	1	148	0.82	72	4	1	1	77	5
Thermotogae	0.64	156	22	2	1	180	0.68	80	9	4	1.3	93	32
Fusobacteria	0.57	166	21	2	11.5	189	0.73	75	6	1	4	82	20
Negativicutes	0.56	204	18	4	1.5	226	0.60	94	8	3	1.7	105	12
Epsilonproteobacteria	0.56	178	27	7	2.4	212	0.70	104	11	4	1.8	119	253
Cyanobacteria	0.55	312	59	24	21.4	395	0.61	173	34	187	20.9	394	94
Tenericutes	0.53	124	17	6	2.2	147	0.35	53	8	4	5.5	65	141
Verrucomicrobia	0.52	212	48	12	1.4	272	0.54	121	18	4	2	143	10
Planctomycetes	0.51	249	46	16	1.5	311	0.59	118	22	13	1.8	153	8
Acidobacteria	0.51	274	45	10	1.2	329	0.60	130	24	4	1	158	8
Nitrospirae	0.50	142	18	1	1	161	0.40	92	8	0	-	100	8
Chloroflexi	0.41	249	45	9	1.1	303	0.45	138	17	9	3.1	164	26
Spirochaetes	0.39	312	58	12	3.2	382	0.47	144	24	12	6.8	180	66
Aquificae	0.39	121	16	1	3	138	0.39	84	6	2	3	92	16
Erysipelotrichia	0.38	144	16	2	1	162	0.31	68	3	0	-	71	8
Bacilli	0.36	439	63	49	8.1	551	0.35	195	30	15	5.5	240	1012
Deltaproteobacteria	0.33	435	149	62	2.8	646	0.38	209	64	44	2.5	317	74
Bacteroidetes	0.33	434	76	66	3.6	576	0.38	186	34	29	3.2	249	185
Actinobacteria	0.33	550	118	157	5.2	825	0.44	289	42	43	9.1	374	593
Clostridia	0.31	364	71	33	2.2	468	0.35	172	27	13	2.7	212	169
Alphaproteobacteria	0.27	515	117	115	31.9	747	0.31	235	48	72	4.5	355	460
Betaproteobacteria	0.27	503	127	49	4.3	679	0.34	246	55	19	2.5	320	493
Gammaproteobacteria	0.25	560	126	147	40.4	833	0.30	273	67	55	23	395	1565
Other bacteria	-	-	-	-	-	420	-	-	-	-	-	210	29
Archaea													
Methanococcales	0.90	190	9	8	4.5	207	0.82	50	7	3	7.7	60	16
Thermococcales	0.88	231	17	11	13.9	259	0.81	43	3	2	7.5	48	27
Methanobacteriales	0.84	218	17	17	7.9	252	0.77	53	4	5	2.6	62	17
Archaeoglobales	0.84	212	24	7	5.3	243	0.78	55	7	6	4.7	68	8
Sulfolobales	0.84	265	6	6	9.5	277	0.87	49	2	1	5	52	29
Methanomicrobiales	0.80	221	23	15	2.7	259	0.67	59	10	13	3.8	82	9
Methanosarcinales	0.48	260	14	29	5.4	303	0.59	75	5	21	7	101	32
Thermoproteales	0.44	223	49	15	3.9	287	0.41	45	5	4	2	54	10
Thermoplasmatales	0.36	220	37	13	2.5	270	0.57	47	18	4	2.5	69	5
Natrialbales	0.28	271	60	13	2	344	0.33	84	15	8	1.8	107	9
Desulfurococcales	0.25	196	45	7	1.9	248	0.35	44	6	5	1.6	55	11
Haloferacales	0.24	259	55	5	2.2	319	0.22	78	13	7	2	98	7
Halobacteriales	0.16	312	28	10	1.2	350	0.18	97	4	8	1.6	109	16
Other archaea	-	-	-	-	-	389	-	-	-	-	-	134	16

Fig. 4.—Identification of prokaryotic sisters in 2,575 eukaryotic–prokaryotic gene trees. (*a*) The individual trees were rooted on the branch leading to the largest prokaryotic clade deriving the sister group to eukaryotes. The average number of sequences in the eukaryotic clade, sister group, and outgroup are indicated. (*b*) The list of bacterial (top) and archaeal (bottom) phyla occurring in the trees exclusive to plant lineages (right) and all other trees (left). Archaeal and bacterial phyla with less than five representative species in the data set were collapsed into "other archaea" and "other bacterial" groups. P_{mono} refers the proportion of trees with a branch (split) separating the species of the phylum from the others; S_{non} refers to the number of occurrence of the phylum only in the outgroup clade; S_{mix} refers to the number of occurrences of the phylum as a mixed sister (more than one phylum in the clade); S_{pure} refers to the number of occurrences of the phylum as pure sister (as the single phylum); $S_{p.avg}$ shows the average size of the sister group when the phylum occurs as a pure sister clade. N_{trees} show the number of occurrences of the phyla across all trees. ID_{gen} refers to the total number of species in each phylum.

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

Table 1

Functional Categories of Genes Duplicated in LECAa

Category[b]	(*n*)	Bacterial	Archaeal	Universal	Eukaryotic
Metabolism	(141)	64	2	58	17
Protein modification, folding, degradation	(89)	30	8	30	21
Ubiquitination		3	1	—	9
Proteases		9	1	7	1
Kinase/phosphatase/modification		12	6	19	9
Folding		6	—	4	2
Novel eukaryotic traits	(61)	8	4	12	37
Cell cycle		1	1	2	5
Cytoskeleton		4	—	1	19
Endomembrane (ER; Golgi; vesicles)		2	2	8	10
mRNA splicing		1	1	1	3
Mitochondrion	(47)	29	—	9	9
Carbon metabolism	(37)	26	—	11	—
Glycolysis		10	—	5	—
Reserve polysaccharides, other		16	—	6	—
Cytosolic translation	(36)	15	7	10	4
Nucleic acids	(55)	13	7	15	20
Histones		—	—	2	8
RNA		8	3	6	4
DNA		5	4	7	8
Membranes (excluding endomembrane)	(46)	18	1	12	15
Transporters, plasma associated		8	1	9	14
Lipid synthesis		10	—	3	1
Redox	(15)	11	—	4	—
Hypothetical	(229)	81	9	61	78
Total		295	38	222	201

Note.—*n*, number of duplicated genes in the corresponding category.

[a]About 475 genes duplicated in LECA and present in all six supergroups plus 281 genes with duplications tracing to the common ancestors of excavates and other supergroups. The annotation, source (bacterial, archaeal, present in bacteria and archaea, eukaryote specific), and the numbers of duplications for each cluster are given in supplementary tables 3 and 4. All categories listed had representatives on both the 475 and the 281 list except mRNA splicing, present in the 475 list only.

[b]The categories do not strictly adhere to KEGG or gene ontology classifications, instead they were chosen to reflect the processes that took place during the FECA to LECA transition. The largest number of duplications in LECA for any individual gene was 12, a dynein chain known from previous studies to have undergone duplications in the common ancestor of plants animals and fungi (Kollmar 2016).

and donations from 13 different archaeal host taxa. With 39 donor phyla, LECA already looks like a grab bag of genes. At the level of genus, the taxon labels of the trees would mean 794 different bacterial donors to LECA under permissive models (Gabaldón 2018), followed by a particularly ad hoc sudden stop of gene influx to eukaryotes after the FECA to LECA transition, because the eukaryotes are monophyletic in these trees. The suggestion of symbiont acquisition and gene transfers without constraints (Gabaldón 2018) carries a hidden and seldom spelled out corollary (Martin 1999). Namely, it entails the strict condition that all of the nonalphaproteobacterial bacterial genes in question not only resided in the genome of members of the 27 different phylum level bacterial taxa at the time of donation to LECA (fig. 4) but furthermore, and crucially, that those genes evolved "vertically" within the chromosomal confines of those respective phyla during the 1.6 Byr since eukaryotes arose. Such unrestricted donor theories (Gabaldón 2018) assume that the present-day phylum taxon label on the gene accurately identifies the donor phylum at the time of transfer. But that is true "if and only if" the gene has been vertically inherited within that phylum (no interphylum LGT) since its donation to LECA (Martin 1999; Esser et al. 2007).

Such theories of unrestricted LGT to eukaryotes with strictly vertical gene evolution among prokaryotes are unlikely and resoundingly rejected by the data. If we look beyond the mere taxon label of the sister group (fig. 4), we see that the putative 27 bacterial donor lineages themselves do not evolve in a vertical manner. The average level of monophyly for bacterial phyla in the 1,649 trees that trace to LUCA is 47% (P_{mono} in fig. 4). Alphaprotebacteria were monophyletic in only 27% of the trees in which they occurred, as were generalists with large genomes such as betaproteobacteria (27%) and actinobacteria (33%). Specialists like chlorobi or chlamydia with more restricted pangenomes were more monophyletic (80% and 72%, respectively). Halophilic archaea, which are known to have acquired many genes from bacteria (Nelson-Sathi et al. 2012), are the least monophyletic prokaryotes

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

sampled (halobacteriales, 16%, fig. 4). For the 926 genes that, based on their distribution, trace to the archaeplastidal common ancestor (fig. 4, right panel), the bacterial phyla have a higher proportion of monophyly (P=0.006, V = 67, using two-tailed Wilcoxon signed-rank test) than for those genes that trace to LECA. Plastids are younger than mitochondria, hence the genes from the ancestral plastid genome have had less time to migrate across prokaryotic genomes than genes from the ancestral mitochondrial genome. For the prokaryotic genes and phyla in question, evolution is not a vertical process. The bacterial reference system against which to infer the origin of eukaryotic genes that stem from the mitochondrion (or the plastid) is a system of mosaic (Martin 1999) or fluid (Esser et al. 2007) chromosomes. These findings are fully consistent with a recent larger scale investigation of gene verticality across genomes (Nagies et al. 2020).

If we accept the evidence that LGT in prokaryotes is real and if we accept the evidence that mitochondria were once endosymbiotic bacteria, then the expectation for the phylogeny of a gene that was acquired from the mitochondrion is that it traces to a single origin in LECA, which the genes in this study do, but "not" that it traces to alphaproteobacteria. This is because LGT among prokaryotes preceding and subsequent to the origin of mitochondria generates the illusion of many donors by shuffling the taxon labels attached to genes in mosaic bacterial chromosomes (Martin 1999). Most current studies still equate mitochondrial origin with an alphaproteobacterial sister group relationship (Vosseberg et al. 2021), but if we look at all the data, it is clear that such an interpretation is too strict. For example, Vosseberg et al. (2021) found that about 7% of the eukaryotic protein-domains that they examined branched with alphaproteobacterial homologs. But looking beyond the eukaryotic branch, Nagies et al. (2020) found that only about 35% of alphaproteobacterial genes recover alphaproteobacteria monophyly to begin with, and only 16% of the 220 trees in which alphaproteobacteria appeared as the sole sister of all eukaryotes recovered aphaproteobacteria as monophyletic among prokaryotes. To investigate mitochondrial origin from the standpoint of genes, it is not enough to identify the relationship of eukaryote genes to prokaryotic homologs. One has also to investigate the relationship of prokaryotic homologs to each other, because they are the reference system for comparison.

It is because of LGT among prokaryotes that many different groups are implicated as donors of genes to LECA (fig. 4; see also Nagies et al. 2020). There is no evidence independent of gene phylogenies to suggest or support theories for the participation of spirochaetes (Margulis et al. 2006), actinobacteria (Cavalier-Smith 2002), cyanobacteria (Cavalier-Smith 1975), deltaproteobacteria (López-García and Moreira 1999), planctomycetes (Cavalier-Smith and Chao 2020), or multiple donor lineages (Gabaldón 2018) at eukaryote origin (Embley and Martin 2006). One could of course argue that those conflicting theories for contributions from many different prokaryotic lineages are all simultaneously true, but then theories for eukaryogenesis would no longer be constrained by observations in data, and any assertion about eukaryote origin would be permissible as a line of evidence, an untenable state of affairs. The same sets of considerations apply to the cyanobacterial origin of plastids (fig. 4).

If we let go of the belief that sister group relationships between eukaryotic genes and prokaryotic homologs (fig. 4) identify the prokaryotic lineages that donated genes (Martin 1999; Nagies et al. 2020), and take into account the functions encoded by nuclear genes of bacterial origin that were duplicated in LECA (figs. 2 and 4; table 1), the simplest interpretation of the data in our view is that the bacterial duplicates in LECA were donated by the mitochondrion. Other more complicated interpretations are imaginable, but these interpretations do not simultaneously account for the phylogenetic behavior of the bacterial reference phylogeny set, which we have done here and elsewhere (Nagies et al. 2020). Our data furthermore show that eukaryotic genes are of monophyletic origin. With large genomic samples spanning thousands of reference prokaryotic genomes, eukaryotic gene evolution is clearly vertical, both in terms of lineage-specific distribution of gene duplications (fig. 1) and in terms of likelihood ratio tests (Nagies et al. 2020).

Can Positive Selection Explain Excess Bacterial Duplications?

The vast excess of bacterial duplications (fig. 3) and the phylogenies of 2,575 genes that would address the question of gene origin (fig. 4) speak in favor of bacterial acquisition in LECA from a single-resident endosymbiont, the mitochondrion, prior to the origin of eukaryote complexity. Yet one could still imagine numerous individual gene acquisitions in LECA from different donors with a blanket ad hoc hypothesis of "positive selection" increasing the copy number of bacterial-related functions to account for the excess of bacterial-derived duplications (table 1). However, the selection proposal would not explain the excess of bacterial over archaeal or eukaryote-specific genes with the same functional category, as is widely observed in table 1. That is, selection would have to be invoked as a special plea on a bacterial-gene-for-bacterial-gene basis, requiring yet one additional corollary of positive selection for each duplication. Because we observe over 900,000 duplications in the present data, the selection theory to account for duplications carries a burden of too many corollary assumptions.

On the other hand, it is possible that duplications are fundamentally mechanistic in origin, via chromosome mispairing, translocations, genome duplications, or via duplicative transfers from a resident endosymbiont as we argue in this paper. In a context of mosaic, fluid bacterial genomes (Martin 1999; Esser et al. 2007) permitting LGT among prokaryotes (fig. 4) (Nagies et al. 2020), we would require no corollary

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

assumptions of ad hoc selection. The mechanism of transfer from the endosymbiont generates the excess of bacterial duplications and does so across all functional categories (table 1).

The Functions of Bacterial Duplicates Polarize Events at LECA's Origin

Gene duplications speak to more than phylogeny. Gene duplications are a standard proxy for the evolution of complexity, as diversification of function and form is canonically underpinned by gene family expansion (Ohno 1970). Accordingly, we observe that the morphologically most complex multicellular eukaryotes—plants, animals, and fungi—harbor the largest numbers of duplications (fig. 1). As outlined above, the simplest interpretation of the present data is that complexity started with the mitochondrion. That is not only true for the present data on duplications, is also true from a purely physiological standpoint (Martin et al. 2017) and a bioenergetic standpoint (Lane and Martin 2010).

The functions of genes that were duplicated in LECA help to polarize events in LECA's evolution. For example, LECA had a mitochondrion. LECA's gene duplications in 47 genes with mitochondrial functions include pyruvate dehydrogenase complex, enzymes of the citric acid cycle, components involved in electron transport, a presequence cleavage protease, the ATP–ADP carrier, and seven members of the eukaryote-specific mitochondrial carrier family that facilitates metabolite exchange between the mitochondrion and the cytosol (table 1 and supplementary tables 3 and 4). A recent study estimated that some genes for mitochondrial function were probably duplicated in LECA, but interpreted the data as evidence for mitochondria-intermediate hypothesis (Vosseberg et al. 2021). The methodology used in Vosseberg et al. has major limitations because: 1) the timing of gene duplications was inferred using an approach that equates branch-lengths from phylogenetic trees to time, which is expected to be valid "only if" the evolutionary rate is constant across genes (substitutions and gene loss, for example); 2) prokaryotic sequences were arbitrarily removed from gene trees, inflating the estimates of duplications in genes of archaeal origin; 3) the use of trees for which the same gene sequence can be represented simultaneously in multiple trees, biasing the estimates of duplications and their origin; and 4) the use of too liberal thresholds for gene clustering which result in aberrantly large gene families (see supplementary fig. 5, Supplementary Material online), a potential source of tree reconstruction errors. By contrast, we do not infer time from branch lengths, we did not remove sequences that did not fit our expectations, and gene membership in our gene families is always unique.

Our findings clearly indicate that canonical energy metabolic functions of mitochondria were established in LECA, underscored by additional functions performed by mitochondria in diverse eukaryotic lineages: ten genes for enzymes of the lipid biosynthetic pathway (typically mitochondrial in eukaryotes; Gould et al. 2016), the entire glycolytic pathway (mitochondrial among marine algae; Río Bártulos et al. 2018), and 11 genes involved in redox balance are found among bacterial duplicates. The largest category of duplications with annotated functions concerns metabolism and biosynthesis (table 1).

Many products of bacterial-derived genes operate in the eukaryotic cytosol (Martin et al. 1993; Esser et al. 2004). This is because at the outset of gene transfer from the endosymbiont, there was no mitochondrial protein import machinery (Martin and Müller 1998; Dolezal et al. 2006), and no nucleus, such that the products of genes transferred from the endosymbiont were active in the compartment where the genes were cotranscriptionally translated (French et al. 2007). Gene transfers in large, genome sized fragments from the endosymbiont, as they occur today (Timmis et al. 2004; Portugez 2018), furthermore, permitted entire pathways to be transferred, because the unit of biochemical selection is the pathway and its product, not the individual enzyme (Martin 2010). In the absence of upstream and downstream intermediates and activities in a pathway, the product of a lone transferred gene is generally useless for the cell, expression of the gene becomes a burden, and the transferred gene cannot be fixed (Martin 2010).

Bacterial-derived duplications are present in functions that underpinned the origin of cell compartmentation in LECA (table 1). LECA possessed an endomembrane system consisting of bacterial lipids, as symbiogenic models predict (Gould et al. 2016). Bacterial duplicates, not archaeal duplicates, dominate lipid synthesis and membrane biogenesis (table 1). Functions of bacterial duplicates are also involved in mRNA splicing, a selective force at the origin of the nucleus (Garg and Martin 2016; Eme et al. 2017). The origin of protein import into mitochondria was essential to mitochondrial origin (Dolezal et al. 2006) and encompasses many bacteria-derived duplicates (table 1). LECA's duplicates of bacterial origin are also involved in the origin of eukaryotic-specific traits, including the cell cycle, the cytoskeleton, endomembrane system, and mRNA splicing (table 1). Eukaryote complexity required intracellular molecular movement in the cytosol, which is realized by motor proteins. The protein with the most duplications found in LECA is a light chain dynein with 12 duplications (supplementary table 3), in agreement with previous studies of dynein evolution that document massive dynein gene duplications early in eukaryote evolution (Kollmar 2016).

Notably, ten of the 20 genes encoding cytoskeletal functions that were duplicated in LECA (supplementary tables 3 and 4) encode dynein or kinesin motor proteins (see also Tromer et al. 2019). The bacterial duplicate contribution vastly outnumbers the archaeal contribution to these categories, which are dominated by eukaryote-specific genes, indicating that eukaryotes not only acquired genes, but they also

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

invented new ones as well (Lane and Martin 2010). Duplications in LECA depict bacterial carbon and energy metabolism in an archaeal host supported by genes that were recurrently donated by a resident symbiont, in line with the predictions of symbiotic theories for the nature of the first eukaryote (Martin and Müller 1998; Martin et al. 2017; Imachi et al. 2020). The functions of duplications are consistent with the predictions of symbiogenic theories but contrast with gradualist theories positing eukaryote origin from an archaeal lineage that attained eukaryote-like complexity in the absence of the mitochondrial endosymbiont (Cavalier-Smith 2002; Booth and Doolittle 2015; Pittis and Gabaldón 2016; Hampl et al. 2019).

What Does This Say about the Biology of LECA?

Gene transfers from the mitochondrion can generate duplications of bacterial-derived genes. What mechanisms promoted genome-wide gene duplication at the prokaryote–eukaryote transition? Population genetic parameters such as variation in population size (Zachar and Szathmáry 2017) apply to prokaryotes and eukaryotes equally, hence they would not affect gene duplications specifically in eukaryotes, but recombination processes (Garg and Martin 2016) in a nucleated cell could. Because LECA possessed meiotic recombination (Speijer et al. 2015), it was able to fuse nuclei (karyogamy). Karyogamy in a multinucleate LECA would promote the accumulation of duplications in all gene classes and promote genome expansion to its energetically permissible limits (Lane and Martin 2010) because unequal crossing between imprecisely paired homologous chromosomes following karyogamy generates duplications (Ohno 1970; Scannell et al. 2006; Hittinger and Carroll 2007; Van De Peer 2009). At the origin of meiotic recombination, chromosome pairing and segregation cannot have been perfect from the start; the initial state was likely error-prone, generating nuclei with aberrant gene copies, aberrant chromosomes, and even aberrant chromosome numbers. In cells with a single nucleus, such variants would have been lethal; in multinucleate (syncytial or coenocytic) organisms, defective nuclei can complement each other through mRNA in the cytosol (Garg and Martin 2016). Multinucleate forms are present throughout eukaryotic lineages (fig. 5), and ancestral reconstruction of nuclear organization clearly indicates that LECA itself was multinucleate (fig. 5 and supplementary fig. 1, Supplementary Material online). The multinucleate state enables the accumulation of duplications in the incipient eukaryotic lineage in a mechanistically nonadaptive manner, whereby duplications are implicated in the evolution of complexity (Ohno 1970; Scannell et al. 2006; Hittinger and Carroll 2007; Van De Peer 2009), as observed in the animal lineage (fig. 1). The syncytial state presents a viable intermediate state in the transition from prokaryote to eukaryote genetics.

Conclusion

Serial transfers of mitochondrial DNA to the chromosomes of the host are not only a mechanism of gene duplication, they are a form of endosymbiont genome duplication in which an original copy is retained in the organelle and remains functional. Gene duplications in LECA support an early origin of mitochondria and record the onset of the eukaryotic gene duplication process, a hallmark of genome evolution in mitosing cells (Ohno 1970; Scannell et al. 2006; Hittinger and Carroll 2007; Van De Peer 2009; Treangen and Rocha 2011).

Materials and Methods

Protein Clustering and Tree Reconstruction for Gene Duplication Inferences

Protein sequences for 150 eukaryotic genomes were downloaded from NCBI, Ensembl Protists, and JGI (see supplementary data 1 for detailed species composition). To construct gene families, we performed an all-vs-all BLAST (Altschul et al. 1997) of the eukaryotic proteins and selected the reciprocal best BLAST hits with e-value $\leq 10^{-10}$. The protein pairs were aligned with the Needleman–Wunsch algorithm (Rice et al. 2000) and the pairs with global identity values $<25\%$ were discarded. The retained global identity pairs were used to construct gene families with the Markov clustering algorithm (Enright et al. 2002) (version 12-068) with default parameters. Because in this study we were interested in gene duplications, we considered only the gene families with multiple gene copies in at least two eukaryotic genomes. Our criteria retained a total of 24,571 multicopy gene families.

Protein-sequence alignments for the individual eukaryotic multicopy gene families were generated using MAFFT (Katoh 2002), with the iterative refinement method that incorporates local pairwise alignment information (L-INS-i, version 7.130). The alignments were used to reconstruct maximum likelihood trees with IQ-tree (Nguyen et al. 2015), using default settings (version 1.6.5), and the trees were rooted with MAD (Tria et al. 2017) (supplementary data 2).

Inference of Gene Duplication

Gene duplications were inferred from gene trees by assigning duplication events to internal nodes in the rooted topologies. Given a rooted gene tree with n leaves, let S be the set of species labels for the leaves. For the case of paralogous gene trees, there is at least one leaf pair, a and b, such that $s_a{=}s_b$. Assigning a gene duplication to the last common ancestor of the pair a and b corresponds to the evolutionary scenario that minimizes paralog losses in the gene tree. For each rooted gene tree, we performed pairwise comparisons of all leaf pairs with identical species labels to infer all the internal nodes corresponding to gene duplications using the minimal loss criterion for each leaf pair. Note that, this approach considers

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

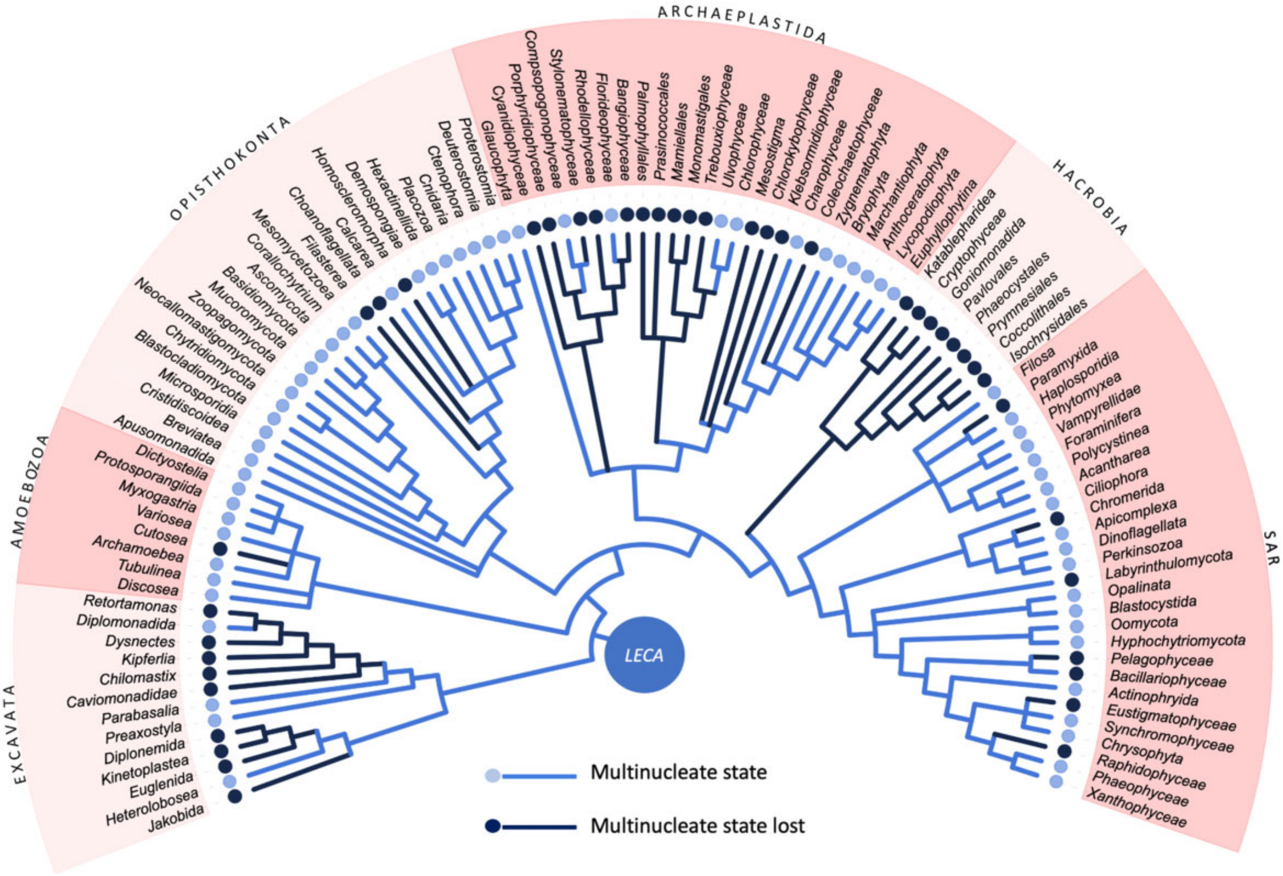


FIG. 5.—Ancestral state reconstruction for nuclear organization in eukaryotes. Presence and absence of the multinucleate state in members of the respective group are indicated. Resolution of the branches (polytomy vs. dichotomy) does not alter the outcome of the ancestral state reconstruction, nor does position of the root on the branches leading to Amoebozoa, Excavata, or Opisthokonta. LECA was a multinucleate, syncytial cell, not uninucleate (see supplementary fig. 1, Supplementary Material online). Together with mitochondrion and sex, the multinucleate state is ancestral to eukaryotes and fostered accumulation of duplications (see text).

the possibility of multiple gene duplications per gene tree (supplementary fig. 2, Supplementary Material online). We summarized the gene duplication inferences from all gene trees by evaluating the distribution of descendant paralogs across the eukaryotic supergroups for each gene duplication event (fig. 2).

The inferences of gene duplications in the present work are based on trees that were rooted with MAD (Tria et al. 2017). A recent comparison of MAD with other methods showed that MAD performs better than other rooting methods currently in use (Wade et al. 2020).

Inference for the Origin of Eukaryotic Duplicates

For identification of homologs in prokaryotes, we used all protein-coding genes from 5,656 prokaryotic genomes downloaded from RefSeq (Pruitt et al. 2007) (see supplementary data 3) and compared them against eukaryotic protein-coding genes using Diamond (Buchfink et al. 2015) to perform sequence searches with the "more-sensitive" parameter. A eukaryotic gene family was considered to have homologs in prokaryotes if at least one gene of the eukaryotic family had a significant hit against a prokaryotic gene (e-value $<10^{-10}$ and local identity $\geq$25%). Gene families with homologs only in archaeal genomes were considered as genes of archaeal origin and similarly for bacteria. Gene families with significant hits in both archaea and bacteria (universal) could have originated from either archaea or bacteria.

We purposefully avoided using trees to inferring the origin of eukaryotic genes because of low levels of sequence conservation entailing a large number of prokaryotic homologs. Note, however, that we reconstructed trees for the subset of eukaryote–prokaryote genes with sufficient sequence conservation (see below). We found that the presence–absence of homologs across prokaryotic taxa remarkably recapitulates the distribution of prokaryotic sisters derived from phylogenetic trees serving, thus, as a validation of our approach (supplementary table 5).

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

Prokaryote–Eukaryote Protein Clustering and Tree Reconstruction

To assemble a data set of conserved genes for phylogenies linking prokaryotes and eukaryotes, eukaryotic, archaeal, and bacterial protein sequences were first clustered separately before homologous clusters between eukaryotes and prokaryotes were identified as described (Ku et al. 2015). Eukaryotic sequences for the 150 genomes (supplementary data 1) were clustered with MCL (Enright et al. 2002) using global identities from best reciprocal BLAST (Altschul et al. 1997) hits for protein pairs with e-value $\leq 10^{-10}$ and global identity $\geq 40\%$. The clusters with genes distributed in more than one eukaryotic genome were retained. Similarly, prokaryotic protein sequences from 5,655 genomes (see supplementary data 3, except for MK-D1 for which the genome was unavailable by the time the data were compiled) were clustered using the best reciprocal BLAST for protein pairs with e-value $\leq 10^{-10}$ and global identity $\geq 25\%$, for archaea and bacteria separately. The resulting clusters with gene copies in at least five prokaryotic genomes were retained. The most universally distributed clusters comprise 20–40 proteins, the majority of which are involved in translation (supplementary fig. 4, Supplementary Material online). Eukaryotic and prokaryotic clusters were merged using the reciprocal best cluster procedure. We merged a eukaryotic cluster with a prokaryotic cluster if $\geq 50\%$ of the eukaryotic sequences in the cluster have their best reciprocal BLAST hit in the same prokaryotic cluster and vice versa (cut-offs: e-value $\leq 10^{-10}$ and local identity $\geq 30\%$). We refer to the merged cluster as eukaryotic–prokaryotic cluster (EPC).

Protein-sequence alignments for 2,575 EPCs were generated using MAFFT (Katoh 2002) (L-INS-i, version 7.130). The alignments were used to reconstructed maximum-likelihood trees with IQ-tree (Nguyen et al. 2015) (version 1.6.5) employing default settings (supplementary data 4).

Tests for Eukaryote Monophyly

For 475 gene trees where eukaryotes were not recovered as monophyletic, we conducted the Shimodaira–Hasegawa (Shimodaira and Hasegawa 1999) (SH), Kishino–Hasegawa (Kishino and Hasegawa 1989) (KH), and approximately unbiased (AU) test (Shimodaira 2002) to determine whether the observed nonmonophyly was statistically significant. We reconstructed trees constraining eukaryotic sequences to be monophyletic, but not imposing any other topological constraint, using FastTree (Price et al. 2010) (version 2.1.10 SSE3) and recording all trees explored during the tree search with the "-log" parameter (supplementary data 5). The sample of monophyletic trees was used as input in IQ-tree (Nguyen et al. 2015) (version 2.0.3; parameter: "-zb 100000 –au") to perform the SH, KH, and AU tests against the unconstrained tree (nonmonophyletic). If the best-constrained tree did not show significant difference relative to the unconstrained tree ($P < 0.05$), then we considered that eukaryotic monophyly cannot be rejected.

Inference of Prokaryotic Sisters

To infer prokaryotes sisters to eukaryotes in the gene trees we used the unconstrained tree if eukaryotes were recovered as monophyletic and the constrained tree if eukaryotes were not recovered as monophyletic, since the SH test did not reject eukaryote monophyly for any gene tree (see main text). Note that in unrooted trees for which eukaryotes are monophyletic, the prokaryotic side of the tree is bisected by one internal node into two prokaryotic subclades, each subclade being the potential sister to eukaryotes (see fig. 4*a*). We considered the prokaryotic subclade with the smallest number of leaves for our inferences of sister-relations and the prokaryotic phyla present in the sister clade and outgroup clade was recorded for each tree. The sister clades were scored as a "pure" sister when only a single prokaryotic phylum was present in the clade or as "mixed" sister when more than one phylum was present.

Ancestral Reconstruction of Eukaryotic Nuclear Organization

Ancestral state reconstructions were performed on the basis of a morphological character matrix, using maximum parsimony as implemented in Mesquite 3.6 (https://www.mesquiteproject.org/, accessed June 2019). The reference eukaryotic phylogeny includes 106 taxa (ranging from genus to phylum level) to reflect the relations within the eukaryotes and reduce taxonomic redundancy. The phylogeny includes members of six supergroups: Amoebozoa (Mycetozoa), Archaeplastida, Excavata, Hacrobia, Opisthokonta, and SAR, and was constructed by combining branches from previous studies (Burki et al. 2010; Yoon et al. 2010; Adl et al. 2012; Powell and Letcher 2014; Burki et al. 2016; Cavalier-Smith et al. 2016; Derelle et al. 2016; Spatafora et al. 2016; Yang et al. 2016; Archibald et al. 2017; Krabberød et al. 2017; McCarthy and Fitzpatrick 2017; Roger et al. 2017; Spatafora et al. 2017; Bass et al. 2018; Cavalier-Smith et al. 2018; Tedersoo et al. 2018; Irwin et al. 2019). The nuclear organization for each taxon was coded as 0 for nonmultinucleate, 1 for multinucleate or 0/1 if ambiguous according to the literature (Byers 1979; Willumsen et al. 1987; Barthel and Detmer 1990; Daniels and Pappas 1994; Walker et al. 2006; Steiner 2010; Yoon et al. 2010; Adl et al. 2012; Niklas et al. 2013; Maciver 2016; Spatafora et al. 2016; Archibald et al. 2017; Bloomfield et al. 2019) (supplementary data 6). In order to account for uncertainties of lineage relations among eukaryotes, we used a set of phylogenies with alternative root positions (Vossbrinck et al. 1987; Stechmann and Cavalier-Smith 2002; Katz and Grant 2015) (altogether a total of 15 different roots) as well as the consideration of polytomies for debated branches (supplementary data 6). All ancestral state reconstruction

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

rendered LECA as multinucleated, with no ambiguity. Ambiguous reconstructions, however, were observed within supergroups in some topologies but did not pose ambiguity to the reconstructed state in LECA.

Supplementary Material

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

Acknowledgments

We thank the European Research Council (Grant No. 666053), the Volkswagen Foundation (Grant No. 93 046), and the Moore Simons Initiative on the Origin of the Eukaryotic Cell (Grant No. 9743) for financial support. We also thank Damjan Franjević (Department of Biology, University of Zagreb, Croatia) for helpful discussions.

Author Contributions

All authors conceived and designed the study. J.B. and J.S. prepared the data sets with contribution from all the authors. F.D.K.T. performed gene duplication inferences, functional annotation of genes, and the tests for eukaryotic monophyly. J.B. and F. N. performed the analyses of eukaryotic sisters. J.S. compiled the eukaryotic phylogenies and performed ancestral state reconstructions. All authors wrote the paper.

Data Availability

Supplementary tables and data used in this study are available under the link https://doi.org/10.6084/m9.figshare.12249260.

Code Availability

Custom Matlab scripts used to perform data analysis are available upon request.

Literature Cited

Adl SM, et al. 2012. The revised classification of eukaryotes. J Eukaryot Microbiol. 59(5):429–493.

Albalat R, Cañestro C. 2016. Evolution by gene loss. Nat Rev Genet. 17(7):379–391.

Allen JF. 2015. Why chloroplasts and mitochondria retain their own genomes and genetic systems: colocation for redox regulation of gene expression. Proc Natl Acad Sci U S A. 112(33):10231–10238.

Altschul SF, et al. 1997. Blast and Psi-Blast: protein database search programs. Nucleic Acid Res. 25:2289–4402.

Andersson JO, et al. 2003. Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. Curr Biol. 13:94–104.

Archibald JM, et al. 2017. Handbook of the protists. Cham: Springer Nature.

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

Barthel D, Detmer A. 1990. The spermatogenesis of *Halichondria panicea* (Porifera, Demospongiae). Zoomorphology 110:9–15.

Bass D, et al. 2018. Clarifying the relationships between microsporidia and cryptomycota. J Eukaryot Microbiol. 65(6):773–782.

Betts HC, et al. 2018. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nat Ecol Evol. 2:1556–1562.

Bloomfield G, et al. 2019. Triparental inheritance in *Dictyostelium*. Proc Natl Acad Sci U S A. 116(6):2187–2192.

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

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12(1):59–60.

Burki F, et al. 2010. Evolution of Rhizaria: new insights from phylogenomic analysis of uncultivated protists. BMC Evol Biol. 10:377.

Burki F, et al. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc R Soc Lond B. 283:20152802.

Byers TJ. 1979. Growth, reproduction, and differentiation in Acanthamoeba. Int Rev Cytol. 61:283–338.

Cavalier-Smith T, Chao EE. 2020. Multidomain ribosomal protein trees and the planctobacterial origin of neomura (eukaryotes, archaebacteria). Protoplasma 257(3):621–753.

Cavalier-Smith T, et al. 2016. 187-gene phylogeny of protozoan phylum Amoebozoa reveals a new class (Cutosea) of deep-branching, ultrastructurally unique, enveloped marine Lobosa and clarifies amoeba evolution. Mol Phylogenet Evol. 99:275–296.

Cavalier-Smith T, et al. 2018. Multigene phylogeny and cell evolution of chromist infrakingdom Rhizaria: contrasting cell organisation of sister phyla Cercozoa and Retaria. Protoplasma 255(5):1517–1574.

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

Cavalier-Smith T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. Int J Syst Evol Microbiol. 52(Pt 2):297–354.

Daniels EW, Pappas GD. 1994. Reproduction of nuclei in *Pelomyxa palustris*. Cell Biol Int. 18(8):805–812.

de Duve C. 2007. The origin of eukaryotes: a reappraisal. Nat Rev Genet. 8(5):395–403.

Derelle R, et al. 2016. Phylogenomic framework to study the diversity and evolution of Stramenopiles (= Heterokonts). Mol Biol Evol. 33(11):2890–2898.

Dolezal P, Likic V, Tachezy J, Lithgow T. 2006. Evolution of the molecular machines for protein import into mitochondria. Science 313(5785):314–318.

Doolittle FW. 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.

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

Eme L, et al. 2017. Archaea and the origin of eukaryotes. Nat Rev Microbiol. 15(12):711–723.

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

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

Esser C, Martin W, Dagan T. 2007. The origin of mitochondria in light of a fluid prokaryotic chromosome model. Biol Lett. 22:180–184.

French SL, Santangelo TJ, Beyer AL, Reeve JN. 2007. Transcription and translation are coupled in Archaea. Mol Biol Evol. 24(4):893–895.

Gabaldón T. 2018. Relative timing of mitochondrial endosymbiosis and the "pre-mitochondrial symbioses" hypothesis. IUBMB Life. 70(12):1188–1196.

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

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

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

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

Hampl V, Čepička I, Eliáš M. 2019. Was the mitochondrion necessary to start eukaryogenesis? Trends Microbiol. 27(2):96–104.

Hampl V, et al. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic 'supergroups'. Proc Natl Acad Sci U S A. 106(10):3859–3864.

He D, et al. 2014. An alternative root for the eukaryote tree of life. Curr Biol. 24(4):465–470.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449(7163):677–681.

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

Irwin NA, et al. 2019. Phylogenomics supports the monophyly of the Cercozoa. Mol Phylogenet Evol. 130:416–423.

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

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22(4):225–231.

Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059–3066.

Katz LA, Grant JR. 2015. Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. Syst Biol. 64(3):406–415.

Keeling PJ, Palmer LD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 9(8):605–618.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol. 29(2):170–179.

Kollmar M. 2016. Fine-tuning motile cilia and flagella: evolution of the dynein motor proteins from plants to humans at high resolution. Mol Biol Evol. 33(12):3249–3267.

Krabberød AK, et al. 2017. Single cell transcriptomics, mega-phylogeny, and the genetic basis of morphological innovations in Rhizaria. Mol Biol Evol. 34(7):1557–1573.

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

Ku C, Martin WF. 2016. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70% rule. BMC Biol. 14(1):89.

Lake JA, Rivera MC. 1994. Was the nucleus the first endosymbiont? Proc Natl Acad Sci U S A. 91(8):2880–2881.

Lane N, Martin W. 2010. The energetics of genome complexity. Nature 467(7318):929–934.

Leger MM, et al. 2018. Demystifying eukaryote lateral gene transfer. Bioessays 40(5):e1700242.

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

López-García P, Moreira G. 2015. Open questions on the origin of eukaryotes. Trends Ecol Evol. 30(11):697–708.

Maciver SK. 2016. Asexual amoebae escape Muller's ratchet through polyploidy. Trends Parasitol. 32(11):855–862.

Makarova KS, et al. 2005. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. Nucleic Acids Res. 33(14):4626–4638.

Margulis L, Chapman M, Guerrero R, Hall J. 2006. The last eukaryotic common ancestor (LECA): acquisition of cytoskeletal motility from aerotolerant spirochetes in the Proterozoic Eon. Proc Natl Acad Sci U S A. 103(35):13080–13085.

Margulis L, et al. 2000. The chimeric eukaryote: origin of the nucleus from the karyomastigont in amitochondriate protists. Proc Natl Acad Sci U S A. 97(13):6954–6959.

Martin W. 1999. Mosaic bacterial chromosomes: a chalenge en route to a tree of genomes. Bioessays 21:99–104.

Martin W. 2010. Evolutionary origins of metabolic compartmentalization in eukaryotes. Philos Trans R Soc Lond B Biol Sci. 365(1541):847–855.

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 U S A. 90(18):8692–8696.

Martin W, et al. 2001. An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. Biol Chem. 382(11):1521–1539.

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

Martin WF, Garg S, Zimorski V. 2015. Endosymbiotic theory for eukaryote origin. Philos Trans R Soc Lond B. 370:20140330.

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

McCarthy CG, Fitzpatrick DA. 2017. Multiple approaches to phylogenomic reconstruction of the fungal kingdom. Adv Genet. 100:211–266.

Mereschkowsky C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol Centralbl. 25:593–604. English translation in 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.

Nagies FSP, Brueckner J, Tria FDK, Martin WF. 2020. A spectrum of verticality across genes. PLoS Genet. 16(11):e1009200.

Nei M, Gu X, Sitnikova T. 1997. Evolution by birth and death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A. 94(15):7799–7806.

Nelson-Sathi S, et al. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc Natl Acad Sci U S A. 109(50):20537–20542.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1):268–274.

Niklas KJ, et al. 2013. The evo-devo of multinucleate cells, tissues, and organisms, and an alternative route to multicellularity. Evol Dev. 15(6):466–474.

Ohno S. 1970. Evolution by gene duplication. Heidelberg (Berlin): Springer.

Parfrey LW, et al. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci U S A. 108:1364–13629.

Pittis AA, Gabaldón T. 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature 531(7592):101–104.

Poole AM, Gribaldo S. 2014. Eukaryotic origin: how and when was the mitochondrion acquired? Cold Spring Harb Perspect Biol. 6(12):a015990.

Portugez S, Martin WF, Hazkani-Covo E. 2018. Mosaic mitochondrial-plastid insertions into the nuclear genome show evidence of both non-homologous end joining and homologous recombination. BMC Evol Biol. 18(1):162.

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

Powell MJ, Letcher PM. 2014. 6 Chytridiomycota, Monoblepharidomycota, and Neocallimastigomycota. In: McLaughlin DJ, Spatafora JW, editors. 2nd ed. The Mycota Part VII A. Systematics and evolution. Heidelberg (Berlin): Springer. p. 141–175.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One 5(3):e9490.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35(Database issue):D61–D65.

Ren R, et al. 2016. Phylogenetic resolution of deep eukaryotic and fungal relationships using highly conserved low-copy nuclear genes. Genome Biol Evol. 8(9):2683–2701.

Rice P, et al. 2000. EMBOSS: the European Molecular Biology Open software suite. Trends Genet. 16(6):276–277.

Río Bártulos C, et al. 2018. Mitochondrial glycolysis in a major lineage of eukaryotes. Genome Biol Evol. 10(9):2310–2325.

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

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440(7082):341–345.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51(3):492–508.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114–1116.

Spang A, et al. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521(7551):173–179.

Spatafora JW, et al. 2016. A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. Mycologia 108(5):1028–1046.

Spatafora JW, et al. 2017. The fungal tree of life: from molecular systematics to genome-scale phylogenies. Microbiol Spectr. 5(5):1–32.

Speijer D, Lukeš J, Eliáš M. 2015. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. Proc Natl Acad Sci U S A. 112(29):8827–8834.

Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science 297(5578):89–91.

Steiner JM. 2010. Technical notes: growth of *Cyanophora paradoxa*. J Endoc Cell Res. 20:62–67.

Tedersoo L, et al. 2018. High-level classification of the fungi and a tool for evolutionary ecological analyses. Fungal Div. 90:135–159.

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

Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7(1):e1001284.

Tria FDK, Landan G, Dagan T. 2017. Phylogenetic rooting using minimal ancestor deviation. Nat Ecol Evol. 1:0193.

Tromer EC, van Hooff JJE, Kops GJPL, Snel B. 2019. Mosaic origin of the eukaryotic kinetochore. Proc Natl Acad Sci U S A. 116(26):12873–12882.

Van De Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10(10):725–732.

Vossbrinck CR, et al. 1987. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. Nature 326(6111):411–414.

Vosseberg J, et al. 2021. Timing the origin of eukaryotic cellular complexity with ancient duplications. Nat Ecol Evol. 5(1):92–100.

Wade T, et al. 2020. Assessing the accuracy of phylogenetic rooting methods on prokaryotic gene families. PLoS One 15(5):e0232950–e0233022.

Walker G, et al. 2006. Ultrastructural descripton of *Breviata anathema*, n. gen., n. sp., the organism previously studied as "*Mastigamoeba invertens*". J Eukaryot Microbiol. 53(2):65–78.

Wallin IE. 1925. On the nature of mitochondria. IX. Demonstration of the bacterial nature of mitochondria. Am J Anat. 36:131–139.

Willumsen NB, et al. 1987. A multinucleate amoeba, *Parachaos zoochlorellae* (Willumsen 1982) comb, nov., and a proposed division of the genus *Chaos* into the Genera *Chaos* and *Parachaos* (Gymnamoebia, Amoebidae). Archiv Protist. 134:303–313.

Yang EC, et al. 2016. Divergence time estimates and the evolution of major lineages in the florideophyte red algae. Sci Rep. 6:21361.

Yoon HS, et al. 2010. Evolutionary history and taxonomy of red algae. In: Seckbach, JChapman, DJ, editors. Red algae in genomic age. Dordrecht: Springer. p. 27–45.

Zachar I, Szathmáry E. 2017. Breath-giving cooperation: critical review of origin of mitochondria hypotheses. Biol Direct. 12:19.

Zaremba-Niedzwiedzka K, et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541(7637):353–358.

Zmasek CM, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. 12(1):R4.

Associate editor: Ellen Pritham

Downloaded from https://academic.oup.com/gbe/article/13/5/evab055/6178794 by Universitaets und Landesbibliothek user on 08 June 2021

4 Bibliography

Adam PS, Borrel G, Gribaldo S. 2018. Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proc Natl Acad Sci U S A*; 115: E1166–E1173.

Adams DA, Nelson RR, Todd PA. 1992. Perceived usefulness, ease of use, and usage of information technology: A replication. *MIS Q*; 16: 227–247.

Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Pryszcz LP, Schreiber F, da Silva AS, Szklarczyk D, Train CM, Bork P, Lecompte O, von Mering C, Xenarios I, Sjölander K, Jensen LJ, *et al*. 2016. Standardized benchmarking in the quest for orthologs. *Nat Methods*; 13: 425–430.

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol*; 5: e1000262.

Altenhoff AM, Glover NM, Train CM, Kaleb K, Warwick Vesztrocy A, Dylus D, de Farias TM, Zile K, Stevenson C, Long J, Redestig H, Gonnet GH, Dessimoz C. 2018. The OMA orthology database in 2018: Retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res*; 46: D477–D485.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*; 215: 403–410.

Alva V, Söding J, Lupas AN. 2015. A vocabulary of ancient peptides at the origin of folded proteins. *Elife*; 4: e09410.

Aminov RI. 2010. A brief history of the antibiotic era: Lessons learned and challenges for the future. *Front Microbiol*; 1: 134.

Aravind L, Iyer LM, Koonin EV. 2006. Comparative genomics and structural biology of the molecular innovations of eukaryotes. *Curr Opin Struct Biol*; 16: 409–419.

Arndt NT, Nisbet EG. 2012. Processes on the young earth and the habitats of early life. *Annu Rev Earth Planet Sci*; 40: 521–549.

Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U, Simonyan V, Kimchi-Sarfaty C. 2017. A new and updated resource for codon usage tables. *BMC Bioinformatics*; 18: 391.

Balaban M, Moshiri N, Mai U, Jia X, Mirarab S. 2019. TreeCluster: Clustering biological sequences using phylogenetic trees. *PLoS One*; 14: e0221068.

Bar-On YM, Phillips R, Milo R. 2018. The biomass distribution on earth. *Proc Natl Acad Sci U S A*; 115: 6506–6511.

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2014. GenBank. *Nucleic Acids Res*; 42: D32–D37.

Betts HC, Puttick MN, Clark JW, Williams TA, Donoghue PCJ, Pisani D. 2018. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol*; 2: 1556–1562.

Brocchieri L. 2001. Phylogenetic inferences from molecular sequences: Review and critique. *Theor Popul Biol*; 59: 27–40.

Brocchieri L, Karlin S. 2005. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res*; 33: 3390–3400.

Brueckner J, Martin WF. 2020. Bacterial genes outnumber archaeal genes in eukaryotic genomes. *Genome Biol Evol*; 12: 282–292.

Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*; 12: 59–60.

Cantarel BL, Morrison HG, Pearson W. 2006. Exploring the relationship between sequence similarity and accurate phylogenetic trees. *Mol Biol Evol*; 23: 2090–2100.

Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: Models and estimation procedures. *Evolution*; 21: 550–570.

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

Chatton E. 1925. *Pansporella perplexa*: Amoebien à spores protégées parasite des daphnies. Réflexions sur la biologie et la phylogénie des protozoaires. *Ann Sci Nat Zool*; 8: 5–85.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*; 2: e383.

Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: A program for dating gene duplications and optimizing gene family trees. *J Comput Biol*; 7: 429–447.

Chor B, Tuller T. 2005. Maximum likelihood of evolutionary trees: Hardness and approximation. *Bioinformatics*; 21: i97–i106.

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

Coleman GA, Davín AA, Mahendrarajah T, Spang A, Hugenholtz P, Szöllősi GJ, Williams TA. 2020. A rooted phylogeny resolves early bacterial evolution. *bioRxiv*; 2020.07.15.205187.

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 U S A*; 107: 17252–17255.

Crow KD, Wagner GP. 2006. What is the role of genome duplication in the evolution of complexity and diversity? *Mol Biol Evol*; 23: 887–892.

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

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

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

Dalquen DA, Dessimoz C. 2013. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. *Genome Biol Evol*; 5: 1800–1806.

Darwin C. 1859. On the origin of species. London: John Murray.

Datta RS, Meacham C, Samad B, Neyer C, Sjölander K. 2009. Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Res*; 37: W84–W89.

Davies J, Davies D. 2010. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev*; 74: 417–433.

Dayhoff MO, Eck RV. 1968. Atlas of protein sequence and structure. Silver Spring (Md.): National Biomedical Research Foundation.

Dehal PS, Boore JL. 2006. A phylogenomic gene cluster resource: The phylogenetically inferred groups (PhIGs) database. *BMC Bioinformatics*; 7: 201.

DeLuca TF, Cui J, Jung JY, St Gabriel KC, Wall DP. 2012. Roundup 2.0: Enabling comparative genomics for over 1800 genomes. *Bioinformatics*; 28: 715–716.

DeLuca TF, Wu IH, Pu J, Monaghan T, Peshkin L, Singh S, Wall DP. 2006. Roundup: A multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*; 22: 2044–2046.

Dessimoz C, Cannarozzi G, Gil M, Margadant D, Roth A, Schneider A, Gonnet GH. 2005. OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: Introduction and first achievements. Berlin, Heidelberg: Springer.

Dessimoz C, Gil M, Schneider A, Gonnet GH. 2006. Fast estimation of the difference between two PAM/JTT evolutionary distances in triplets of homologous sequences. *BMC Bioinformatics*; 7: 529.

Doolittle RF. 1986. Of URFs and ORFs: A primer on how to analyze derived amino acid sequences. Mill Valley (Ca.): University Science Books.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*; 26: 2460–2461.

Edman P. 1950. Method for determination of the amino acid sequence in peptides. *Acta Chem Scand*; 4: 283–293.

Enright A, van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*; 30: 1575–1584.

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

Fan L, Wu D, Goremykin V, Xiao J, Xu Y, Garg S, Zhang C, Martin WF, Zhu R. 2020. Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. *Nat Ecol Evol*; 4: 1213–1219.

Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ. 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics*; 35: 6.12.1–6.12.19.

Fischer WW, Hemp J, Johnson JE. 2016. Evolution of oxygenic photosynthesis. *Annu Rev Earth Planet Sci*; 44: 647–683.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst Zool*; 19: 99–113.

Fitch WM, Harvey PH, Leigh Brown AJ, Smith JM. 1995. Uses for evolutionary trees. *Philos Trans R Soc Lond B Biol Sci*; 349: 93–102.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. *Science*; 155: 279–284.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*; 28: 3150–3152.

Gabald n T, Dessimoz C, Huxley-Jones J, Vilella AJ, Sonnhammer ELL, Lewis S. 2009. Joining forces in the quest for orthologs. *Genome Biol*; 10: 403.

Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. 2019. Microbial genome analysis: The COG approach. *Brief Bioinform*; 20: 1063–1070.

Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. 2021. COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res*; 49: D274–D281.

Goodman M, Czelusniak J, Koop BF, Tagle DA, Slightom JL. 1987. Globins: A case study in molecular phylogeny. *Cold Spring Harbor Symp Quant Biol*; 52: 875–890.

Goodstadt L, Ponting CP. 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol*; 2: e133.

Graur D. 2016. Molecular and Genome Evolution. Sunderland (Ma.): Sinauer Associates, Inc.

Graur D, Li WH. 2000. Fundamentals of molecular evolution. Sunderland (Ma.): Sinauer Associates, Inc.

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: 1655–1663.

Haeckel E. 1866. Generelle Morphologie der Organismen. Berlin: G. Reimer.

Haeckel E. 1874. Anthropogenie. Leipzig: Engelmann.

Huerta-Cepas J, Bueno A, Dopazo J, Gabald n T. 2008. PhylomeDB: A database for genome-wide collections of gene phylogenies. *Nucleic Acids Res*; 36: D491–D496.

Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabald n T. 2014. PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res*; 42: D897–D902.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*; 47: D309–D314.

Hulsen T, Huynen M, Vlieg J, Groenen P. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol*; 7: R31.

Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P, Cunningham F. 2018. Ensembl variation resources. *Database*; 2018: bay119.

Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, Takano Y, Uematsu K, Ikuta T, Ito M, Matsui Y, Miyazaki M, Murata K, Saito Y, Sakai S, Song C, Tasumi E, Yamanaka Y, Yamaguchi T, Kamagata Y, *et al*. 2020. Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature*; 577: 519–525.

Jaroszewski L, Li W, Godzik A. 2002. In search for more accurate alignments in the twilight zone. *Protein Sci*; 11: 1702–1713.

Javaux EJ. 2019. Challenges in evidencing the earliest traces of life. *Nature*; 572: 451–460.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: The beginning of incongruence? *Trends Genet*; 22: 225–231.

Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P. 2008. EggNOG: Automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*; 36: D250–D254.

Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. 2012. Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc Natl Acad Sci U S A*; 109: 16213–16216.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*; 30: 3059–3066.

Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Grabmueller C, Kumar N, Liu Z, Maurel T, Moore B, McDowall MD, Maheswari U, Naamati G, Newman V, Ong CK, Paulini M, *et al*. 2018. Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res*; 46: D802–D808.

Koonin EV, Novozhilov AS. 2009. Origin and evolution of the genetic code: The universal enigma. *IUBMB Life*; 61: 99–111.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. 2011. Computational methods for gene orthology inference. *Brief Bioinform*; 12: 379–391.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2019. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res*; 47: D807–D811.

Ku C, Martin WF. 2016. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: The 70 % rule. *BMC Biol*; 14: 89.

Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, Hazkani-Covo E, McInerney JO, Landan G, Martin WF. 2015. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature*; 524: 427–432.

Landan G, Graur D. 2007. Heads or Tails: A simple reliability check for multiple sequence alignments. *Mol Biol Evol*; 24: 1380–1383.

Landan G, Graur D. 2009. Characterization of pairwise and multiple sequence alignment errors. *Gene*; 441: 141–147.

Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thévenin A, Stoye J, Hartmann RK, Prohaska SJ, Stadler PF. 2014. Orthology detection combining clustering and synteny for very large datasets. *PLoS One*; 9: e105015.

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

Li H, Coghlan A, Ruan J, Coin LJ, Hériché JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GKS, Zheng W, Dehal P, Wang J, Durbin R. 2006. TreeFam: A curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*; 34: D572–D580.

Li L, Stoeckert Jr CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res*; 13: 2178–2189.

Liang P, Riley M. 2001. A comparative genomics approach for studying ancestral proteins and evolution. *Adv Appl Microbiol*; 50: 39–72.

Lippmann W. 1965. Public opinion. New York: Free Press.

Lupas AN, Alva V. 2017. Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins. *J Struct Biol*; 198: 74–81.

Makarova KS, Wolf YI, Koonin EV. 2015. Archaeal clusters of orthologous genes (arCOGs): An update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. *Life*; 5: 818–840.

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

Martin WF, Garg S, Zimorski V. 2015. Endosymbiotic theories for eukaryote origin. *Philos Trans R Soc Lond B Biol Sci*; 370: 20140330.

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, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*; 393: 162–165.

Mereschkowsky C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biol Centralbl*; 25: 593–604.

Mi H, Muruganujan A, Thomas PD. 2013. PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*; 41: D377–D386.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res*; 49: D412–D419.

Mushegian AR, Koonin EV. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A*; 93: 10268–10273.

Nagies FSP, Brueckner J, Tria FDK, Martin WF. 2020. A spectrum of verticality across genes. *PLoS Genet*; 16: e1009200.

Nakaya R, Nakamura A, Murata Y. 1960. Resistance transfer agents in *Shigella*. *Biochem Biophys Res Commun*; 3: 654–659.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*; 48: 443–453.

Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, Deppenmeier U, Martin WF. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc Natl Acad Sci U S A*; 109: 20537–20542.

Nelson-Sathi S, Sousa FL, Roettger M, Lozada-Chávez N, Thiergart T, Janssen A, Bryant D, Landan G, Schönheit P, Siebers B, McInerney JO, Martin WF. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature*; 517: 77–80.

Nisbet EG, Sleep NH. 2001. The habitat and nature of early life. *Nature*; 409: 1083–1091.

Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I. 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res*; 42: D26–D31.

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

Ohno S. 1970. Evolution by gene duplication. Berlin, Heidelberg: Springer.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farell CM, *et al*. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*; 44: D733–D745.

Pipenbacher P, Schliep A, Schneckener S, Schönhuth A, Schönhuth S, Schomburg D, Schrader R. 2002. ProClust: Improved clustering of protein sequences with an extended graph-based approach. *Bioinformatics*; 18: S182–S191.

Pisani D, Benton M, Wilkinson M. 2007. Congruence of morphological and molecular phylogenies. *Acta Biotheor*; 55: 269–281.

Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol;* 14: 615–623.

Pryszcz LP, Huerta-Cepas J, Gabald n T. 2011. MetaPhOrs: Orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res*; 39: e32.

Remm M, Storm CEV, Sonnhammer ELL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*; 314: 1041–1052.

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

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

Rollo IM, Williamson J, Plackett RL. 1952. Acquired resistance to penicillin and to neoarsphenamine in *Spirochaeta recurrentis*. *Br J Pharmacol Chemother*; 7: 33–41.

Rost B. 1999. Twilight zone of protein sequence alignments. Protein Eng; 12: 85–94.

Roth ACJ, Gonnet GH, Dessimoz C. 2008. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*; 9: 518.

Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol*; 4: 406–425.

Sanger F. 1945. The free amino groups of insulin. *Biochem J*; 39: 507–515.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*; 74: 5463–5467.

Sanger F, Thompson EOP. 1953. The amino-acid sequence in the glycyl chain of insulin 1. The investigation of lower peptides from partial hydrolysates. *Biochem J*; 53: 353–366.

Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A. 2014. TreeFam v9: A new website, more species and orthology-on-the-fly. *Nucleic Acids Res*; 42: D922–D925.

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

Semple C, Steel M. 2009. Mathematical aspects of the 'tree of life'. *Math Horizons*; 16: 5–9.

Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. 2002. PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief Bioinform*; 3: 265–274.

Sleep NH, Bird DK, Pope EC. 2011. Serpentinite and the dawn of life. *Philos Trans R Soc Lond B Biol Sci*; 366: 2857–2869.

Sleep NH. 2018. Geological and geochemical constraints on the origin and evolution of life. *Astrobiology*; 18: 1199–1219.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol*; 147: 195–197.

Sojo V, Dessimoz C, Pomiankowski A, Lane N. 2016. Membrane proteins are dramatically less conserved than water-soluble proteins across the tree of life. *Mol Biol Evol*; 33: 2874–2884.

Sonnhammer ELL, Östlund G. 2015. InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res*; 43: D234–D239.

Steinegger M, Söding J. 2018. Clustering huge protein sequence sets in linear time. *Nat Commun*; 9: 2542.

Storm CEV, Sonnhammer ELL. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*; 18: 92–99.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics*; 4: 41.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*; 28: 33–36.

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

Tekaia F, Yeramian E, Dujon B. 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: A global picture with correspondence analysis. *Gene*; 297: 51–60.

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.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res*; 13: 2129–2141.

Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet*; 7: e1001284.

Tria FDK, Brueckner J, Skejo J, Xavier JC, Kapust N, Knopp M, Wimmer JLE, Nagies FSP, Zimorski V, Gould SB, Garg SG, Martin WF. 2021. Gene duplications trace mitochondria to the onset of eukaryote complexity. *Genome Biol Evol*; in press.

Tria FDK, Landan G, Dagan T. 2017. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol*; 1: 193.

van der Heijden RTJM, Snel B, van Noort V, Huynen MA. 2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics*; 8: 83.

van Dongen SM. 2000. Graph clustering by flow simulation. Utrecht: PhD thesis, University of Utrecht.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*; 19: 327–335.

Watson JD, Crick FHC. 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*; 171: 737–738.

Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF. 2016. The physiology and habitat of the last universal common ancestor. *Nat Microbiol*; 1: 16116.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, *et al*. 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*; 36: D13–D21.

Whitman WB, Coleman DC, Wiebe WJ. 1998. Perspective prokaryotes: The unseen majority. *Proc Natl Acad Sci U S A*; 95: 6578–6583.

Williams TA, Szöllősi GJ, Spang A, Foster PG, Heaps SE, Boussau B, Ettema TJG, Embley TM. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc Natl Acad Sci U S A*; 114: E4602–E4611.

Woese CR. 1987. Bacterial evolution. *Microbiol Rev*; 51: 221–271.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eucarya. *Proc Natl Acad Sci U S A*; 87: 4576–4579.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci U S A*; 74: 5088–5090.

Wolf YI, Koonin EV. 2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol Evol*; 4: 1286–1294.

Zahn-Zabal M, Dessimoz C, Glover NM. 2020. Identifying orthologs with OMA: A primer. *F1000Res*; 9: 27.

Zhang J. 2000. Protein-length distributions for the three domains of life. *Trends Genet*; 16: 107–109.

Zhao J, Grant SFA. 2011. Advances in whole genome sequencing technology. *Curr Pharm Biotechnol*; 12: 293–305.

Zmasek CM, Eddy SR. 2002. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*; 3: 14.

Zuckerkandl E, Pauling L. 1965. Molecules as documents of evolutionary history. *J Theor Biol*; 8: 357–366.