Patterns and causes of gene evolution over long evolutionary distances in prokaryotes

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

Falk Sascha Per Nagies

From Werdohl, Germany

Düsseldorf, May 2023

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 Professor Dr. William F. Martin
Co-supervisor Professor Dr. Laura Rose

Date of the oral examination: 03.07.2023

" **Nothing in biology makes sense except in the light of evolution**"
- Dobzhansky, 1973

Publications themed in this thesis:

Nagies, F. S. P., Brueckner, J., Tria, F. D. K., & Martin, W. F. (2020). A spectrum of verticality across genes. *PLOS Genetics*, *16(11). e1009200.*

Tria, F. D. K., Brueckner, J., Skejo, J., Xavier, J. C., Kapust, N., Knopp, M., Wimmer, J. L. E., Nagies, F. S. P., Zimorski, V., Gould, S. B., Garg, S. G., & Martin, W. F. (2021). Gene duplications trace mitochondria to the onset of eukaryote complexity. *Genome Biology and Evolution*, *13(5). evab055.*

Nagies, F. S. P., Wimmer, J. L. E., Mrnjavac, N., Knopp, M. R., Kapust, N., Gerhards, R. E., Trost, K., Modjewski, L., Bremer, N., Degli Esposti, M., Misrahi, I., Allen, J. F. & Martin, W. F. (2023) Earth's atmospheric oxygen levels were governed by three enzymes and drove biochemical evolution through oxygen inhibition. *Unpublished Manuscript.*

Martin, W. F., Nagies, F. S. P., & do Nascimento Vieira, A. (2021). To what inanimate matter are we most closely related and does the origin of life harbor meaning? *Philosophies, 6(2). 33-52.*

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 while in the candidature for a research degree at the Heinrich Heine University Düsseldorf and has not been submitted in this form or similar form to other institutions. I have not previously failed a doctoral examination procedure.

Düsseldorf, Falk Sascha Per Nagies

Table of Contents:

Abstract

Horizontal gene transfer (HGT) allows prokaryotes to adapt, but it also allows for new lineages to be created by gene influx. Transfers occur even among distantly related groups. As a result, species barriers are not well defined in prokaryotes. Nevertheless, there are vertical components of prokaryotic evolution. These vertically inherited genes, often genes for ribosomal components, are sufficiently resistant to transfer that phylogeneticists can use them as a tool to determine relationships among prokaryotic species. The verticality of genes will be one of the main themes throughout this thesis. It will be shown that more universal genes generally have high verticality.

Contrary to vertical genes like those for ribosome biogenesis, some genes are apparently prone to horizontal transfers among lineages. In some cases, these massive transfers may have been an ecological necessity for lineages to survive shifts in the environment. Such a shift was the Great Oxygenation Event that occurred around 2.4 Ga years ago and altered Earth on a global scale. Photosynthesis by cyanobacteria led to the accumulation of O_2 in the atmosphere. This new molecule was both detrimental and beneficial to many of the existing species since oxygen can be lethal to anaerobic organisms, which all life was before the Great Oxygenation Event, but it also allows for the decomposition of compounds that are otherwise hard to break down. As a result, it is reasonable to expect genes for enzymes that utilize oxygen to have spread quickly among the prokaryotic world and thus show lower verticality than comparable enzymes.

A different way for organisms to adapt is to evolve new genes, which can arise by gene duplications. However, this process is much more common among eukaryotes than prokaryotes. The first eukaryotes are the result of a symbiosis between two prokaryotes, likely an archaeal host and a bacterial symbiont that would later become the mitochondrion. During this transition from two prokaryotes to one entity – the eukaryotes – gene duplications may have played a more prominent role than previously in the evolution of either lineage. Moreover, the original symbionts also were prokaryotes and hence performed HGT. These processes should leave detectable signals in today's eukaryotic genomes by virtue of ancient gene transfers.

But it remains hard to interpret these signals. Long time frames make it hard to interpret all the possible clues. For very ancient divergences approaching the origin of life, the problems of interpreting the signals in gene data become even greater, nearly insurmountable. By taking the origin of life as one example of ancient evolution, it is possible to show that competing hypotheses can have various aspects in common but differ in the timing of their common elements.

Zusammenfassung

Der horizontale Gentransfer (HGT) ermöglicht es Mikroben sich anzupassen, aber er erlaubt auch die Entstehung neuer Arten durch massive Genübertragungen. Angesichts dieser hohen Variabilität von Genen zwischen entfernt verwandten Gruppen sind die Artengrenzen bei Prokaryonten nicht klar definiert. Es wurde jedoch berichtet, dass bestimmte Gene eher vertikal von einer Generation zur nächsten weitergegeben werden. Diese Gene, häufig ribosomale Proteine, sind in der Tat so resistent gegen eine Übertragung, dass die Phylogenetik sie als Instrument zur Bestimmung der Verwandtschaftsverhältnisse zwischen prokaryotischen Arten nutzt. Die Vertikalität von Genen wird eines der Hauptthemen dieser Arbeit sein.

Im Gegensatz dazu sind einige Gene anscheinend anfällig für horizontale Transfers zwischen den Abstammungslinien. In einigen Fällen könnten diese massiven Transfers eine ökologische Notwendigkeit gewesen sein, damit die Abstammungslinien Veränderungen in der Umwelt auf globaler Ebene überleben konnten. Eine solche Veränderung ist das Great Oxygenation Event vor etwa 2,4 Milliarden Jahren. Die Photosynthese der Cyanobakteria führte zu einer Anreicherung der Atmosphäre mit ihrem Abfallprodukt: Sauerstoff. Dieses neue Molekül war für viele der vorhandenen Arten sowohl schädlich als auch vorteilhaft, da Sauerstoff für anaerobe Organismen tödlich sein kann, aber auch eine enorme Freisetzung von Energie und die Zersetzung von ansonsten schwer abbaubaren Verbindungen ermöglicht. Daher kann man davon ausgehen, dass sich Enzyme, die Sauerstoff verwerten, in der Welt der Prokaryonten schnell verbreitet haben.

Eine andere Möglichkeit der Anpassung von Organismen besteht in der Schaffung neuer Gene, die durch Genduplikationen entstehen können. Dieser Prozess ist jedoch bei Eukaryonten viel häufiger als bei Prokaryonten. Es ist bekannt, dass die ersten Eukaryonten eine Symbiose zwischen zwei Prokaryonten waren, wahrscheinlich einem archäischen Wirt und einem bakteriellen Symbionten, aus dem sich später die Mitochondrien entwickelten (Endosymbiose). Während dieses Übergangs von zwei Prokaryonten zu einer Einheit, den Eukaryonten, könnten Duplikationen eine größere Rolle als bisher in der Evolution der beiden Linien gespielt haben. Außerdem wird oft übersehen, dass die ursprünglichen Symbionten ebenfalls Prokaryoten waren und daher wahrscheinlich HGT betrieben. Diese Prozesse sollten aufgrund der ursprünglichen Gentransfers nachweisbare Signale in den heutigen eukaryotischen Genomen hinterlassen.

Doch auch wenn diese Signale nachweisbar sind, bleibt es schwierig, sie zu interpretieren. Lange Zeiträume machen es schwierig, alle möglichen Hinweise richtig zu interpretieren. In der Folge wird es fast unüberwindbar schwieriger, Daten aus noch länger zurückliegenden Zeiten zu

interpretieren: Die Entstehung des Lebens. An diesem Beispiel lässt sich aber auch zeigen, dass konkurrierende Hypothesen zwar immer noch verschiedene Gemeinsamkeiten haben, sich aber in der Ausführung und im Zeitpunkt dieser Gemeinsamkeiten unterscheiden.

Part 1. Theoretical basis of verticality

A closer look at species in prokaryotes

Species are one of the most important units in evolutionary biology. However, what species are, is problematic upon closer inspection (Hey 2001). At first, observation of a species usually involves precise morphological forms occurring in the same time and space (Simpson 1951, Cohan 2002, de Queiroz 2007). While humans have an intrinsic understanding of this species concept, it fails regarding edge cases (Hey 2001, Reydon and Kunz 2019). As such, the morphological grouping of species is sometimes necessary, e.g., in paleontology; however, it is unusable in other cases, e.g., for many prokaryotes (Rosselló-Mora and Amann 2001, Cohan 2002, Gevers et al. 2005). Once this was widely recognized, different species concepts were developed, with one of the most influential being the biological species concept, conceptualized by Ernst Mayr (Mayr 1996, 1999).

The biological species concept focuses on populations considered part of the same species if they can interbreed and create viable offspring (Mayr 1996, 1999, Cohan 2002, de Queiroz 2005). This concept is again helpful for most macroscopic organisms. Yet, this definition fails to explain the emergence and diversification of asexual species (Cohan 2002, Barraclough et al. 2003), with the most common example again being prokaryotes (Rosselló-Mora and Amann 2001, Cohan 2002). Prokaryotes mostly give genes from one generation to the next by cell division but exchange genes between unrelated lineages, a process called horizontal gene transfer (HGT). While HGT frequencies between lineages could allow using the biological species concept in prokaryotes (Kloesges et al. 2011, Skippington and Ragan 2012), these tendencies are difficult, if not impossible, to test *in vivo*. On top of that, gene exchanges can occur over long evolutionary distances (Brown 2003). Nonetheless, if HGT is present, then vertical inheritance of genes also exists (Ochman et al. 2000).

Vertical inheritance is the basis for various prokaryotic species being recognized under pre-defined taxonomic criteria (Fox et al. 1992, Cohan 2002, Oren and Garrity 2021). These criteria show that specific individuals have enough similarity in gene content, lifestyle, and gene sequence identity to be recognized as belonging to the same species (Barraclough et al. 2003). However, how closely related strains of a species are depends on the phylogenetic age of the taxon, with some species being barely different (Kosoy et al. 2012) and others showing high variability more akin to distinct higher taxon (Collins and East 1998, Yu et al. 2019). “That is, species groups may be real, but the species rank is not.” (Wright and Baum 2018). Prokaryotes having species was not

certain historically, as with the continuing discovery of more HGT events, the question of whether prokaryotes can be grouped as species was discussed (Ward 1998, Rosselló-Mora and Amann 2001, Hanage et al. 2005, Wright and Baum 2018). What then ultimately defines a prokaryotic species? For once, one could argue that the singular identity of the ecological role (or niche) defines a species (Fraser et al. 2009, Philippot et al. 2010, Freudenstein et al. 2017, Baquero et al. 2021). Thus, even if gene exchanges occur, the gene origin is irrelevant to the function of the whole genome. But this definition is incomplete in that it disregards individual variation, differences, and lineage-specific differentiation. However, it can help to include niche as a concept to aid in defining species based on lineage-derived patterns, the genetic similarity (Baquero 2015, Reydon and Kunz 2019, Baquero et al. 2021). And early on, it became apparent that related prokaryotic strains have highly similar genetic parts (Wright and Baum 2018).

The growing number of sequenced genomes showed that prokaryotic genomes form a so-called pangenome of variously distributed genes (Medini et al. 2005, Tettelin et al. 2008, Vernikos et al. 2015). This pangenome consists of a so-called core and accessory genome. Related strains had parts of their genetic repertoire present in every strain at hand, the "core" of the species, and a variable genetic part of only a few strains, the accessory genome (Medini et al. 2005, Tettelin et al. 2008). The identity and content of core genes also aided in defining bacterial species (Chun et al. 2018, Wright and Baum 2018). Thus, a valid approach to defining bacterial species is simply comparing genetic sequences and ordering them into species this way (Konstantinidis and Tiedje 2005a, 2005b, Chun et al. 2018). But this general approach can still be difficult for some groups that have human significant strains with high sequence similarity but distinct ecological niches. These would not be grouped into specific species by sequence similarity alone (Kosoy et al. 2012). So, the idea of a bacterial or archaeal species being a cluster of related genomes occupying related ecological niches with a common descent is forming. The vagueness of this definition is a problem of defining species itself (Hey 2001).

These different species definitions still only work under the condition that prokaryotic genomes do not exchange massive amounts of genes, essentially changing the identity of one genome from another (Jiao and Yang 2021). If such huge HGT events happened, especially on the core genome of a species, all pre-established species concepts would fail. Instead of having a specific phylogenetic tree of dividing species, we have a network of genomes that exchange their species identity upon gene exchange (Popa et al. 2011). On a microscopic scale, we can observe this by phylogenetic trees of single genes being different from their reference species trees (Dykhuizen and Green 1991). But many of the more critical bacterial genes seem to show stable similarity to the presumed species tree. Incidentally, these same genes are also used to identify

species (Hansmann and Martin 2000) and often are the only criterium to group newly isolated strains into established species. This grouping is possible because these genes seem to resist the tendency of HGT (Dykhuizen and Green 1991, Ochman et al. 2000, Nagies et al. 2020). In this thesis, they will be defined as vertically evolving genes (Nagies et al. 2020). Thus, species can be recognized based on factors like sequence identity. This is in stark contrast to the prokaryotic ability to exchange genes.

Based on the Linnean taxonomy, researchers classified prokaryotic species into higher-level groups based on their degree of relatedness. This relatedness is often determined by comparing their gene sequences (Faircloth et al. 2012, Oren and Garrity 2021, Nouioui and Sangal 2022). One of the most frequently used criteria is gene sequence identity (Jain et al. 2018, Hedlund et al. 2022). But what is the basis of grouping species into groups? The number of possible sequence variants for a genome of only 1000000 bp is $4^{1000000}$, which far exceeds the number of sequence variants that could have been formed since the beginning of life up until now (Martin et al. 2021). If all of these potential variants existed at once, taxonomy would be impossible because instead of the pre-established 'clearly' defined species groups, there would be a continuum of all possible variants. However, there are gaps in this continuum, allowing prokaryotic species to be distinguished based on gene similarity and ecological factors (Cohan 2002, Reydon and Kunz 2019, Parks et al. 2021).

These species do not form spontaneously. Every species group has an ancestor (Darwin 1859) that either transformed over time into the current state (Malmgren et al. 1983) or split into two different lineages, one of which evolved into the present species. The latter is the basic principle of cladistics (Kluge and Wolf 1993). While this principle has some exceptions (e.g., Hybrid plant species, (Yang et al. 2019)), it still holds that related species must share certain similarities to other species of the same taxonomic rank (Reydon and Kunz 2019). This is known as phylogenetic inertia (Blomberg and Garland 2002). The more distant the past split between two species, the greater the potential variation in their genetic identity. Thus, closely related species share more genetic and ecological similarities, as they have similar gene repertoires and adaptations (Wright and Baum 2018).

This relatedness is the basis for the taxonomic ordering of species into genera, families, orders, classes, phyla, domains, and corresponding sub- and super-groups (Kluge and Wolf 1993, Blomberg and Garland 2002, Reydon and Kunz 2019). One can observe this relatedness as gradients in the similarity between different species, such that species of the same genera are more similar than those of the same class. However, horizontal gene transfer (HGT) can still occur between distantly related groups (Brown 2003). Furthermore, there are many cases where lineages

benefitted from entirely foreign DNA. For example, archaea of the haloarchaeal group gained crucial genetic adaptions due to HGT (Nelson-Sathi et al. 2015). These genes stem from bacterial lineages, meaning a different domain and as far alienated from the archaeal lineage as possible, resulting in the invasion of the archaeal recipients into completely new niches (Nelson-Sathi et al. 2015). As a result, HGT not only changes the gene repertoire of established groups but can also create new groups (Nelson-Sathi et al. 2015). Therefore, understanding the mechanics of HGT is essential to grasp its limits and significance for the evolution of life.

HGT in prokaryotes – an ongoing process

Horizontal gene transfer is common in prokaryotes (Fuchsman et al. 2017, Acar Kirit et al. 2020, Arnold et al. 2021). There are four main mechanisms of HGT in prokaryotes: Conjugation, transduction, transformation, and the use of gene transfer agents by some groups (Ochman et al. 2000). Conjugation and transduction tend to occur within closely related lineages, as plasmids used in conjugation are usually adapted to particular groups (Ochman et al. 2000, Popa and Dagan 2011), while viruses that enable transduction are often adapted to infect a limited range of host species (Bahir et al. 2009, Rodríguez-Beltrán et al. 2021). Gene transfer agents, on the other hand, are produced by cells to facilitate more efficient gene transfer to related cells (Lang et al. 2012). The greatest exception to this pattern of transfers occurring between related groups is transformation.

During transformation, a cell will incorporate environmental DNA into its genome (Ochman et al. 2000). However, the randomness of transformation raises a paradox: Widely distributed genes, like the essential ribosomal genes, should be the most common gene material. If a cell incorporates a new copy of these genes and loses the old copy, lineage identity could be changed or lost. However, as discussed, this seemingly does not, or at least only rarely, occur in nature. Cases of functional gene replacements after long evolutionary distances have been documented (Dykhuizen and Green 1991, Gehring and Ikeo 1999). Hence, the process is possible but is selected against or stopped altogether. How is this regulated, and what genes are affected by this process that seems to preserve vertical evolution? A possible answer could lie in what happens after gene incorporation. For many genes, cells suppress expression after incorporation (Park and Zhang 2012). Hence, cells can specifically select for the benefit of the cell by up- or downregulating a gene's expression. For example, cells may benefit from the expression of antibiotic-resistance genes, leading to a positive feedback loop that eventually results in permanent expression and incorporation of the gene (Park and Zhang 2012, Soucy et al. 2015). In contrast,

no such benefits exist for the expression of genes already present. Instead, the original versions are preserved. There may even be a maladaptation to other interacting proteins in a cell so that there is selection against any new variant (Cohen et al. 2011). Such a selection force is in prokaryotic populations often significant enough for a selective effect as is observable with codon adaption (Gao et al. 2017), which would ultimately mean: Once a gene is present, a second (foreign) copy is not needed.

However, prokaryotes experience high gene flux in each generation due to the loss and acquisition of genes (Vernikos et al. 2015, Conrad et al. 2021). Not all of these gene variants are necessarily beneficial to the host. For example, some may be part of a toxin-antitoxin system (Hernández-Arriaga et al. 2014), a selfish gene variant that duplicates itself despite possible drawbacks to the host (Rodríguez-Beltrán et al. 2021), or simply be neutral if not slightly deleterious (Brockhurst et al. 2019). These variants will be selected against if no other force maintains their presence in the genome (Ohta and Gillespie 1996). Thus, for HGT to benefit a cell, there must be at least some selection benefit for incorporating different genetic variants (Niehus et al. 2015, McInerney et al. 2017). The most beneficial variants will be more prevalent in specific niches and can be identified by sampling various strains of the same species. In addition, some genes are essential for the survival of any cell from a particular species lineage. These are the accessory and core genomes of the species, respectively (Tettelin et al. 2008). The core is a set of genes unique to a species' identity (Wright and Baum 2018), but there is no hard rule on what has to be part of the core. In fact, with more strain sampling, the core genome often is reduced (Medini et al. 2005, Tettelin et al. 2008, Vernikos et al. 2015), and what is essential to a strain depends on genetic background (Beavan and McInerney 2022). On the contrary, each new gene variant in a strain increases the size of the accessory genome of a species (McInerney et al. 2017, Beavan and McInerney 2022).

The core and accessory genomes comprise what is called the pangenome. With the prefix 'pan' (Greek) meaning "whole." It characterizes all gene variants found in one lineage (Medini et al. 2005, Tettelin et al. 2008). Nevertheless, not all of the gene variants have to be beneficial. As outlined above, specific HGT mechanisms, more specifically transformation, can lead to the undirected uptake of any DNA strand that can then be found within the environment of a cell. These do not have to be functional or even expressed (Park and Zhang 2012, Andreani et al. 2017, Brockhurst et al. 2019). Do they then remain in the genome simply by chance (Ohta and Gillespie 1996)? Upon closer examination, most genes in pangenomes are rare, often singletons, meaning genes that occur only once in a sample of related genomes (Medini et al. 2005, Nagies et al. 2020). Such singletons appear only once over the whole strain selection, which suggests that they are less

important for the species overall. If these variants are functional, they are likely restricted to particular niches or strains. Genes with clear selection benefits should be present across several strains of a lineage and may even cross the species barrier into different distantly related lineages (Brockhurst et al. 2019). The core genes, on the other hand, may have been inherited from an ancestor or become lineage-defining after an initial HGT event (Nelson-Sathi et al. 2012, 2015). Thus, constant HGT among lineages creates the vast prokaryotic pangenome. Only a select few genes seem present in all genomes due to vertical inheritance.

Verticality as a concept - and as a mathematical unit

Contrary to horizontal gene transfer, genes are duplicated during cell division and are transmitted to the next generation by vertical gene inheritance. By default, each daughter cell has a nearly identical copy of the parental genome. Mutations of genes result in natural variation for Darwinian evolution (Koonin and Wolf 2009). The problem with this process in isolation is that mutations accumulate in genes over time. One of these mutations might not be enough to be deleterious, but in a changing environment and with enough slightly deleterious mutations, a non-recombining lineage encounters mutational burden termed genetic load. This process is called Muller's ratchet (Haigh 1978). It is one of the reasons why HGT is advantageous across prokaryotic lineages, which lack sexual recombination found in eukaryotes. Selection after HGT has several effects. First, genes readily exchanged between groups should be present in several higher-level taxonomic groups. In this case, the ecological benefit creates the wide distribution of the gene. Genes that are only useful in specific niches should be present in the few groups that occupy these niches, regardless of the genes' origin. Genes that are very important for the cell, like ribosomal genes, are already widely distributed due to them having some of the most fundamental functions needed in a cell (Hansmann and Martin 2000). However, this also means that these widely distributed genes provide more gene material so that there is more opportunity for random HGT events even across lineage barriers.

Left unchecked, HGT would disintegrate a species-tree over time. But this is not observed. Instead, there exists a core of genes that preserve species and lineage identity. These vertically evolving genes should preserve monophyly more often, while genes that undergo HGT should have gene trees that contradict the species tree. More specifically, monophyly in these will decay over time. This creates a basis for a unit that describes the verticality of the gene. This verticality measure is defined by the number of monophyletic taxonomic groups, which also means these groups must be determined beforehand. For this, higher taxonomic ranks like phylum, class, or

order – depending on the dataset – are applicable. Once groups are defined in a set of phylogenetic trees, each monophylum can be determined and counted to create a raw version of the verticality measure. If no transfers occur, then this basic verticality value V_B would have a maximum equal to the number of pre-defined groups present.

$$V_B = \sum_{1}^{n} M_T$$

with M_T as the count of monophyletic taxonomic groups in a tree.

However, for given a sample of several strains from which the protein sequences at hand are derived, some strains may not have all the genes in question (Beavan and McInerney 2022). That means that for some protein sequence trees, not all strains of a group are present – if the group is represented at all. An easy way to work around this is by summing up the fraction of the whole group in the tree if the respective taxonomic group is monophyletic. This creates a more robust measure that also considers the higher likelihood of a group being monophyletic if only a fraction of the whole group is present. With this, we can derive the following components to create a verticality measure $\boldsymbol{V}$:

$$\boldsymbol{V} = \sum_{1}^{n} P_T$$

With:

- $\boldsymbol{V}$: The verticality index
- P_T : The relative proportion of the taxonomic group in the tree

To utilize the verticality measure, it is necessary to group protein sequences into homologous groups. In more extensive datasets, this becomes an impossible task to do manually. Hence, algorithmic approaches were implemented that often use the inherent similarity of related sequences to cluster them (Enright et al. 2002). This similarity can come from common descent but is also preserved due to the same function since genes or the derived proteins can be classified as complex units performing specific functions in a cell. The proteins are defined by their amino acid sequence (Wetlaufer and Rustow 1973). However, the individual amino acids are not as crucial as the three-dimensional structure created by the protein *in vivo* (Wetlaufer and Rustow 1973). These structures are very specific for certain functions; they can be classified as domains, meaning particular patterns that reoccur in proteins of the same function. The combination of

domains and close phylogenetic relationships makes it possible to create groups of related protein sequences: the protein family (Dayhoff et al. 1978). Accordingly, it is reasonable to calculate the verticality measure with a clustering of protein sequences from several high-quality genomes. In this thesis, an established clustering was used based on the genomes of the RefSeq database (O'Leary et al. 2016). In the following, the theory behind the clustering process will be presented as it bears upon the verticality measure.

Part 2. Processes of protein sequence clustering used in this thesis

Finding homologs among sequences

The first step to grouping protein sequences into protein families is identifying homologs. However, the term homology is rooted in morphological research and can be ambiguous with molecular data (Patterson 1988). First used by Richard Owen (1848) on morphological patterns, homology broadly describes two independent evolutionary structures with the same origin — simply by virtue of common ancestry. The forelimbs of different animals are a classic example; all are homologous structures inherited from their ancestor, but some evolved into wings, others into legs, arms, and flippers (Owen 1848). However, converting this concept to molecular data becomes problematic since genes may be present in duplicates in any genome. Thus, the homologous sequences between genomes can also be homologous to genes in the same genome.

Different terms were introduced to differentiate the different homology cases in molecular data (Patterson 1988). First, homologous genes between genomes that are equivalent units to each other are coined as orthologous sequences. This is based on the haploid genome and differentiation of species (or other taxonomic units) (Patterson 1988). Paralogous sequences are the second form of homology (Fitch 1970), which include homologous sequences related by gene duplications. It is important to note that two sequences in the same genome are paralogous, but each sequence might have an orthologous sequence in another genome. It is also possible that only one paralogous has an orthologous sequence in another genome. This is determined by the relative time point of the duplication event and the lineage split between the different genomes. If the duplication occurred after the lineage split, no orthologous gene is present in older lineages for the new paralog.

Another special case is the so-called xenolog. These homologous sequences are created in different lineages via HGT and are thus of particular interest in prokaryotic genomics. Detecting HGT events makes it possible to infer xenologs instead of orthologs between genomes. However, reliable and automatic HGT detection is difficult to perform. In many cases, the necessary gene data to differentiate a single HGT event from orthology between distantly related genomes is lacking. Instead, orthologs can be determined, and xenologs can then be annotated in a later step. Considering this, the similarity between two sequences is not enough to postulate orthology but rather a necessary criterion (Patterson, 1988; Stevens, 1984). Since the relative timing of duplications that create paralogous sequences varies, higher similarity should be expected between true orthologous sequences, even if several sequences show high similarity (Patterson 1988). Additionally, there should be congruence in the descent between orthologous sequences (Patterson 1988). This pattern allows the postulation of orthologs by crosswise comparison, and the orthologous pairs should show higher similarity between each other than the paralogous versions. The reciprocal best hits approach implements this concept (elaborated below). Coupling this with a subsequent clustering based on similarity matrices of sequences allows for reliable ortholog determination and grouping of protein sequences into protein families (Dayhoff et al. 1978).

Creating a protein similarity matrix for the clustering of protein sequence data

The clustering of data is the algorithmic process of ordering specific objects based on traits and similarities into distinct groups (Enright et al. 2002). Over the last decades, it became possible for increasingly complex networks to be ordered (Enright et al. 2002, Funahashi et al. 2008, Lercher and Pál 2008, Popa et al. 2011). One of the first described clustering processes was k Means which uses the ordering of elements into *k* groups while minimizing means between comparison values in these groups (Macqueen 1967). Most of these early described algorithms can be done by hand, which is often doable in a timely manner for small datasets. However, with the development of computers, the algorithmic clustering processes have become faster and more efficient. Starting with this, people developed and searched for new clustering methods, and in modern times machine learning has taken up the spotlight. The algorithm used for the clustering in this thesis has a unique history: It was “discovered” by accident and later adapted for biological sequence data (van Dongen 2000). However, as input for these algorithms, the similarity between sequences must first be determined.

The most widely used tool to find similar sequences is the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990), which compares distinct patterns of sequences to find the

pairs that are the most similar (Altschul et al. 1990). The BLAST algorithm begins with a database of sequences (DNA or protein) and an input sequence. It ends with local alignments of the input sequence to all database sequences that show similarity over a pre-defined threshold. As the final output, BLAST shows all database pairs and alignment results for each query to all database sequences (hits) lower than pre-defined thresholds, like the e-value (a measure approximating how likely a hit was with a given database size). The calculated alignment is done with the Smith-Waterman algorithm (Smith and Waterman 1981). This process is much faster than calculating alignments of *n* sequences to each other to find the most similar sequences. The first step in the utilized clustering pipeline was to BLAST all sequences against each other to find the most similar pairs.

These pairs are usually filtered by one or several quality values of the calculated alignments (alignment length, identity between alignments, e-value; here, an e-value of 10^{-10} (1e-10) was used as a filtering threshold). Nevertheless, these different hits still include paralogous sequences. The difference is subtle as all sequence pairs are related, but only orthologous sequences are the most meaningful pairs. Depending on the requirements of the pipeline, it might be better to filter for orthologs; sometimes, this is only done to reduce data size further. In prokaryotic data, one technique can be used for fast orthologous sequence detection with high quality: the aforementioned reciprocal Best Hits (RBH) (Moreno-Hagelsieb and Latimer 2008). In this algorithm, the best hit (sequence 1) for each query sequence (sequence 2) is compared to the best hit of the original best hit (sequence 1). If the different sequences have each other as their best hit, they are seen as orthologous sequences. It is important to note that for this to work, the query and hit sequence comparison must be done on a genome-against-genome basis. It is only for direct genomic comparison that RBH finds orthologous sequences (Moreno-Hagelsieb and Latimer 2008).

For many pipelines, these different hit pairs can already be used as input for clustering algorithms. However, this does not consider problems generated by utilizing local alignments in BLAST and several unique properties in biological sequence data. Most importantly, even with several completely different functions, protein sequences have overlapping sequence motifs due to folding and functional constraints. These motifs, called protein domains, are distinct patterns occurring in protein sequences with a specific function, e.g., DNA binding (Berg 1990). Proteins involved in different processes and completely unrelated can therefore have similar local subsequences. The RBH method should filter these hits out, but this might fail in rare cases due to data gaps. Furthermore, due to the randomness of character states, similar regions between sequences might be created by chance. Especially in low identity alignments, the possibility of

false positive hits rises. This identity threshold is also called the twilight zone of protein sequence alignments (Rost 1999).

However, false positive hits can be further avoided by introducing a step of calculating a global alignment of sequence pairs. Global alignments consider all of the information between two sequences and thus lower the chances of two sequences having a high identity due to random chance (Needleman and Wunsch 1970). By filtering all of these alignments for their identity, it is presumably possible to avoid most false-positive hits and get a thorough list of orthologous sequence pairs. These pairs and their identity can be given as a similarity matrix to a clustering algorithm. To cluster the final data amassment used for the works in this thesis, very stringent filter thresholds had to be implemented. The resulting implementation will be described below.

Specifics of protein sequence clustering of the RefSeq database

Different algorithms can cluster different protein sequences into usable protein families. One such algorithm that is very popular due to its robustness is the Markov Cluster Algorithm (short: MCL) (Enright et al. 2002). However, the results of this are highly dependent on the data that is used. For the calculation of the clustering used for the verticality calculations in this thesis, a dataset based on the genomes from the RefSeq 2016 September release was used (O'Leary et al. 2016). This dataset encompassed 5,655 genomes spanning several bacterial and archaeal phyla and was clustered by Julia Brückener as part of her PhD thesis work (Brückner 2021). These phyla were sorted into 40 taxonomic groups. For this, the Proteobacteria and Firmicutes were split into classes due to overrepresentation and the Archaea into orders due to the underrepresentation of diversity in the dataset. Each group had to have at least five genomes to be viable in the analyses. This way, comparable taxonomic units were defined in the dataset.

The MCL algorithm cannot use the 19 million protein sequences of the dataset directly as input. Instead, the relations between the sequences are used. As described above, homologous sequence pairs must be identified first to create these relations. For this, the RBH method was needed (Moreno-Hagelsieb and Latimer 2008), which included a filtering process of an e-value of 1e-10 (Altschul et al. 1990, Camacho et al. 2009). All sequence pairs were filtered again with a cutoff of local identity of at least 25 percent between pairs. This bypasses the 'Twilight zone' of sequence identity (Rost 1999). More specifically, sequence pairs with less than 25 percent sequence identity are likely false positives. In the end, all of these filtered sequence pairs were given as input for the MCL algorithm. To understand how protein sequence groups are created,

looking deeper into how the algorithm works is essential. The Markov Chain algorithm is called so because it utilizes so-called Markov matrices, which will be explored in the following passage.

MCL: The algorithm for protein sequence clustering

The MCL algorithm has become one of the central tools in protein sequence clustering (Enright et al. 2002). It is an iterative process that uses a network of edges between protein sequences coded as nodes, and from these nodes, a final disjunct network (or graph) is created. Every single subunit of the network is a cluster that, by being based on similar protein sequences, is hypothesized to be a protein family. Enright et al. (2002) and other studies (Bernardes et al. 2015) showed that the protein families postulated by MCL are robust. To understand how this clustering works and its potential pitfalls, it is necessary to dissect this iterative process. First, the matrix of nodes and edges has to be created. The general structure of this stochastic matrix is called a Markov matrix in mathematics; specifically, the numbers of each column are normalized beforehand to sum up to one. Enright et al. (2002) used the e-values of protein sequences taken from the all-against-all BLAST matrix as edges. But in the clustering based on the RefSeq2016 database, global identities of protein sequences were instead utilized. This matrix is then the starting point of the algorithm, which is more refined with every iteration that alters the edges into being stronger or weaker until they are finally pruned. Edges are changed in two steps called expansion and inflation. During expansion, the matrix product of the Markov matrix is calculated. This way, connections of two separate nodes are weighted against each other. The second step of inflation calculates the Hadamard power of this matrix, which means taking the power of each singular entry. Finally, the matrix is scaled again so that a stochastic matrix is reacquired.

MCL works because of the underlying cluster structure already present in the connectivity graph through unequal weight distribution of graph edges. Subsequently, the clustering process enhances these. For example, the expansion step is often described as ‘random walks’ through the network. During these walks, it is more likely that intra-cluster connections are taken than inter-cluster connections; thus, the cluster structure is amplified. Following this, the Inflation step amplifies all resulting connections: strong connections get relatively stronger, while weak connections weaken. The strength of this is given by the inflation operator that can be defined beforehand. This corresponds to the number by which the Hadamard Product of the Matrix is calculated (e.g., power of 2, power of 3, etc.). The stronger the inflation operator is, the more amplified the present disparities of weak and strong connections are. A higher inflation operator leads to higher cluster granularity as weaker connections are lost. As Enright et al. (2002) put it:

"Inflation will then have the effect of boosting the probabilities of intra-cluster walks and will demote inter-cluster walks." After the recovery of a stochastic matrix due to the scaling of the acquired values, the process starts anew. Several rounds of both steps quickly lead to an equilibrium state in the matrix, after which no discernible differences are created anymore. Experiments have shown this to apply after three to ten rounds (the graph converges to this equilibrium) (Enright et al. 2002). But it is difficult to prove an actual equilibrium state (meaning global convergence of all values with further iterations) so that an alternative end condition is defined. If the input graph of a new iteration (expansion and inflation) is symmetric, then the matrix is assumed to converge (van Dongen 2000) and the algorithm ends.

The clustering of protein families is not perfect

The clustering process orders protein sequences into putative protein families. To understand when and how the process could fail, it is best to examine specific extreme cases that illustrate how protein families are ordered in nature. A protein sequence does not exist in a vacuum. There are always closely related sequences in the natural world. This relatedness is always present since either other cells of the same lineage or different lineages include copies of the gene, often with slight to extensive alterations. These relations are at the heart of determining the homology of protein sequences (Patterson 1988). How significant can the differences between two such sequences be? As an example, one could create an idealized prokaryotic gene. It has around 1000 base pairs, translating to roughly 333 amino acids for the associated protein sequence. Between any two variants, two of these protein sequences can differ in any of the amino acids. Furthermore, all 333 amino acids in one sequence can be one of 20 states, the usual proteinogenic amino acids. This means there are 20^{333} possible combinations possible. For comparison, there are approximately 10^{80} protons in the universe. Moreover, the specific amino acid can also be more or less similar to each other. For example, aromatic amino acids are more similar to each other biochemically than polar amino acids (Woese 1965).

These differences allow testing for the relatedness of protein sequences. Only two differences are closer related than four differences, at least if mutation rate differences are disregarded. The clustering process uses these relations to see which groups of sequences seem to form natural 'clouds' in the space of all possible sequences. But what happens if all possible sequence combinations exist in a sample? In this case, all possible connections are present, and the algorithm cannot differentiate between different subgroups of protein sequences. Thus, all groups are presented as one big cluster. In order to 'create' clusters from this sample, it would be

necessary to delete interconnecting sequences until subgroups start to separate from one another. This happens naturally due to factors like the relation of sequences, the selection to keep specific functions, and the diversification force of mutations (Rost 1999, 2002, Enright et al. 2002). Not all possible sequences are viable, and it is even the case that not all sequences exist that would theoretically be possible due to the sheer number of possibilities (Martin et al. 2021). Thus, even in bigger sequence samples from nature, we can always extract cluster or putative protein families from these groups. Even more so because sequences can have different lengths and gaps that further distinguish them.

But any sample we take is also limited. As there can never be a sample encompassing all existing variation, gaps within a cluster naturally exist. If the samples are complete enough, these gaps should not affect the grouping of sequences. However, in practice, this is not always the case. Particular old subgroups could exist for some protein families that differentiated so long ago that they seem to be completely different clusters, e.g., ribosomal genes from bacteria and archaea. On the opposite side, some clusters that are, in practice, completely distinct could be combined into one big but false cluster. There are several reasons why this could happen. First, sometimes genes fuse, creating new genes with new functions, or even just one protein sequence being more efficient by performing both functions at once (Snel et al. 2000). Yet, these fusions create direct links between distinct clusters. Second, re-occurring sequence patterns can be observed in protein sequences (Rost 1999, Ponting and Russell 2003). These usually include domains that perform specific tasks in the function of the protein (Snel et al. 2000, Ponting and Russell 2003). But domains are not exclusive to one protein and may reoccur in other protein sequences. These domains can be due to the selective force being very similar in otherwise dissimilar proteins, which could again lead to false connections between unrelated protein sequences. A way to mitigate this is by including a global alignment filter step. Third, if sequences did not diverge long ago, the interconnection of these may still be strong enough that a clustering algorithm recovers these as only one clustering object.

In short, the clustering process is not perfect. However, benchmarking of the resulting quality shows that these processes are still reliable (Bernardes et al. 2015). Especially in a large data set, the error can get neglectable small if natural biases in the data set are considered (Bernardes et al. 2015). However, in certain situations, it might be beneficial to try to combine clusters. For example, the number of independent origins or protein families in ancient genes might be smaller than anticipated from the cluster count of proteins with the same potential function. This is due to the divergence of sequences and missing links between distant sequence variants. This was implemented in the works of this thesis to cluster highly divergent sequences of

eukaryotes and prokaryotes. The reciprocal best cluster approach was utilized for this, as explained in the next chapter (Ku et al. 2015).

Generating Eukaryotic-Prokaryotic-Clusters

When clustering sequences of prokaryotes and eukaryotes origin, the process can be problematic because of many homologous but distantly related sequences. For example, ribosomal protein sequences strongly diverge between these groups, and due to lacking interconnections, two or more singular clusters instead of one large cluster might be created for any ribosomal protein. In addition, eukaryotes have archaeal ribosomes in the cytosol and bacterial ribosomes in mitochondria and chloroplasts, creating additional copies of distantly related genes within the eukaryotic genome. To make matters worse, the unpredictability of sequence evolution and trade-offs during the clustering process may produce this effect for some protein families while others are represented by one singular, clear cluster. One solution might entail an approach of combining generated clusters on the basis of sequence similarity post-clustering. But, since sequence dissimilarity was the original criterion for the protein family falling apart into distinct clusters, a cutoff of arbitrarily low level might be necessary for this approach, with the consequence that clusters are erroneously combined. Without secondary criteria (which can also be subject to human error), such an approach of combining protein clusters to account for high-variability protein families can leads to protein family size inflation.

Another solution is to cluster the sequences of each domain separately and then only combine sequences that meet specific stringency criteria of similarity between clusters. This approach was implemented by Ku et al. (2015) and utilized in the works of this thesis to create Eukaryotic-Procaryotic-Clusters, hereafter called EPCs. First, bacterial, archaeal, and eukaryotic sequences were clustered individually with the same procedure already outlined to cluster sequences (see: “Specifics of protein sequence clustering of the RefSeq database”). In eukaryotes, the global identity cutoff was raised to 40%, which creates fewer false positive connections due to domain overlap and gene duplications in eukaryotes (Ku et al. 2015, Nagies et al. 2020). For prokaryotic clusters, only those with at least five sequences were kept for the inter-domain cluster fusion. Many of these smaller clusters are created by the diversity of sequences in the prokaryotic world and cannot be uniquely assigned to a functional category, diminishing their utility for prokaryotic-eukaryotic comparisons. As a solution, the best-hits of BLASTs of eukaryotic sequences against the prokaryotic sequences were compared. If at least 50% of the sequences from a eukaryotic cluster had the sequences of the prokaryotic cluster as a best hit, it was marked as a

potential EPC. If then at least 50% of the sequences from this prokaryotic cluster had the sequences from the same eukaryotic cluster as a best hit, then these clusters were fused into an EPC (a reciprocal best cluster approach). This procedure was done separately for eukaryotes vs. bacteria and eukaryotes vs archaea, but if one eukaryotic cluster fused with one of the bacterial and one of the archaeal clusters, all three were fused. This is the reciprocal best clustering approach of Ku et al. (2015).

Assignment of functional categories to protein family clusters

Before genes can be compared for different patterns regarding their function, they need to be annotated with functions. This can be done based on the clustering used in this thesis. All of these clusters are putative protein families. Hence, comparing sequences with established functions to the clustered sequences makes it possible to assign functions to clusters. As a source of sequences with annotated functions, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al. 2016) can be used. The Kanehisa Laboratories established KEGG in 1995 and have since regularly updated the database. By downloading the sequences of KEGG and blasting them against the original RefSeq sequences used in the clustering, it is possible to obtain homology relationships to sequences with functional annotation. This means again that the BLAST results have to be filtered, this time for an e-value of 1e-10. It can then be determined how many sequences have similarities to sequences in the KEGG database for each cluster.

KEGG sequences usually include a so-called KO-number (Kegg Orthology), which provides information about reactions, pathways, and functions that a gene involves. Thus, a cluster can be assigned with a specific KO to determine the cluster function/annotation through the relation of KEGG sequences. A KO can be determined for each sequence in a cluster, from which the total number can be summed up. Because enzymes and sequences can have different similarities and overlapping domains, and because KOs in KEGG have redundancy, one specific enzyme can have several KOs, and sequences of the same cluster can be assigned with different KOs. Therefore, annotating a cluster with all possible KOs is possible. Nevertheless, assigning only one KO per cluster with a majority approach has proven more efficient by reducing redundancy. The KO with the highest assignments to a cluster is the annotation of the cluster, but only if a certain percentage of the sequences in a cluster have an assigned KO to ensure the correct annotations. This threshold was set to 25% in the investigations of this thesis.

Aim of this thesis

With the help of information contained in all genes from a large sample of genomes, the goal of this thesis was to address three central questions of early microbial evolution from the perspective of gene clusters.

The first question concerns the factors that govern the degree to which genes are transferred across prokaryotic genomes in evolution. For this, clusters were used to generate trees for all genes in a sample of 5,655 prokaryotic genomes, from which values of verticality were obtained. The values of verticality were found to correlate significantly with the density of distribution for a given prokaryotic gene, indicating that the presence of a preexisting gene impedes the frequency at which transferred genes become fixed in genomes and incorporated into the descendant lineage. The EPCs also enabled this approach to investigate the contribution of the free-living progenitors of mitochondria and plastids to the eukaryotic genomic lineage. This is the topic of the first paper in this thesis (Nagies et. al. 2020).

The second question concerns a fundamental difference in the mechanisms by which eukaryotes and prokaryotes generate sequence variation. In prokaryotes cell division is clonal whereby LGT introduces novel sequences that can be deleted or fixed; recombination is unidirectional from donor to recipient. In eukaryotes, novel sequences are introduced via reciprocal recombination at meiosis, the pairing of homologous chromosomes brought together by gamete fusion, and by gene duplications. In prokaryotes gene duplications are rare, in eukaryotes they are one of the main sources of new genes. Using trees generated from eukaryotic clusters permitted investigations into the role of gene duplications at the onset of the eukaryotic lineage. It was found the genes contributed by the mitochondrion were more frequently duplicated in the genome of the last eukaryotic common ancestor than genes from the archaeal host, indicating a role for duplicative transfer during the relocation of genetic material from the mitochondrial to the host genome. This is the topic of the second paper of the thesis (Tria et al., 2021).

The third topic concerns the role that molecular oxygen, O_2, played in the evolution of prokaryotic genomes. For this, the clusters and trees that map to O_2 dependent enzyme reactions in KEGG were identified and compared to genes and trees for O_2 independent reactions. The working hypothesis was that O_2 dependent reactions were rapidly transferred in the wake of the Great Oxygenation Event that occurred 2.5 billion years ago. Such an effect was detected. The main findings were that O_2 enabled substrate utilization and biosynthesis in environments where O_2 inhibited the preexisting anaerobic pathways. The inhibition of enzymes by O_2 likely exerted major impact on the ability of organisms to inhabit O_2 containing environments. This is the topic of the third paper of the thesis (Nagies et al., unpublished).

Looking back in evolution, prokaryotic life has a single origin and this raises the question of how and where life might have arisen. The fourth paper of the thesis (Martin et al. 2021) presents a comparative and philosophical perspective on that topic, focusing on the role that the environment — the physical setting in which life is posited to have arisen — plays in different theories for the origin of life.

Publications

Publication 1:

A spectrum of verticality across genes

Authors: Nagies, F. S. P., Brueckner, J., Tria, F. D. K., & Martin, W. F
Published: 2020 in *PLOS Genetics*, *16*(*11). e1009200.*

Contribution of Falk Sascha Per Nagies

Shared first author

The theory of the verticality value was conceptualized and tested by me. I then calculated the measure for all possible of the present protein sequence clusters. Following that, I performed a complete analysis of the data. During this I also annotated the clusters, calculated various additional measures, and compared sister groups. Finally, the Manuscript was written and edited with the shared first and last author.

PLOS GENETICS

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

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

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

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.

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 ***V*** for that tree or cluster. For *n* phyla, ***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 ***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 ***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 ***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 ***V*** encode ribosomal proteins, which have many interactions, many ribosomal proteins have modest values of ***V***. We found that the majority of highly vertical genes are soluble proteins as opposed to being components of macromolecular complexes, and that verticality ***V*** strongly correlates with the gene's distribution frequency across genomes, as shown in **Fig 1**, where the value of ***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

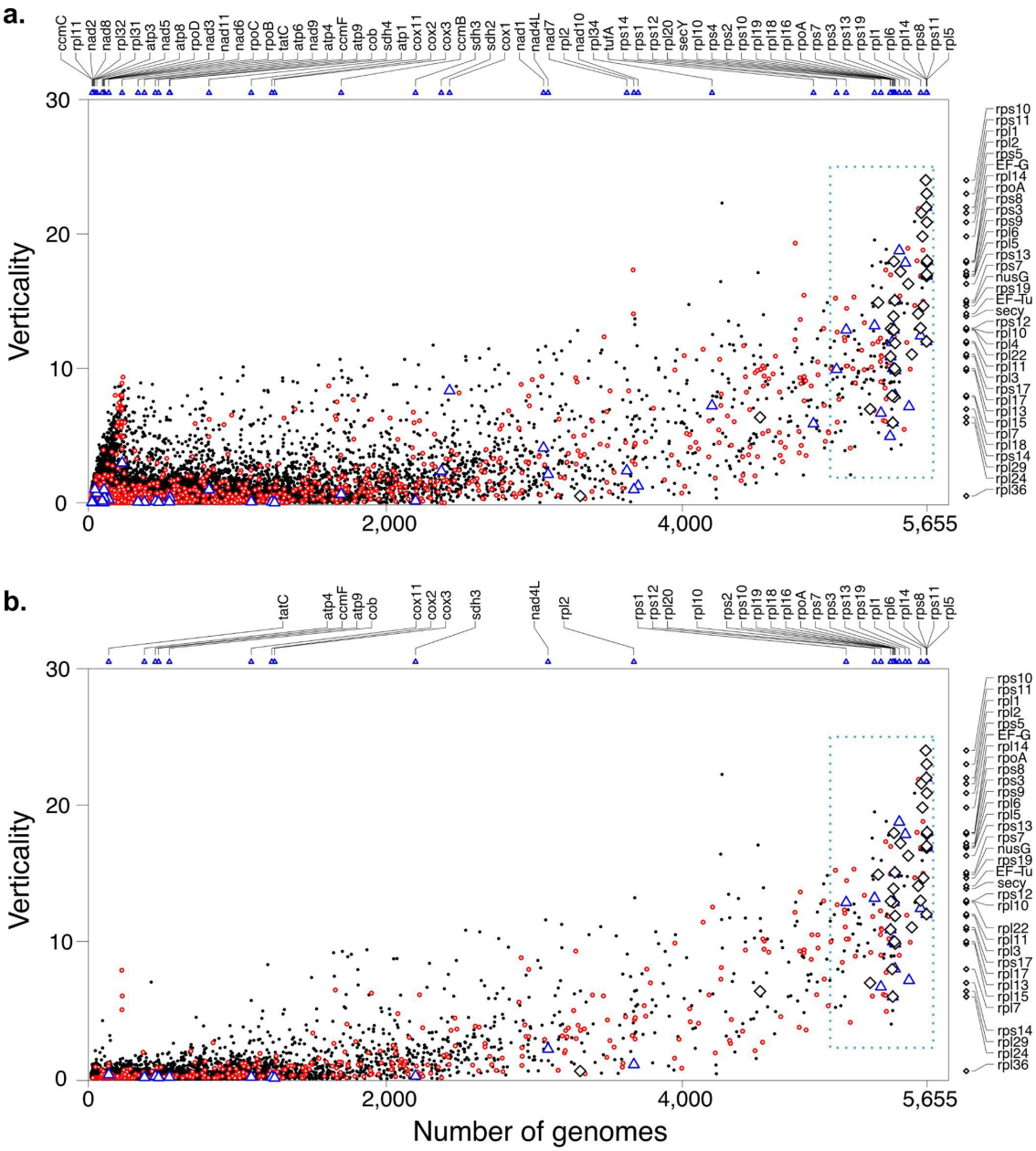


Fig 1. Comparison of estimated verticality and number of genomes in a protein cluster for **a.** all clusters (n = 101,422) and **b.** all conserved clusters (average branch length $\geq$ 0.1; n = 8,547). Unrooted trees were analyzed if at least two taxonomic groups were present. Verticality was calculated 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 $\boldsymbol{V}$ 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, $\boldsymbol{V_w}$; 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 $\boldsymbol{V_w}$ 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

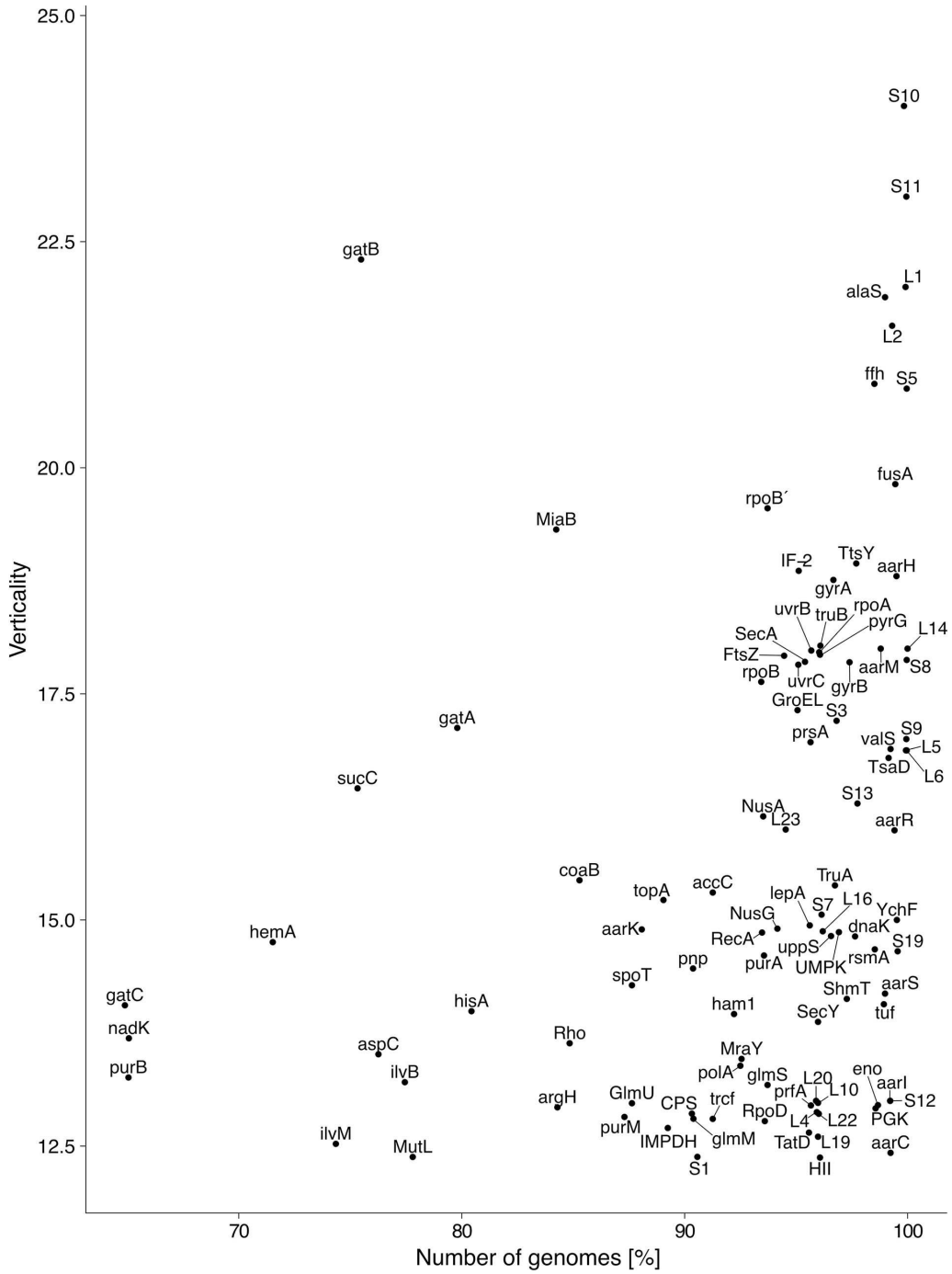


Fig 2. Comparison of estimated verticality and number of genomes [%] for the 100 most vertical clusters. Identity and Annotation of clusters can be found in S6 Table. This is a representation of some of the clusters shown in the blue rectangle of Fig 1A.

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 ***V***, an indication that our measure of ***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, 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

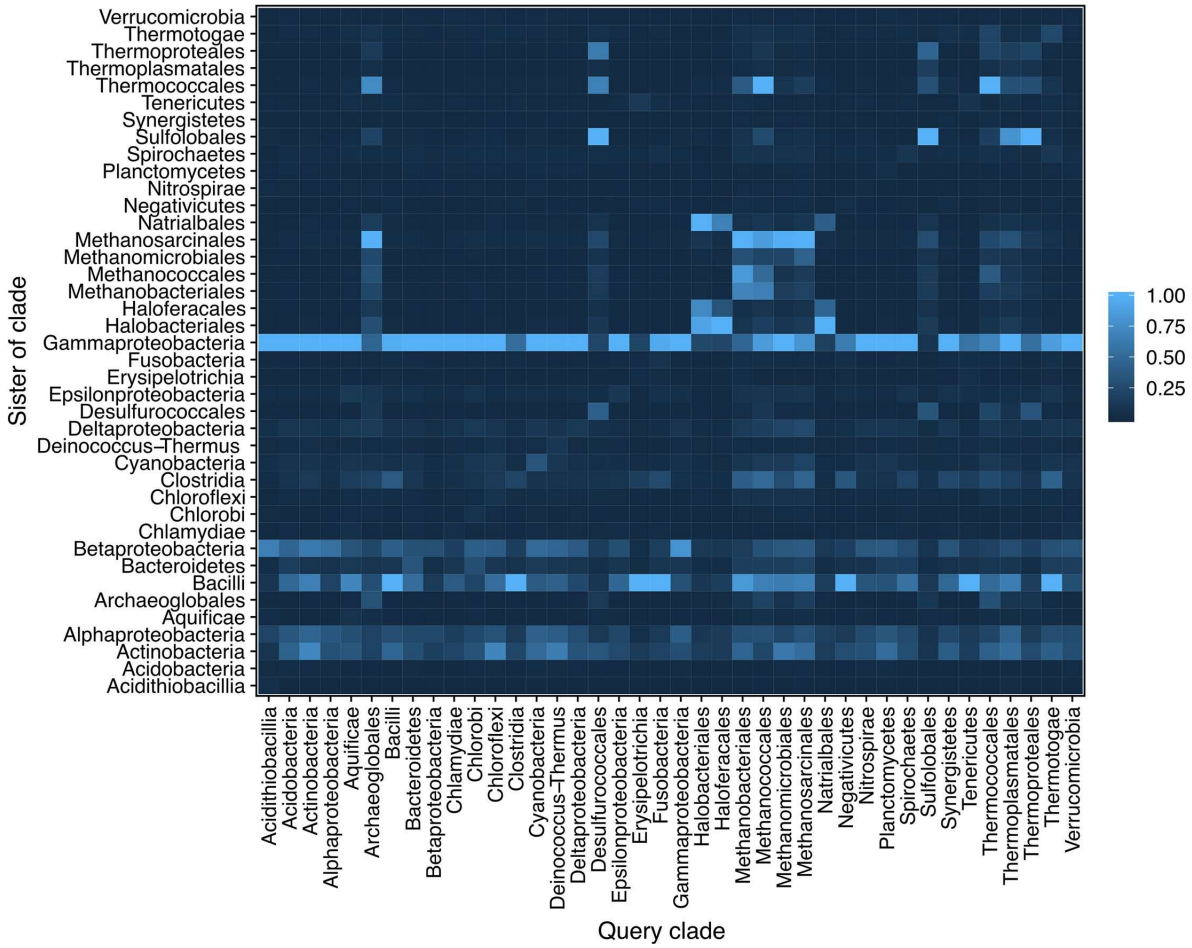


Fig 3. Relative occurrence of a taxonomic group as the sister group of each clade in the unrooted trees. For each taxonomic group in a cluster the sister was determined and counted. Multiple occurrences of different groups in the sister were accounted for by their relative occurrence. If the taxonomic 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

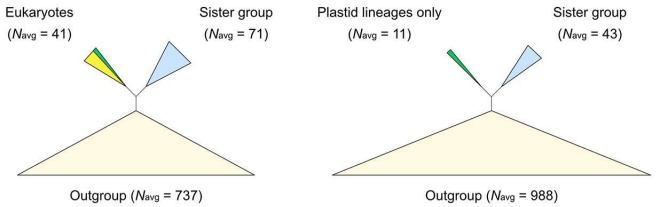
In an ideal world of vertically inherited genes and infallible phylogeny, all genes would produce the same tree and all eukaryotic genes would trace to the same alphaproteobacterium (the mitochondrion) and the same are archaeon (host), plus the same cyanobacterium in the case of eukaryotes with plastids. 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

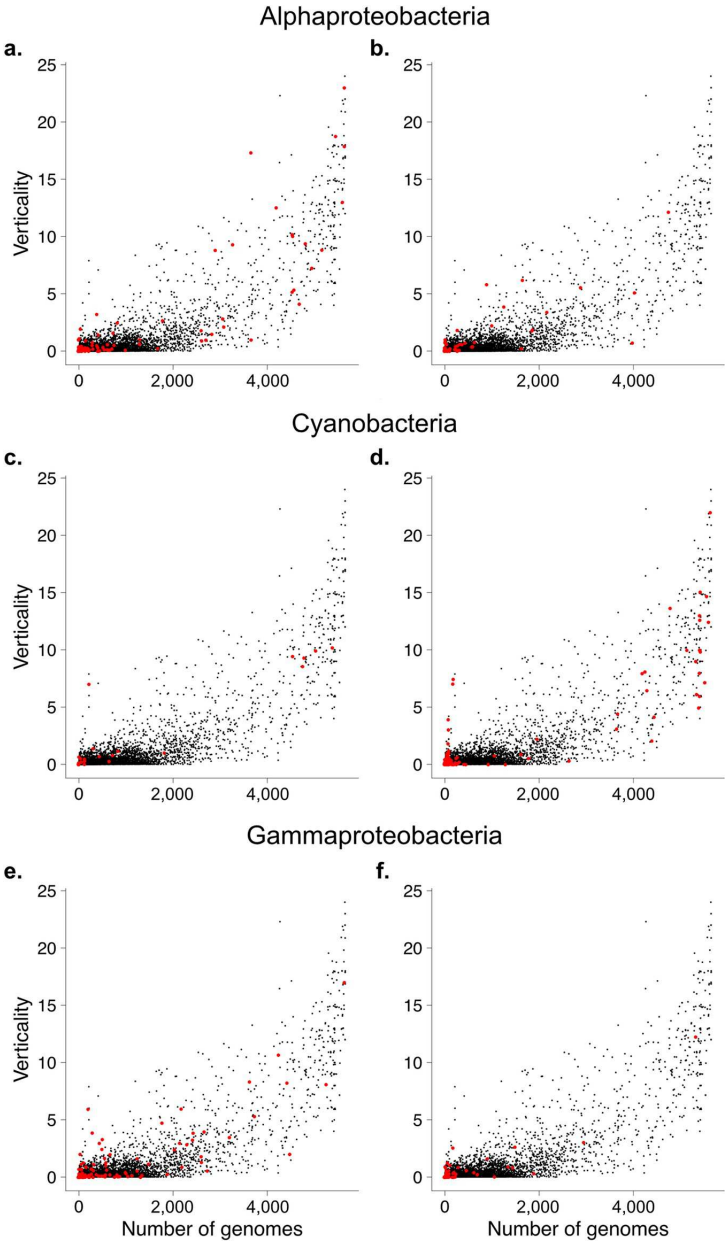


Fig 5. Mapping of EPCs to prokaryotic clusters. The EPCs were separated according to the pure sister group of eukaryotes in the unrooted trees: **a.** and **b.** Alphaproteobacteria, **c.** and **d.** Cyanobacteria, **e.** and **f.** Gammaproteobacteria. The left panel shows EPCs that may include all 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

Publication 2:

Gene Duplications Trace Mitochondria to the Onset of Eukaryote Complexity.

Authors: Tria, F. D. K., Brueckner, J., Skejo, J., Xavier, J. C., Kapust, N., Knopp, M., Wimmer, J. L. E., Nagies, F. S. P., Zimorski, V., Gould, S. B., Garg, S. G., & Martin, W. F.

Published: 2021 in *Genome Biology and Evolution*, *13(5). evab055.*

Contribution of Falk Sascha Per Nagies:

During this work I performed a thorough analysis of the sister groups of the given eukaryotes.

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. Martin[1]

[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.

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

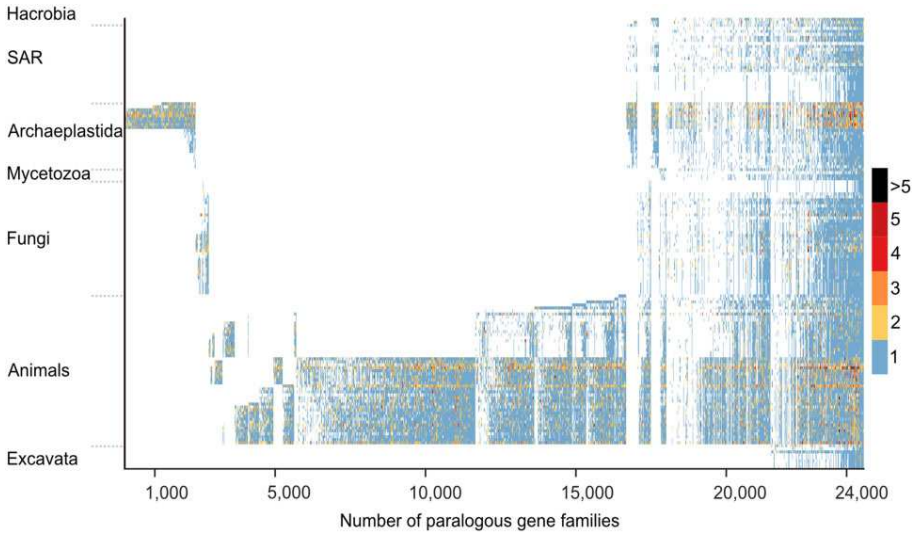


Fig. 1.—Distribution of multicopy genes across 150 eukaryotic genomes. All eukaryotic protein-coding genes were clustered, aligned, and used for phylogenetic inferences. The resulting gene families present as multiple copies in more than one genome are plotted (see Materials and Methods). The figure 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

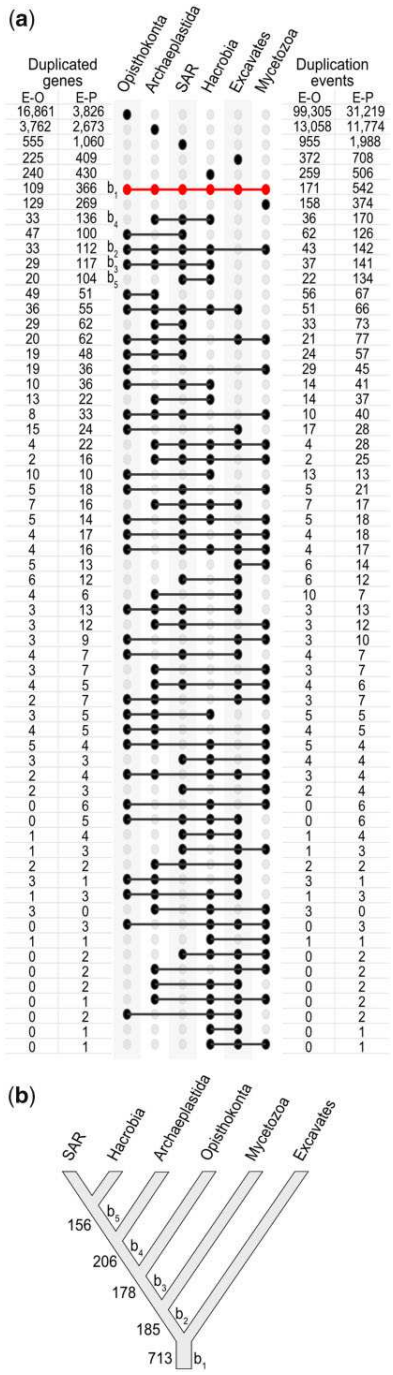


Fig. 2.—Distribution of paralogs descending from gene duplications across six eukaryotic supergroups. (*a*) The figure shows the distribution of paralogs resulting from gene duplications in eukaryotic-specific genes (E-O) and eukaryotic genes with prokaryotic homologs (E-P) (see Materials and Methods 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).

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

(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

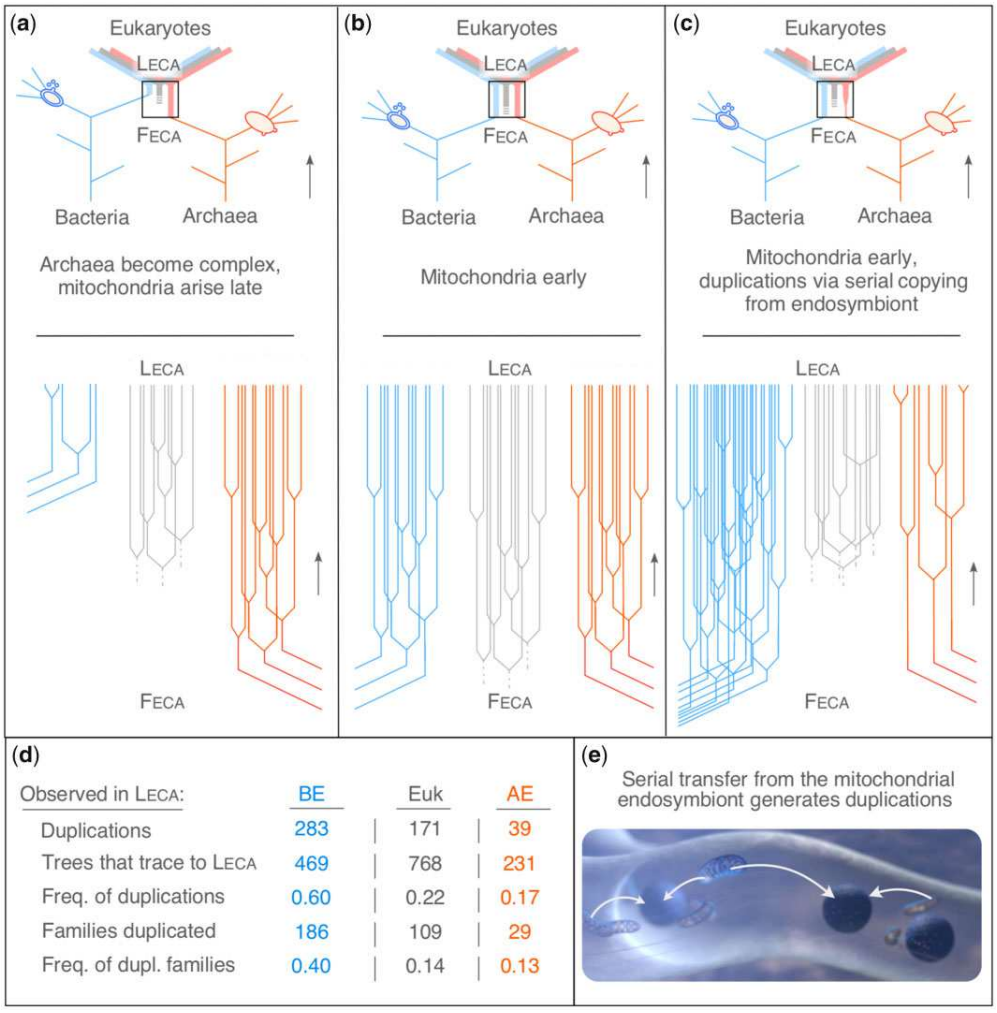


FIG. 3.—Alternative models for eukaryote origin generate different predictions with respect to duplications. In each panel, gene duplications during the FECA to LECA transition (boxed in upper portion) are enlarged in the lower portion of the panel. (*a*) Cellular complexity and genome expansion in an 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

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*).

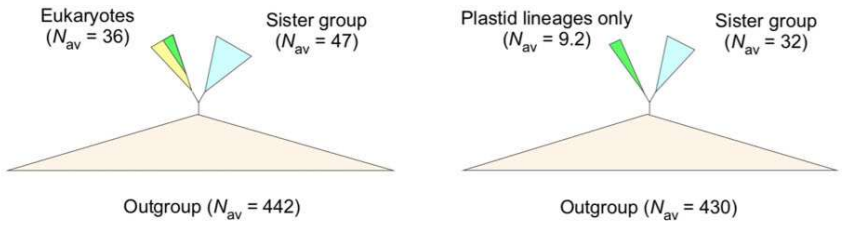
Do Bacterial Genes in LECA Stem from the Mitochondrion?

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 $\geq$1,649 out of the 2,575 genes shared by prokaryotes and eukaryotes to LECA, and the origin of $\leq$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,

(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.

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

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

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

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

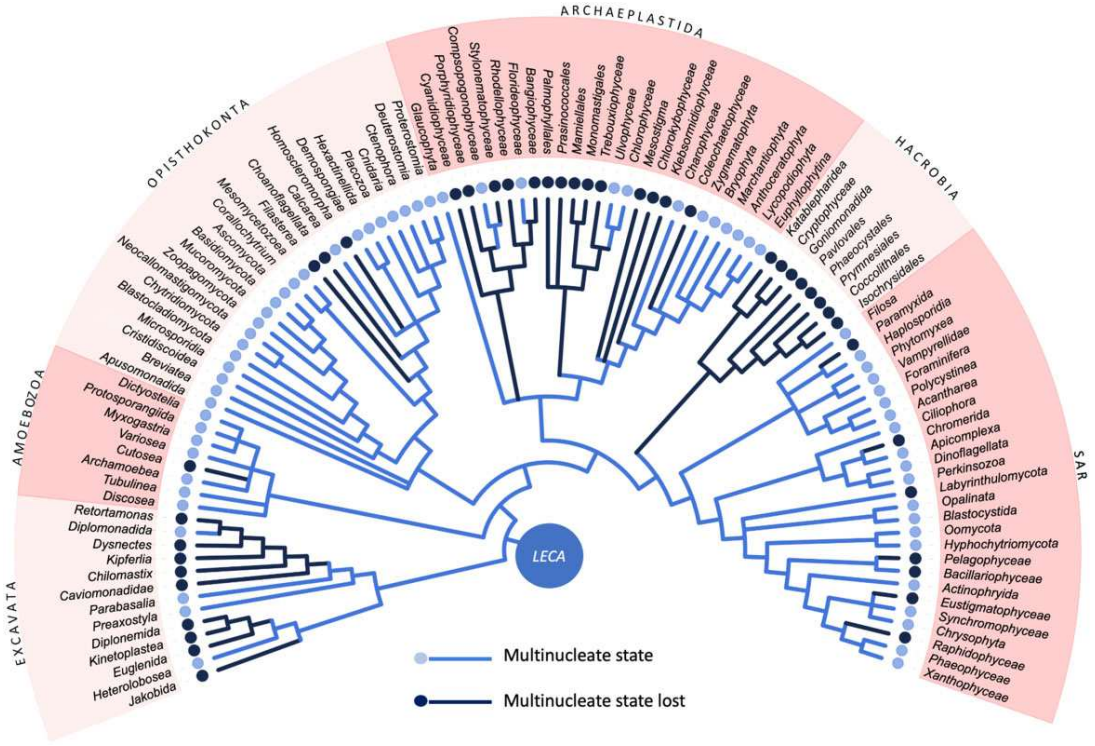


Fig. 5.—Ancestral state reconstruction for nuclear organization in eukaryotes. Presence and absence of the multinucleate state in members of the respective group are indicated. Resolution of the branches (polytomy vs. dichotomy) does not alter the outcome of the ancestral state reconstruction, nor does 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).

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

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 alpha-proteobacteria 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.

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.

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

Manuscript 3:

Earth's atmospheric oxygen levels were governed by three enzymes and drove biochemical evolution through oxygen inhibition

Authors: Nagies, F. S. P., Wimmer, J. L. E., Mrnjavac, N., Knopp, M. R., Kapust, N., Gerhards, R. E., Trost, K., Modjewski, L., Bremer, N., Degli Esposti, M., Misrahi, I., Allen, J. F. & Martin, W. F.

Published: Unpublished Manuscript.

Contribution of Falk Sascha Per Nagies:

Shared first author

I curated with several of the other authors the data and analyzed the data in regard to HGT patterns in oxygen dependent and independent enzymes. I conceptualized with several of the other authors the first version of the manuscript. Then I participated in analyzing or analyzed myself the patterns visible in the data in regard to function or distribution among prokaryotes. Finally, I contributed to the writing, editing and reviewing of the final manuscript.

Earth's atmospheric oxygen levels were governed by three enzymes and drove biochemical evolution through oxygen inhibition

Short title: The history of atmospheric oxygen and its impact on biochemical evolution in prokaryotes

Falk S. P. Nagies[1¶], Jessica L. E. Wimmer[1*], Natalia Mrnjavac[1¶], Michael R. Knopp[1], Nils Kapust[1], Rebecca E. Gerhards[1], Katharina Trost[1], Luca Modjewski[1], Nico Bremer[1], Mauro Degli Esposti[2], Itzhak Mizrahi[3], John F. Allen[4], William F. Martin[1]

[1] Department of Biology, Institute for Molecular Evolution, Heinrich Heine University of Duesseldorf, Duesseldorf, Germany

[2] Center for Genomic Sciences, UNAM Campus de Cuernavaca, Morelos, Mexico

[3] Department of Life Sciences, Ben-Gurion University of the Negev and the National Institute for Biotechnology in the Negev, Marcus Family Campus, Be'er-Sheva, Israel

[4] Research Department of Genetics, Evolution and Environment, University College London, Gower Street, London, U.K.

[*] Corresponding author: Jessica L. E. Wimmer

E-mail: jessica.wimmer@hhu.de

[¶] These authors contributed equally to this work.

Abstract

Enzymes that use molecular oxygen as a substrate arose as a consequence of cyanobacterial O_2 production in the Great Oxidation Event (GOE) 2.4 billion years ago. The genes for enzymes encoding 365 prokaryotic reactions that use O_2 as a substrate and that map to gene clusters underwent LGT more frequently than genes for enzymes of oxygen-independent chemical reactions. Though traditional views posit that the main selective advantage that O_2 provided to prokaryotes was increased energy conversion in respiratory pathways, many oxygen reductases of prokaryotic respiratory chains do not conserve energy via proton pumping. The O_2 consuming reactions surveyed have an average estimated $\Delta G'$ of –243 $kJ \cdot mol^{-1}$, but only one out of 365 reactions conserves energy via substrate level phosphorylation. The most common functions among O_2-dependent enzymes in prokaryotes involve secondary metabolism and the cleavage of stable bonds in organic (biogenic) substrates using the electron extracting power of O_2 in redox reactions. The history of Earth's atmospheric O_2 accumulation reflects three phases: the emergence of O_2 at the GOE 2.4 billion years ago, its subsequent stasis at about 1 % of modern levels until roughly 500 million years ago, followed by its accumulation to modern levels. Various geochemical processes have been proposed to have caused oxygen's three-phase accumulation. Physiological data suggest that the history of O_2 accumulation on Earth was not governed by geochemical processes but by only three enzymes in O_2-producing lineages: i) the oxygen evolving complex of cyanobacterial photosystem II, which generated Earth's O_2, ii) cyanobacterial nitrogenase, O_2-inhibition of which kept atmospheric O_2 levels low for 2 billion years, and iii) cellulose synthase of land plants, which allowed O_2 to rise to its current levels of 21 %. The evolutionary advantage of O_2-dependent reactions was circumvention of O_2-inhibition in preexisting anaerobic enzymes that involved O_2-sensitive cofactors, typically FeS clusters, or O_2-sensitive

48 substrates. This allowed microbes to synthesize essential cofactors in O_2-containing
49 environments and provided them access to carbon and nitrogen nutrients in habitats where O_2-
50 sensitive reactions were inhibited.

51

52 Introduction

53 Discovered by Priestley as the agent produced by green leaves that allows air to support
54 the burning of a candle and the life of a mouse (Priestley, 1772) and by Scheele as a product
55 of heated mercuric oxide (Scheele, 1777), oxygen was named by their contemporary Lavoisier
56 as *oxygène* (acid-former) (Lavoisier, 1781). The finding by Pasteur (1861) that fermentations
57 are performed by organisms that grow in the absence of oxygen divided the living world into
58 aerobes and anaerobes, a dichotomy that gave rise to the 100-year-old idea that the first
59 organisms on Earth arose and lived as anaerobic chemolithoautotrophs (Mereschkowsky 1910;
60 Preiner et al., 2020). The Great Oxidation Event, GOE (Holland, 2002), punctuates Earth's
61 history close to its midpoint. Roughly 2.4 billion years ago (Ga), photosynthetic prokaryotes
62 with two chlorophyll-based photosystems linked in series—cyanobacteria—invented the
63 molecular tools needed to extract electrons from H_2O, so that those electrons could be used to
64 fix CO_2 and N_2 for growth. The byproduct of that reaction was O_2, produced by the Mn
65 containing oxygen-evolving complex (OEC) of photosystem II.

66 The environmental setting for the origin of O_2 was global anoxia, because from the
67 origin of the first microbes at roughly 4 Ga to the onset of the GOE (Kump, 2008), the Earth's
68 oceans and atmosphere were effectively devoid of O_2. Prior to the GOE, all microbial ecology
69 and evolution occurred in a world of anaerobes. During the GOE, cyanobacterial
70 photosynthesis allowed O_2 to accumulate in Earth's oceans and atmosphere from around zero
71 to approximately 1 % of its present atmospheric level (PAL) (Ślesak et al., 2019). Following

72 the GOE, atmospheric O_2 did not continue to accumulate to higher levels, rather it stayed more
73 or less constant at 1 % PAL for roughly 1.8 billion years (Lyons et al., 2014; Allen et al., 2019)
74 until about 580 million years ago, when O_2 levels started to rise further, approaching modern
75 levels with the advent of land plants about 450 million years ago (Lenton et al., 2016; Stolper
76 and Keller, 2018; Mills et al., 2022). The protracted low oxygen phase of Earth history from
77 2.4 Ga to 0.58 Ga has been called the 'boring billion' (Buick et al., 1995; Mukherjee et al.,
78 2018) to emphasize the lack of geologically interesting events during that period of O_2 stasis,
79 but it has also been called the 'Pasteurian' era of life's history (Martin et al., 2020; Mills et al.,
80 2022) to emphasize the biologically crucial observation that O_2 levels of 1 % PAL during 'the
81 boring billion' correspond to the Pasteur point—the level of ambient oxygen (ca. 1 % PAL or
82 0.2 % v/v) at which facultative aerobes switch their terminal acceptors from anaerobic to
83 aerobic respiration. The boring Pasteurian billion was an era during which anaerobes learned
84 to manage the toxicity of O_2 and to use O_2 for their benefit.

85 The main evolutionary impact of the origin of O_2 is traditionally viewed either from the
86 standpoint of animal evolution (Cloud, 1968; Planavsky et al., 2014; Reinhard et al., 2016) or
87 from the standpoint of eukaryote origin (Margulis, 1970; Raymond and Segrè, 2006). Yet
88 animals appear in the fossil record at the base of the Cambrian, almost 2 billion years later than
89 the appearance of O_2 (Budd, 2008), and eukaryotes appear roughly 1 billion years after the
90 advent of O_2 (Zimorski et al., 2019; Mills et al., 2022). Both eukaryotes and animals were born
91 into a world that already contained small amounts of O_2 (Müller et al., 2012; Zimorski et al.,
92 2019; Mills et al., 2022), hence their immediate ancestors were already equipped,
93 enzymatically, to deal both with anoxia and with O_2 as a toxin and a substrate (Müller et al.,
94 2012). The fossilization of early invertebrates required the existence of O_2 for the synthesis of
95 collagen by proline hydroxylases to generate hard body parts (Towe, 1970; Shoulders and
96 Raines, 2009). The origin of large animals also required the existence of O_2 (Payne et al., 2011;

Harrison et al., 2010) and life on land, above the soil line, entailed changes in the way that eukaryotes transfer electrons, namely a transition from ferredoxin-dependent one-electron transport chains to NADH-based two-electron transfers (Gould et al., 2019). The impact of O_2 on prokaryote evolution involved the origin of O_2-dependent enzymes long before the origin of eukaryotes (Fig 1).

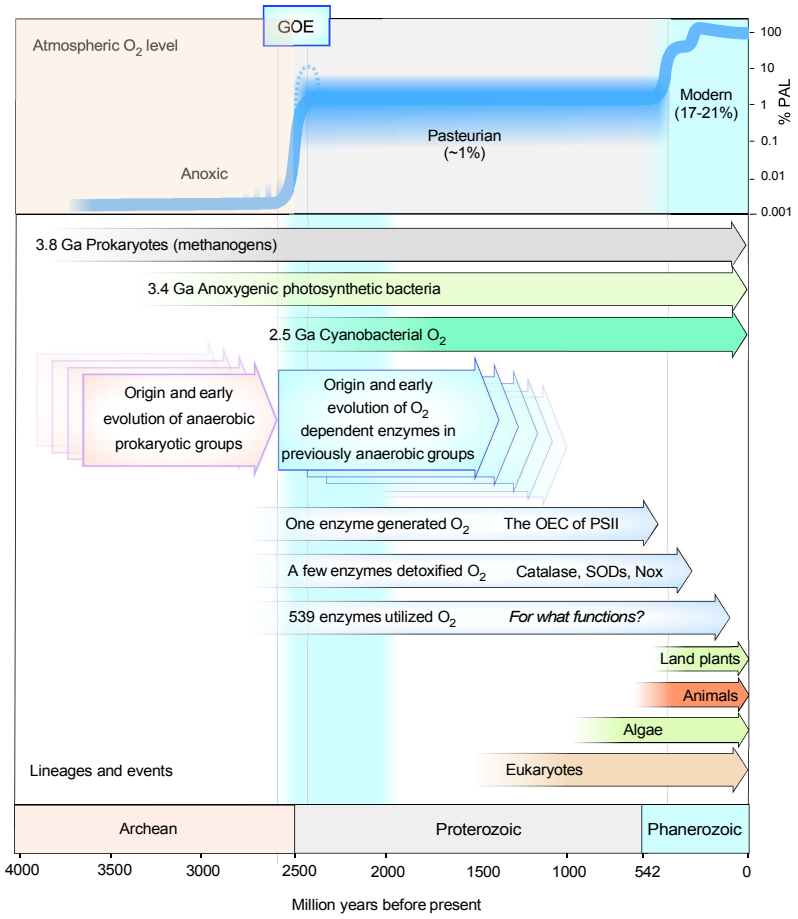


Fig 1. A timeline of Earth history with the rise of O_2 and the appearance of some relevant groups of eukaryotes and prokaryotes indicated. Blue boxes emphasize the GOE, the onset of O_2-dependent enzymes, and questions concerning their evolution and functions. The number of 547 stems from the

present paper (see text). SOD: superoxide dismutase; Nox: NADH oxidases (diaphorases). Data from Lyons et al. (2014); Mills et al. (2022); Stolper and Keller (2018); Brocks et al. (2017); Arndt and Nisbet (2012). The reasons why O_2 levels remained near the Pasteur point for 1.8 billion years are still discussed. Numerous geological causes (Poulton et al., 2004a; Alcott et al. 2019; Klatt et al., 2021) and one biological cause (Allen et al., 2019) for the existence of the boring billion have been proposed (see text). It is largely undebated that cyanobacteria generated the global supply of O_2 via one single enzyme and one single enzyme activity: the conserved Mn-containing oxygen evolving complex of photosystem II.

Prokaryotes were eyewitnesses to the GOE. They have existed in uninterrupted continuity starting possibly as early as 3.95 Ga (Tashiro et al., 2017), probably as early as 3.8 Ga (Ueno et al., 2006), and surely by 3.5 Ga at the very latest (Arndt and Nisbet, 2012; Nisbet and Sleep, 2001; Westall et al., 2011). By 3.4 to 3.3 Ga, anoxygenic photosynthetic prokaryotes were in existence generating stromatolites in aerial settings (Allen, 2016; Arndt and Nisbet, 2012). Since prokaryotes encompass the only lineages that actually experienced the GOE (Fig 1), we reasoned that their O_2-dependent enzymes might hold clues about the impact of O_2 on physiological evolution.

Studies that address the role of O_2 in evolution generally focus on its role as a terminal electron acceptor in respiratory chains (Castresana et al., 1994; Castresana and Moreira, 1999; Brochier-Armanet et al., 2009; Murali et al., 2022). Such investigations are hampered by the circumstance that terminal oxidases have been subject to lateral gene transfer (LGT) in evolution, yet to an unknown extent. Data from genomes show that essentially all genes in prokaryotic genomes have been subject to LGT at some point in evolution (Dagan et al., 2008; Nagies et al., 2020), and oxygen-dependent enzymes are no exception (Passardi et al., 2007). Terminal oxidases are important in the evolution of energy metabolism, but represent only about 1 % of known prokaryotic O_2-utilizing enzyme families (Sousa et al., 2016).

132 The responses of anaerobic microbes to O_2 via O_2 detoxification enzymes such as NADH
133 oxidases and rubredoxin:oxygen oxidoreductase are well studied (Lu and Imlay, 2021), as is
134 the impact of O_2 on prokaryotic gene expression via DNA binding proteins such as the FeS
135 cluster containing O_2 sensor FNR (for fumarate and nitrate respiration) (Unden and Bongaerts,
136 1997) or the ArcAB two-component system (for anoxic redox control or aerobic respiratory
137 control), which responds to the oxidation state of the quinone pool (Brown et al., 2022). The
138 impact of O_2-dependent enzymes on the expansion of eukaryotic biochemical pathways has
139 been surveyed (Raymond and Segrè, 2006), yet the majority of enzymes surveyed in that study
140 were members of the same family—eukaryotic cytochrome P_{450} enzymes having different
141 substrate specificities—whereby the vast majority of cytochrome P_{450} enzymes arose very late
142 evolution, in the plant, animal and fungal lineages (Nelson, 2018).

143 The OEC of photosystem II not only extracts electrons from H_2O to supply the
144 photosynthetic electron transport chain of cyanobacteria, it also stores massive amounts of
145 chemical energy in its enzymatic reaction product, O_2. The reduction potential for the O_2/H_2O
146 pair, i.e. the potential for the complete four-electron reduction of dioxygen, is +815 mV at pH
147 7 (Wood, 1988). The enthalpy change in combustion reactions of O_2 with organic compounds
148 (generating CO_2) are thus typically on the order –400 kJ per mol of O_2 regardless of the organic
149 compound undergoing combustion, because the energy released is almost entirely the energy
150 stored in the energy-rich O_2 molecule, not in the organic reactant (Schmidt-Rohr, 2015). The
151 Gibbs energy for these reactions is also highly negative, with a calculated average of –243 kJ
152 per mol of O_2. What did cells do upon their encounter with this virtually unlimited source of
153 concentrated chemical energy (Sousa et al., 2016)? Here we investigate the impact of O_2-
154 dependent enzymes on microbial evolution in terms of the functions that these prokaryotic
155 enzymes fulfill, examining the nature of the physiological traits that O_2-dependent enzymes

imparted upon prokaryotes, and in terms of their dispersal among lineages via lateral gene transfer (LGT) following the GOE.

Results

Genes for O_2-dependent enzymes occur most frequently in generalist genomes

Among the 11,804 reactions in the KEGG database, we identified all O_2-dependent reactions and mapped them to gene clusters generated for 5565 sequenced prokaryotes. Of 1,949 reactions that use O_2 as a substrate in KEGG, 540 occur in prokaryotes, of which 365 mapped to 792 clusters of prokaryotic genes (see Material and methods). The distribution of O_2-dependent enzymes across prokaryotic lineages in the present sample is not fundamentally different from that observed in earlier studies (Sousa et al., 2016; Jabłońska and Tawfik, 2019). The prokaryotic phyla with the most oxygen-dependent reactions are metabolic generalists with large genomes from the Betaproteobacteria, followed by Actinobacteria, Cyanobacteria, Alphaproteobacteria and Gammaproteobacteria (Table 1). Most members of those groups are, however, facultative anaerobes. Of the 5655 genomes in our sample, 4165 are classified as being able to grow in aerobic conditions based on their terminal oxidases (Sousa et al. 2011). The largest number of O_2-dependent reactions is found in the genome of the actinobacterium *Rhodococcus jostii* (158), followed by *Rhodococcus opacus* (151). Among archaea, the largest number of O_2-dependent reactions were found in Sulfolobales, Natrialbales, Halobacteriales and other halophiles. Archaeal halophiles are aerobic heterotrophs that are evolutionarily derived from anaerobic autotrophic methanogens via gene acquisitions, from bacteria, for heterotrophy and O_2 respiration (Nelson-Sathi et al. 2012). The smallest number of O_2-

dependent reactions per genome was found in methanogens, Tenericutes and Thermotogae. Anaerobes can harbour O_2-dependent enzymes (McCord and Fridovich, 1969; Brioukhanov and Netrusov, 2007; Jabłońska and Tawfik, 2019). Of the 5655 genomes in our sample, 249 harbour no O_2-dependent reactions (S1 Table). These are often, but not always, small genomes, and they are always identified as anaerobes using the classifier of Sousa et al. (2011) (see Material and methods). The most frequent phylum among genomes harbouring no O_2-dependent reactions is the Tenericutes, which includes the heavily sampled mycoplasmas, obligate anaerobic pathogens that lack a cell wall. The taxonomic distribution among O_2-dependent and O_2-independent reactions is shown in Fig 2. Fig 3 shows the number of O_2-dependent reactions normalized to genome size. Anaerobes have significantly fewer O_2-dependent reactions per genome (Welch's t-test: $p < 1 \cdot 10^{-300}$; mean anaerobes = 4.5, mean aerobes = 37.3). They also have significantly smaller genomes as a result of their ancestral specialization (Welch's t-test: $p < 1 \cdot 10^{-300}$; mean anaerobes = 2000.6, mean aerobes = 3858.4, S2A Table).

Table 1. O_2-dependent reactions per genome across groups.

Taxonomic group	$R_{ox.}$	R_a	N_g	P_{An}
Betaproteobacteria	65	4-143	493	0.02
Actinobacteria	58	0-159	593	0.19
Cyanobacteria	46	18-82	94	0.02
Alphaproteobacteria	46	5-128	460	0.03
Gammaproteobacteria	45	1-140	1565	0.08
Acidobacteria	45	25-67	8	0.13
Deinococcus-Thermus	37	11-60	26	0
Sulfolobales	32	7-44	29	0
Bacteroidetes	32	0-87	185	0.26
Natrialbales	31	17-49	9	0
Planctomycetes	30	19-76	8	0
Deltaproteobacteria	27	1-106	74	0.32
Bacilli	26	0-83	1012	0.13
Nitrospirae	24	7-42	8	0.41
Haloferacales	20	13-26	7	0
Halobacteriales	18	11-31	16	0.13
Verrucomicrobia	17	2-46	10	0.30
Chloroflexi	17	2-60	26	0.65
Chlorobi	11	7-18	12	0.58
Aquificae	11	8-13	16	0.19
Clostridia	10	0-34	169	0.98
Epsilonproteobacteria	10	6-31	253	0.01
Spirochaetes	9	0-49	66	0.76
Erysipelotrichia	7	1-10	8	1
Methanosarcinales	6	2-10	32	0.97
Negativicutes	6	2-16	12	1.00
Archaeoglobales	6	1-11	8	0.63
Thermococcales	5	2-7	27	1
Fusobacteria	5	0-16	20	1.00
Chlamydiae	5	3-36	113	0.96
Thermoproteales	4	2-7	10	0.80
Methanomicrobiales	3	0-6	9	1
Desulfurococcales	2	0-5	11	1
Thermotogae	2	0-6	32	1
Methanobacteriales	2	0-6	17	1
Tenericutes	1	0-19	141	1
Methanococcales	0	0-1	16	1

R_{ox}: Average number of O_2-dependent reactions per genome in the corresponding phylum, rounded to the nearest integer; R_a: Range of O_2-dependent reactions per genome in the corresponding phylum; N_g: Number of genomes; P_{An}: Proportion of anaerobes in the corresponding phylum according to the classifier of Sousa et al. (2011).

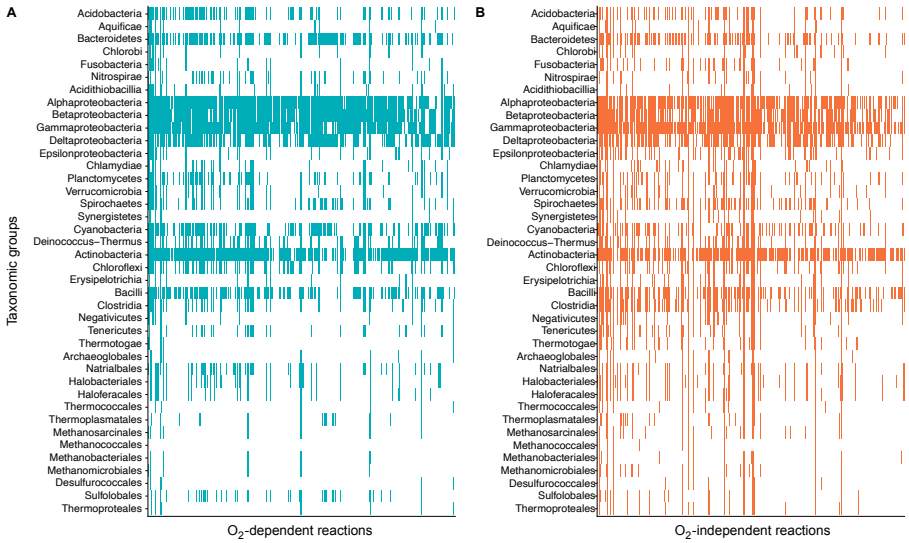


Fig 2. Taxonomic distribution among O_2-dependent and O_2-independent reactions. Panel (**A**) shows 365 O_2-dependent reactions in blue and panel (**B**) 3,018 O_2-independent reactions in orange for comparison. The reactions are sorted by decreasing average verticality value (V_{avg}) of the corresponding clusters.

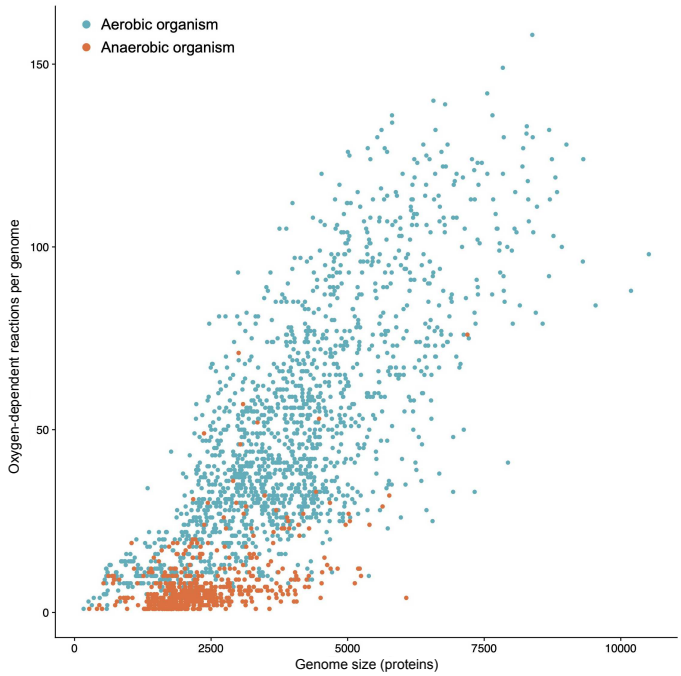


Fig 3. Reaction frequency versus genome size for O_2-dependent reactions in prokaryotes. For each species in the dataset, only the strain with the largest genome was used. Anaerobic genomes are colored in orange and aerobic organisms in blue (aerobe vs. anaerobe classification according to Sousa et al., 2011).

Oxygen-dependent enzymes oxidize organic substrates

Oxygen is a strong but stable oxidant. The O_2 molecule is a diradical with two unpaired electrons. The shorthand structure of O_2 is sometimes written as ·O–O· instead of O=O to underscores its diradical nature. Though most radicals are extremely reactive, O_2 is generally unreactive (Lu and Imlay, 2021). The O_2 diradical is kinetically stable because each of the unpaired electrons in O_2 is delocalized over a two-center, three-electron π bond, resulting in a very large resonance stabilization energy, and consequently a high activation energy barrier

(Borden et al., 2017). Despite the kinetic stability imparted by its two-center, three-electron π bonds, O_2 has a very weak σ bond (Borden et al., 2017). From the thermodynamic standpoint, this makes O_2 an extremely energy-rich molecule (Schmidt-Rohr, 2015), so energy-rich that it undergoes exergonic reactions with every element except gold (Brewer, 1952).

Based upon the frequency of O_2-dependent biochemical functions among prokaryotic enzymes, the main utility of oxygen appears to be the chemical hallmark of O_2 itself: high-energy oxidation, affording microbes that possess the corresponding genes accessibility to substrates in the presence of O_2. The most frequent reaction types of prokaryotic O_2-dependent enzymes are aromatic degradation and amine oxidation (Table 2). Dioxygenases, which incorporate both atoms of O_2 into the reaction product (EC 1.13.11.-), are the most common enzyme type (n=57), followed by NAD(P)H-dependent monooxygenases (n=54), which incorporate one atom from O_2 into the product (E.C. 1.14.13.-). The enzymes from both groups act mainly on aromatic substrates (Table 2). The dioxygenases usually disrupt aromatic rings (Vaillancourt et al. 2006). The next most common enzyme type were amine oxidases acting on $CH–NH_2$ groups (EC 1.4.3.-) (n=48), copper or flavin containing proteins that catalyze the oxidation of primary amines, polyamines and amino acids (Gaweska and Fitzpatrick 2011). NAD(P)H-dependent dioxygenases (n=38) (EC 1.14.12.-) were the next most common category, all of which acted on aromatic substrates (Fuchs et al., 2011), followed by iron sulfur dependent monooxygenases (Vanoni, 2021) (n=28) (E.C. 1.14.12.-) (Table 2).

Of the 365 reactions that could be linked to protein families (see Material and methods), 194 (53 %) act on aromatic substrates. The predominance of aromatic substrates for O_2-dependent reactions is generally not surprising because these substrates are chemically quite stable, requiring either a strong oxidant or a very strong reductant (Huwiler et al., 2019) to disrupt the aromatic ring. Aromatic degradation does not strictly require O_2, because many microbes harbor O_2-independent, CoA-dependent enzymatic pathways for anaerobic

degradation of aromatic compounds (Fuchs et al., 2011; Fuchs, 2008). Even though the ring activating reaction of O_2-dependent aromatic degradation is more exergonic than the reductive O_2-independent route catalyzed by benzoyl-CoA reductases, generating midpoint potentials around –622 mV (Huwiler et al., 2019), the growth rates of aerobic and anaerobic aromatic degrading microbes are not fundamentally different, both having doubling times on the order of 4-6 hours (Fuchs et al., 2011). This is noteworthy. The comparable growth rates for microbes employing the O_2-dependent and the O_2-independent routes suggest that the advantage of the O_2-dependent aromatic degradation pathways might simply lie in the O_2-sensitivity of the (older) anaerobic routes, which were inhibited in oxic environments, an interpretation (Gomez Maqueo Chew and Bryant, 2007) that we will encounter again in the context of essential cofactor biosynthesis.

Table 2. Most common reaction types and reactants of prokaryotic O_2-dependent enzymes.

EC number	n	e^- Donor (substrate)	Fate of O_2 atoms	Dearo.	NH_3 Prod.	Substr. Arom.
1.13.11.-	54	Single donor	Dioxygenase	40	0	45
1.14.13.-	53	Donor + NAD(P)H	Monooxygenase	4	0	38
1.4.3.-	39	CH-NH_2 group	Acceptor	0	27	21
1.14.12.-	38	Donor + NAD(P)H	Dioxygenase	28	2	38
1.14.15.-	28	Donor + FeS cluster	Monooxygenase	0	0	3
1.14.99.-	23	Misc. paired donors	Diverse	0	0	7
1.14.19.-	21	Paired donors	Acceptor	0	0	7
1.14.14.-	18	Flavin and Donor	Monooxygenase	0	0	6
1.1.3.-	14	CH-OH group	Acceptor	0	0	3
1.14.11.-	11	Paired donors	Diverse	0	0	2
1.5.3.-	7	CH-NH group	Acceptor	0	0	4
1.14.18.-	7	Paired donors	Monooxygenase	2	0	6
1.13.12.-	7	Single donor	Monooxygenase	0	0	4
1.14.20.-	6	Donor + 2-OG	Diverse	0	0	2
1.10.3.-	6	Diphenols	Acceptor	2	0	5
1.3.3.-	5	CH-CH group	Diverse	2	0	3
7.1.1.-*	3	Quinones (or cyt *c*)	Acceptor	0	0	3

Notes: **n:** number of reactions; Dearo: number of dearomatizing reactions (the reaction destroys ring aromaticity); Substr. Arom.: number of reactions where the substrate is an aromatic compound (excludes cosubstrates such as $NAD^+(P)H$, flavin etc.); NH_3 Prod.: number of reactions releasing ammonia/ammonium ion as a reaction product; Dioxygenase: both atoms of O incorporated into substrate; Monooxygenase: one atom of O incorporated into substrate; Diverse: O_2 can be incorporated or reduced; Acceptor: O_2 is reduced, not incorporated. * Given out of order in the ranking, there are E.C. categories more common than the translocases (7.1.1.-, terminal oxidases) in the data, but the terminal oxidases are important, hence added to the list. EC numbers and links to all KEGG reactions in S3 Table.

The functions of O_2-dependent reactions across genomes

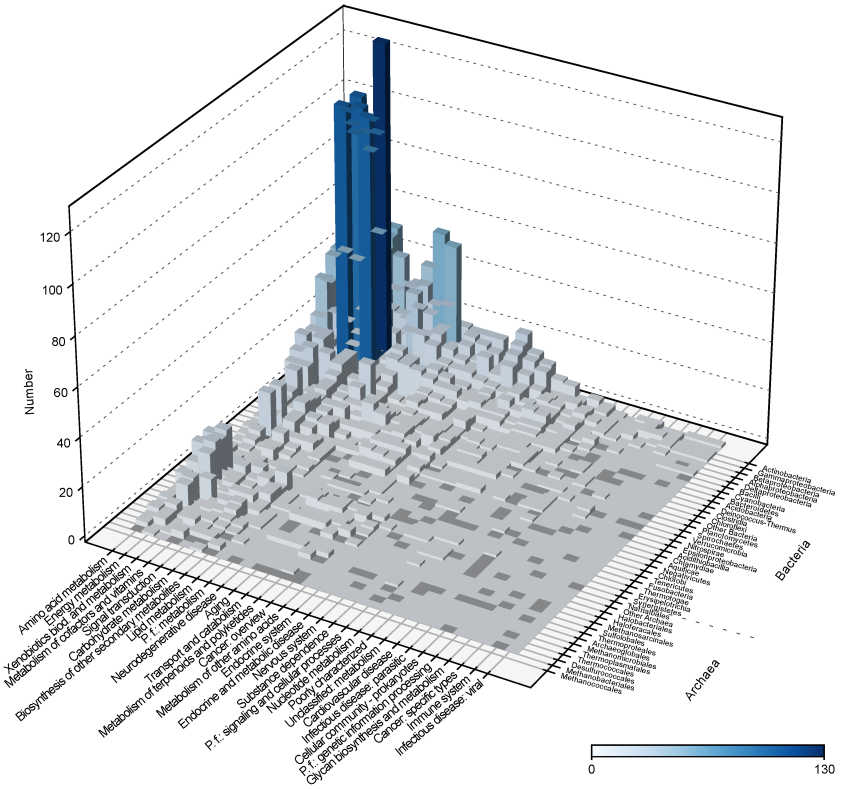


Fig 4. Distribution of KEGG functional categories within prokaryotic phyla associated with protein families catalyzing oxygen-dependent reactions. Thirty-one functional categories (x-axis) were sorted according to the sum of frequencies (y-axis) within each phylum and 14 archaeal / 28 bacterial phyla were ordered according to the sum of frequencies of all functional categories. Both axes

were arranged in ascending order starting at the x/z-intersection, archaea and bacteria were sorted separately. The color scheme underlines the values corresponding to the y-axis from low (light grey) to high (dark blue).

The distribution of genes for O_2-dependent enzymes as assigned to functional categories in KEGG within archaeal and bacterial phyla that appear in 792 protein families associated with 365 O_2-dependent reactions is shown in Fig 4. The most frequent functional categories among oxygen-dependent reactions are amino acid metabolism, energy metabolism, xenobiotics biodegradation and metabolism of cofactors and vitamins. The highest frequencies of O_2-dependent reactions per functional category and *phylum* belong to xenobiotic biodegradation in Actinobacteria, Alpha-, Beta- and Gammaproteobacteria (127, 102, 99, 92, respectively). Amino acid metabolism occurs four times among the top ten values (Alphaproteobacteria=104, Actinobacteria=103, Betaproteobacteria=100, Gammaproteobacteria= 93). The remaining two values that complete the top ten list of O_2-dependent reaction frequencies per functional category and *phylum* belong to energy metabolism in Betaproteobacteria (103) and Alphaproteobacteria (94). Most of the functional category/phylum combinations contain low frequencies of O_2-dependent reactions. Out of 1302 possible combinations, 647 include no O_2-dependent reactions (zero column height, 49.7 %) and 157 contain only one reaction (8 %) (Fig 4).

The frequency of each enzyme among the 5655 genomes in our sample is shown in Table 3 for the 30 most common O_2-dependent enzymes in the current sample (the full list is given in S4 Table). The most common reactions by the measure of gene distributions fall largely into four functional classes: biosynthesis, terminal oxidases, detoxification, and substrate mobilization. In primary metabolism, the O_2-dependent biosynthetic reactions present post-GOE alternatives to O_2-independent reactions that existed in cells before the GOE. The roles

359 of terminal oxidases also present alternatives, post-GOE modifications of anaerobic respiratory
360 chains that used sulfite, metals or other terminal acceptors prior to the advent of O_2. The
361 oxidation of amino acids, sulfonates, or 2-hydroxy acids also present alternatives to preexisting
362 anaerobic pathways. These are metabolic functions that were modified with the advent of O_2,
363 not biochemical novelties. They represent O_2-dependent alternatives that arose in the presence
364 of preexisting (and still existing) anaerobic pathways.

365 Genuinely novel enzymatic functions that did not exist prior to the existence of O_2 are
366 found in the biosynthesis and degradation of secondary metabolites (Sousa et al. 2016;
367 Jabłońska and Tawfik, 2019; Hoffrath et al. 2020). The role of O_2 in the secondary metabolism
368 was already underscored in the study by Raymond and Segrè (2006), who found that the only
369 categories that were significantly ($p=0.05$) enriched in O_2-dependent functions in their
370 simulated metabolic networks were involved in the metabolism of penicillin, limonene
371 (terpenes), indole, flavonoids, sterols, bile acids, androgens and alkaloids. Raymond and Segrè
372 (2006) included eukaryotic enzymes in their survey. Here we are looking only at reactions of
373 prokaryotes.

374 Another novel class of enzymes that did not exist prior to the GOE are the detoxification
375 enzymes that scavenge ROS (reactive oxygen species)—the catalases (Tehrani & Moosavi-
376 Movahedi, 2018; Whittaker, 2012), catalase-peroxidases (Khmelevtsova et al., 2020), and
377 superoxide dismutases (Bafana et al., 2011). The main cofactors of the ROS scavenging
378 enzymes are metals (Ighodaro and Akinloye, 2018; Njuma et al., 2014; Bafana et al., 2011;
379 Boden et al., 2021). Their main function is detoxification of reactive oxygen species that form
380 as a result of O_2-dependent metabolism.

381 It is noteworthy that the most common O_2 reducing terminal oxidases of respiratory chains
382 in the present sample are cytochrome *bd* ubiquinol oxidases, which do not pump protons

(Borisov et al., 2011; Murali et al., 2022), although they do generate proton motive force via scalar protons (the localization of proton-consuming and proton-generating reactions on different sides of the membrane). While *bd* oxidases allow the respiratory chain to operate with O_2, thereby enabling other coupling sites to conserved energy by pumping in the respiratory chain, the same energy conservation strategy is encountered at the level of NADH oxidation in the respiratory chain. Many facultative anaerobes like *E. coli* express alternative forms of NADH dehydrogenases, *nuo* and *ndh* genes. The protein encoded by *nuo* genes pumps protons and is primarily used under anaerobic conditions, whereas that encoded by *ndh* does not pump protons and is primarily used under aerobic conditions (Tran und Unden, 1998; Unden et al., 2002). Hence, both at the terminal oxidase and NADH dehydrogenase level, there is no strict correlation between use of O_2 as a terminal acceptor and energy conservation.

Table 3. Most widely distributed prokaryotic O_2-dependent enzymes (KEGG orthology groups).

Enzyme name	EC number	KO	N_g	*V*
7,8-dihydroneopterin oxygenase	1.13.11.81	K01633	4,424	0.90
Cytochrome *bd* ubiquinol oxidase subunit II [0] (Oxphos)	7.1.1.7	K00425	4,075	0.63
Cytochrome *bd* ubiquinol oxidase subunit I [0] (Oxphos)	7.1.1.7	K00426	4,074	0.57
Catalase (O_2 detoxification)	1.11.1.6	K03781	3,485	0.30
L-aspartate oxidase (NAD^+ synthesis; N mobilization)	1.4.3.16	K00278	3,388	6.22
Pyridoxamine 5'-phosphate oxidase (PLP synthesis)	1.4.3.5	K00275	3,026	0.11
Nitronate monooxygenase (N mobilization)	1.13.12.16	K00459	2,974	0.34
Bacterioferritin (Iron storage)	1.16.3.1	K03594	2,867	0.52
Glycolate dehydrogenase FAD-linked SU (2-OH acids)	1.1.99.14	K00104	2,612	0.36
Cytochrome *o* ubiquinol oxidase subunit I [2] (Oxphos)	7.1.1.3	K02298	2,561	0.14
Coproporphyrinogen III oxidase (Heme synthesis)	1.3.3.3	K00228	2,473	1.18
Cytochrome *c* oxidase subunit I [4] (Oxphos)	7.1.1.9	K02274	2,448	0.86
Cytochrome *c* oxidase subunit II [4] (Oxphos)	7.1.1.9	K02275	2,435	0.39
2-Polyprenylphenol 6-hydroxylase (UQ synthesis)	1.14.13.240	K18800	2,431	1.49
Cytochrome *o* ubiquinol oxidase subunit II [2] (Oxphos)	7.1.1.3	K02297	2,406	0.57
4,5-DOPA dioxygenase (Amino acids & aromatics ox.)	1.13.11.-	K15777	2,266	1.15
Catalase-peroxidase (substrate oxidation and O_2 detox)	1.11.1.21	K03782	2,197	0.14
Superoxide dismutase, Cu-Zn family (O_2 detoxification)	1.15.1.1	K04565	2,180	1.07
2-octaprenyl-6-methoxyphenol hydroxylase (UQ synth.)	1.14.13.-	K03185	2,157	0.72

Malate dehydrogenase, quinone (TCA cycle)	1.1.5.4	K00116	2,095	0.31
Glycine oxidase (Thiamine biosynthesis)	1.4.3.19	K03153	1,886	0.97
(S)-2-hydroxy-acid oxidase (2-OH acids, glycolate DH)	1.1.3.15	K11473	1,849	0.90
Alkanesulfonate monooxygenase (S mobilization)	1.14.14.5	K00299	1,831	0.09
Gamma-glutamyl putrescine oxidase (Amino acid ox.)	1.4.3.-	K09471	1,809	0.09
tRNA 5-MAM-2-thiouridine bifunctional protein (tRNA)	1.5.-.-	K15461	1,789	0.18
4-hydroxyphenylpyruvate dioxygenase (Amino acid ox.)	1.13.11.27	K00457	1,699	0.38
Alkanesulfonate monooxygenase (S mobilization)	1.14.14.5	K04091	1,660	0.02
3-phenylpropanoate dioxygenase (Amino acid ox.)	1.14.12.19	K00529	1,540	0.28
4-hydroxyphenylacetate 3-monooxygenase (AA ox.)	1.14.14.9	K00484	1,539	0.09
protoporphyrinogen oxidase (Heme synthesis)	1.3.3.4	K00231	1,488	0.62
cytochrome c oxidase subunit III [4] (Oxphos)	7.1.1.9	K02276	1,433	0.49

KO: KEGG orthology group identifier; N_g: Number of genomes; *V*: verticality value. Numbers in square brackets correspond to protons pumped. Note that the 7,8-dihydroneopterin oxygenase activity in the KEGG orthology group K01633 is an oxygenase side reaction that proceeds through a carbanion intermediate, generating 7,8-dihydroxanthopterin, which is not a central intermediate in the folate synthesis pathway (Czekster and Blanchard, 2012).

Why O_2-dependent reactions when O_2-independent ones already exist?

In Table 3, the most common genes for O_2-dependent enzymes encode steps in the synthesis of essential cofactors including NAD^+, pyridoxal phosphate (PLP), heme, ubiquinone (UQ), and thiamine (Thi). It is known that there are O_2-dependent and O_2-independent biosynthesis pathways for cobalamin, chlorophyll, heme and PLP (Martens et al., 2002; Gomez Maqueo Chew and Bryant, 2007; Mukherjee et al., 2011; Sousa et al., 2013; Dailey et al, 2017; Bryant et al., 2020). Why should organisms evolve O_2-dependent pathways in the presence of preexisting O_2-independent pathways? In the case of parallel O_2-dependent and O_2-independent pathways for chlorophyll synthesis, Gomez Maqueo Chew and Bryant (2007) concluded: "*When a powerful selective pressure, oxygen, apparently inactivated oxygen-independent enzymes in (B)Chl biosynthesis, unrelated proteins with the same catalytic function but with completely different structures and mechanisms, sometimes even using*

450 *oxygen as a substrate, then evolved.*" In other words, the inactivation by O_2 of a preexisting
451 pathway generated the selection pressure for the evolution of an alternative pathway that
452 tolerated the presence of O_2.

453 That same evolutionary reasoning can probably be applied generally to the origin of
454 O_2-dependent pathways for the synthesis of essential cofactors in the presence of preexisting
455 O_2-independent pathways. The O_2-independent pathways have to be older than the O_2-
456 dependent pathways because many prokaryotes that arose prior to the origin of oxygenic
457 photosynthesis require and synthesize NAD^+ (Ollagnier-de Choudens et al., 2005), PLP
458 (Mukherjee et al., 2011), ubiquinone, UQ (Alexander and Young, 1978; Pelosi et al., 2019),
459 thiamine (ThiO vs. ThiH: Leonardi et al. ,2003; Settembre et al., 2003), heme (Dailey et al.,
460 2017), chlorophyll (Gomez Maqueo Chew and Bryant, 2007) or cobalamine (Martens et al.,
461 2002; Dailey et al., 2017). The preexisting O_2-independent routes typically involve enzymes
462 with O_2-sensitive 4Fe-4S clusters, O_2-sensitive pathway intermediates, or both. Thus, O_2-
463 dependent biosynthetic pathways for essential cofactors offered microbes the tools needed to
464 colonize oxic habitats, from which they would have otherwise been excluded.

465 Why would O_2 itself be a substrate in reactions that present alternatives to O_2-sensitive
466 reactions? All O_2-dependent reactions are redox reactions. The same is true for the preexisting
467 O_2-sensitive reactions for which they substituted. Many of the reactions of anaerobes that are
468 inhibited by O_2 entail a radical—an unpaired electron—as a reaction intermediate, on the
469 enzyme, on the substrate, or both (Buckel and Golding, 2006; Buckel and Golding, 2012).
470 Because of its diradical nature, O_2 can interfere as an inhibitor of radical reactions by extracting
471 the radical, but it can also serve as an effective substrate for the generation of radicals in
472 enzymatic reactions (Buckel and Golding, 2006; Buckel and Golding, 2012). The cytochrome
473 P_{450} enzymes (Nelson, 2018) are a prominent example of O_2-dependent enzymes that generate
474 radicals in a wide variety of substrates during the reaction mechanism (Meunier et al., 2004).

Common reactants and products: Electron donors, H_2O, H_2O_2, CO_2 and NH_3

The reactants and products from the KEGG reaction line reveal that O_2 itself is, of course, the most common reactant among the 365 prokaryotic O_2-dependent reactions, followed by H^+-ions (144), water (50) and the electron donors NAD(P)H and reduced ferredoxin (27) (Table 4). Water and the oxidized form of the electron donors are among the most common products accordingly. Because the chemical identity of the organic substrates themselves is usually unique per reaction, very few are recurrent other than co-substrates such as 2-oxoglutarate, reduced flavins, NADH and the like. The frequency of 2-oxoglutarate as a substrate and succinate as a product correspond because of the reactions of 2-oxoglutarate-dependent dioxygenases, which generate CO_2 and succinate as products. Known for their role in O_2 sensing in eukaryotes via the HIF (hypoxia induced factor) pathway and their involvement in cancer proliferation (Losman et al., 2020), they also catalyze a diversity of reactions in prokaryotes including RNA base modifications (Herr and Hausinger, 2018). Catechol, or 1,2-dihyroxybenzol, is a common intermediate in aromatic degradation, occurring in 8 O_2-dependent reactions.

Table 4. Most common reactants and products across 365 O_2-dependent reactions of prokaryotes.

Compound	Reactant frequency	Compound	Product frequency
Oxygen	365	H_2O	181
H^+	144	$NAD(P)^+$	100
NAD(P)H	100	Hydrogen peroxide	80
H_2O	50	CO_2	33
Reduced ferredoxin	27	Ammonia	29
Reduced acceptor	23	Oxidized ferredoxin	27
2-Oxoglutarate	17	Acceptor	23

Reduced FMN	8	Formaldehyde	21
Catechol*	2	Succinate	17
Ammonia*	1	Catechol*	6

* Not among the list of most common compounds and added manually.

The oxidation of amines is very common among the O_2-dependent reactions of prokaryotes. Ammonia was among the most common reaction products, produced by 29 reactions. The prevalence of NH_3 as a product of O_2-dependent reactions suggests a role for O_2 in mobilizing nitrogen as a growth substrate, with 126 of the gene families for O_2-dependent enzymes (16 %) involved in amino acid degradation. O_2-dependent mobilization of organic N from environmental sources has been discussed as a possible key early function of O_2 in the wake of the GOE, as it would have supplied N sequestered in sediment in an O_2-dependent manner, such that N supply would have been independent of nitrogenase and N_2 fixation (Allen et al., 2019). All of the reactions that generate NH_3 also generate H_2O_2, except one, anthranilate, NADPH:oxygen oxidoreductase, which catalyzes the reaction

$$\text{Anthranilate} + O_2 + \text{NAD(P)H} + H^+ \rightarrow \text{Catechol} + CO_2 + \text{NAD(P)}^+ + NH_3.$$

Roughly 20 % of the 365 O_2-dependent reaction of prokaryotes generate H_2O_2. Hydrogen peroxide is a ubiquitous compound in organisms that encounter O_2. It comes from three sources. The first is environmental: hydrogen peroxide is excreted by competing neighbors, after which it diffuses into the cell resulting in an increase in intracellular H_2O_2 concentrations (Imlay, 2008). Another H_2O_2 source is via intracellular nonenzymatic routes, where H_2O_2 can be formed from spontaneous reactions of O_2 with one-electron donating cofactors such as FADH. Although these reactions were thought to occur mainly on respiratory chain components, new findings suggest that is not the case (Seaver and Imlay, 2004; Imlay, 2006). Non-respiratory flavoprotein autoxidation has been suggested as a major source of H_2O_2

(Imlay, 2013). In *E. coli*, H_2O_2 formation in respiring cells typically occurs at a rate of about 0.5 % of total O_2 consumption, or one out of every 200 O_2 molecules respired (Seaver and Imlay, 2004). A major enzymatic route of H_2O_2 generation is the superoxide dismutase reaction which scavenges $O_2^{\bullet-}$. Alternatively, H_2O_2 can be generated in a 1:1 stoichiometry by enzymatic reactions that use O_2 as a substrate. The most common H_2O_2-metabolizing enzymes in the 5655-genome sample are presented in Table 5.

Table 5. Most frequent H_2O_2-metabolizing enzymes among 365 O_2-dependent reactions.

Reaction	H_2O_2-producing enzyme	genomes	*V*
R02670[a]	catalase-peroxidase	4068	0.29
R00009[a]	catalase	3485	0.31
R00357	L-aspartate oxidase (NAD^+ synthesis)	3388	6.22
R00277	pyridoxal 5'-phosphate synthase	3026	0.12
R00475	(S)-2-hydroxy-acid oxidase	2646	0.64
R00275	Superoxide dismutase	2564	0.34
R00360	malate dehydrogenase (quinone)	2095	0.31
R07415	gamma-glutamylputrescine oxidase	1809	0.09
R08702[b]	tRNA mnm5s2U biosynthesis	1789	0.18
R03222	coproporphyrinogen III oxidase	1488	0.62
R02382	monoamine oxidase	1353	0.36
R07364	acireductone dioxygenase (Met salvage)	1054	1.47
R02173	monoamine oxidase	955	0.35
R00610	sarcosine oxidase	853	0.29
R03139	primary-amine oxidase	547	0.47
R01459	cholesterol oxidase	547	0.07
R12356	acyl-CoA oxidase	301	0.02
R07981	factor-independent urate hydroxylase	284	0.24
R00366	D-amino-acid oxidase	208	0.30

[a] the physiological reaction consumes H_2O_2

[b] mnm5s2U: 5-methylaminomethyl-2-thiouridine

Flavins and iron are the most common cofactors

As a consequence of its kinetically stable triplet diradical state, dioxygen has high activation energy barriers (Borden et al., 2017). Since organic substrates are usually in their singlet ground state, direct reactions of oxygen with organic compounds are often spin-forbidden (Klinman, 2001). Because of this, oxygen needs to be activated in order to react with typical organic substrates. In enzymatic reactions, activation is usually provided by the enzyme-guided donation of a single electron to generate a superoxide radical $O_2^{-\cdot}$ or an electron and a proton to generate a perhydroxyl radical $HO_2\cdot$. The one-electron donor is typically either a metal ion such as copper, iron (sometimes in the form of FeS centers), or manganese (Huang and Groves, 2018; Zhang et al., 2022), or an organic cofactor such as a flavin or a pterin (Palfey et al., 1995; Romero et al., 2018). O_2 so activated by transfer of a single electron is extremely reactive, in O_2-dependent enzymes it readily oxidizes a specific substrate at the active site. The cofactors for the 365 reactions for which data could readily be identified are given in S5 Table. Many but not all of them are involved in O_2 activation. Inconsistencies between S3 Table and S5 Table stem from the fact that the reactants and products were counted as written in the reaction line from KEGG, while cofactors were determined using the BRENDA database and the literature, and listed only if the protein subunit in question directly binds the cofactor (see Material and methods).

Table 6. The most common cofactors across 365 O_2-dependent reactions of prokaryotes.

Cofactors	Frequency
Flavins	156
Iron	152
NAD(P)+	84
Ferredoxin	61
Iron-sulphur clusters	57
Hemes	39
Copper	22
Quinones	17

No cofactor or unknown	21
Coenzyme A	11
Cytochromes	7
Pterins	5
Nickel	3
Flavodoxin	3
Rubredoxin	3
Manganese	2
Tetrahydrofolate	2
Zinc	2
Selenium	1
Magnesium	1
S-adenosyl methionine	1
Metal ion (Fe?)	1
Thiamine-pyrophosphate	1

It was not surprising that the most common cofactors involved in our set of oxygen-dependent prokaryotic reactions were flavins and iron. Flavins are versatile coenzymes that perform both one- and two-electron transfers. They serve as cofactors and interact with dioxygen in several enzyme families, including the flavin monooxygenase family. This family includes several enzymes from our set, such as UbiH and UbiI involved in ubiquinone synthesis, PhzS from the phenazine pathway, the cyclohexanone monooxygenase ChnB and others. The most common activating mechanism of the flavin monooxygenases involves the reduced cofactor reacting with dioxygen to form the characteristic C(4a)-(hydro)peroxyflavin intermediate through a semiquinone radical pair (Wongnate et al., 2014, Romero et al., 2018).

Iron is the most common transition-metal cofactor in oxygenases (Wang et al., 2017). A common transition-metal-dependent enzyme family in our set was the Rieske non-heme iron oxygenase family, including naphthalene-1,2-dioxygenase, the steroid hydroxylase KshAB involved in cholesterol catabolism, phthalate-4,5-dioxygenase and others. The oxygenase component of these enzymes is characterized by the presence of a catalytic mononuclear iron

614 center and a Rieske iron-sulphur cluster. The latter accepts electrons from the two-electron
615 donor NADH through a reductase component (often flavin-dependent), sometimes via an
616 additional electron carrier. The electrons need to be transferred to the mononuclear iron center
617 in order to activate dioxygen. All the reaction intermediates have not yet been characterized,
618 but the proposed mechanism includes rearrangements encompassing several oxidation states
619 of the mononuclear iron center and a variety of radical and non-radical intermediates (Barry
620 and Challis 2013).

621 Both flavins and iron individually were present in over 40 % of the enzymes in our set,
622 sometimes co-occurring. As opposed to flavins, pterins (namely tetrahydrobiopterin) were very
623 rarely employed in catalysis in the reactions we studied, occurring in only 5 cases, with
624 annotations pointing to two enzymes: nitric-oxide synthase and phenylalanine
625 monooxygenase. NAD(P)H was very common as an electron-donor (present in 23 % of the
626 reactions), sometimes indirectly providing electrons to the oxygenase through flavins or the
627 reductase component of a multi-component enzyme, as in the Rieske non-heme iron
628 oxygenases.

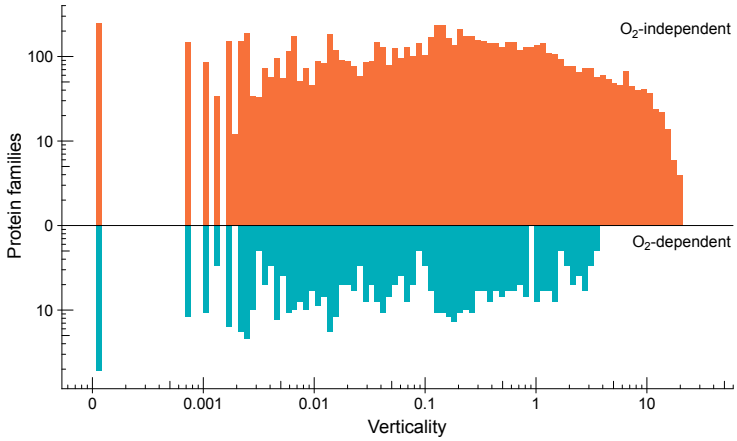
629 Iron-sulphur clusters were present as prosthetic groups in roughly 16 % of the enzymes
630 in our set. Moreover, about 17 % of enzymes from our set bind the soluble electron carrier
631 ferredoxin, a protein in which iron-sulphur clusters represent the redox active chemical group.
632 Cytochromes as electron donors are rare for prokaryotic enzymes. The only S-adenosyl
633 methionine (SAM) binding enzyme identified in our sample was MnmC, a bifunctional enzyme
634 found primarily in γ-Proteobacteria that catalyzes a specific modification of the tRNA wobble
635 base by a mechanism including methylation preceded by oxidative cleavage (Bujnicki et al.,
636 2004, Kim and Almo ,2013).

637 The KEGG list of O_2-dependent reactions in prokaryotes includes 21 having no
638 associated cofactors, or where the cofactors were unknown. A growing number of cofactor-
639 independent oxygenases have recently been described (Widboom et al., 2007, Sciara et al.,
640 2003, Frerichs-Deeken et al., 2004, Grocholski et al., 2010, Baas et al., 2015, Colloc'h et al.,
641 1997). Their mechanisms involve a substrate anion, generally generated by base catalysis,
642 donating a single electron to dioxygen, yielding a stabilized radical pair. The most familiar
643 example of an oxygenase that requires no cofactors to activate dioxygen is the Calvin cycle
644 enzyme RuBisCO, where Mg^{2+} has a role in activating the enzyme, but not in the oxygenase
645 reaction (Bathellier et al., 2020). As the key enzyme of the Calvin cycle, and the most abundant
646 protein in the biosphere, RuBisCO has been extensively studied (Tabita et al., 2008; Tcherkez,
647 2015, Bar-On and Milo, 2019). However, the details of its mechanism, especially its
648 oxygenation activity, are still discussed (Tcherkez 2015, Bathellier et al. 2020). Recent isotope
649 effects studies have shown dioxygen is most likely activated by single-electron transfer from
650 the substrate (Bathellier et al., 2020). After the enolization of ribulose-1,5-bisphosphate, the
651 mechanism presumably proceeds via oxygen addition to the 2,3-enediolate to generate a
652 stabilized peroxo intermediate. From there one oxygen atom is retained in the product
653 phosphoglycolate, the other is lost to the solvent (Tcherkez, 2015). Although the mechanistic
654 details of dioxygen activation are still elusive for many oxygen-dependent enzymes, the
655 chemistry of O_2 prescribes that (one-electron) activation is necessary, which is why O_2 and
656 radical-based reaction mechanisms, which are common to many strict anaerobes (Buckel and
657 Golding, 2006), often lead to enzyme inactivation.

658

659 Genes for O_2-dependent enzymes are transferred more frequently
660 than others

661 To investigate the role of lateral gene transfer (LGT) in the evolution of O_2-dependent
662 enzymes, we utilized values of verticality, V, which provide a measure for how often a gene
663 has been transferred in evolution (Nagies et al., 2020). High values of verticality indicate low
664 levels of LGT for members of a gene family, for example ribosomal proteins, while low values
665 of verticality reflect a high frequency of LGT for a given prokaryotic gene during evolution.
666 With O_2-dependent enzymatic reactions mapped to protein families, we could assign a value
667 of verticality to reactions in order to ask the most pressing question concerning the role of LGT
668 in the evolution of O_2-dependent enzymes: are genes for O_2-dependent enzymes vertically
669 inherited, or are they subject to lateral gene transfer (LGT) like all other prokaryotic genes?
670 The 365 O_2-dependent reactions of prokaryotes map to 792 protein families. Some of these
671 families were too small or too poorly conserved to permit a calculation of verticality. For the
672 547 families of O_2-dependent enzymes in prokaryotes for which we could assign a value of V,
673 the mean verticality or V_{avg} is 0.273 ± 0.626 (avg ± SD). For comparison, the 3,018 O_2-
674 independent reactions in the data map to 11,754 protein families, of which 8,322 remained after
675 discarding small or poorly conserved families. The O_2-independent enzymes have a mean
676 verticality of V_{avg} = 0.781 ± 1.926, the difference in the two distributions is significant at p =
677 $7.43 \cdot 10^{-47}$ (t-value = -14.92, DF = 1,406; Fig 5). The difference remains significant even when
678 higher verticalities (>1) are filtered out ($V \leq 1$, t-value = -6.22, DF = 603, $p = 9.02 \cdot 10^{-10}$; also
679 see S2B and S2C Table).
680



681

682 **Fig 5. Distribution of verticality for protein families catalyzing oxygen-independent and oxygen-**
683 **dependent reactions.** The distribution of 8,322 protein families with verticality values associated with
684 O_2-independent reactions are shown in the upper histogram (orange) and the 547 protein families with
685 verticality values associated with O_2-dependent reactions in the lower histogram (blue). Logarithmic
686 scale, 100 bins. O_2-dependent gene families show a lower verticality due to more frequent violation of
687 monophyly, they have thus been subject to more LGT (see Nagies et al. 2020, t-value = -14.922, DF =
688 1,406.8, $p = 7.43 \cdot 10^{-47}$; S2B Table).

689

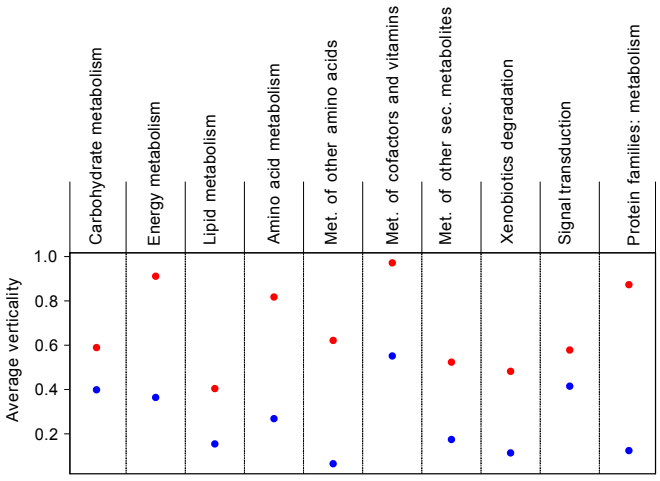
690 To test whether the difference in V_{avg} in Fig 5 might be a result of unequal sample sizes,
691 we used downsampling to generate 100,000 samples of 547 protein families linked to O_2-
692 independent reactions and compared the distribution of V_{avg} of these samples to V_{avg} of the 547
693 protein families associated with O_2-dependent reactions. None of the 100,000 samples
694 generated a value of $V_{avg} \leq 0.273$ (S2D Table). The chance to generate a sample with V_{avg} as
695 extreme as in the O_2-dependent reactions is less than 1/100,000, or $p < 10^{-5}$. Because
696 differences in protein family sizes exist between the two distributions, a binning approach was
697 used to generate samples with protein families of similar size. Again, none of the samples of
698 O_2-independent enzymes reached a value of V_{avg} approaching that of the O_2-dependent

699 reactions (S2D Table). The data clearly indicate that O_2-dependent enzymes have undergone
700 LGT more frequently than O_2-independent enzymes have. Even the more widely distributed
701 protein families and reactions have comparable low verticality (S1 Fig).

702 To see if the observed lateral transfer frequency based on verticality values is an effect of
703 specific functional categories, we compared the verticality of genes for O_2-dependent and O_2-
704 independent reactions within the 10 most frequent functional categories (Fig 6). In each
705 category, O_2-independent reactions have higher verticality than O_2-dependent reactions. The
706 higher the ratio, the more rapidly the genes for the O_2-dependent reactions have spread. The
707 ratios of spread via LGT underestimate the rates of transfer for O_2-dependent reactions because
708 the genes for O_2-independent reactions have had more evolutionary time, up to 4 Ga, to undergo
709 transfer than the O_2-dependent reactions have (up to 2.4 Ga).

710 The impact of O_2 on physiology clearly included respiration employing the six types of
711 terminal oxidases in respiratory chains (energy metabolism, 127 gene families). However, O_2-
712 dependent terminal oxidases do not always conserve energy (see below). Indeed, just as many
713 O_2-dependent enzyme families are involved in amino acid metabolism (126 gene families), or
714 xenobiotic degradation (122 families) (Fig 6) as in energy metabolic pathways. Judging from
715 the frequency of O_2-dependent enzymes across functional categories, the ecological impact of
716 O_2 appears to reside mainly in its utility in providing a high energy co-substrate for the
717 enzymatic oxidation of chemically stable, environmentally available substrates. Relative to O_2-
718 independent reactions within the same functional category, the most rapidly transferred O_2-
719 dependent genes encode products involved in the metabolism of 'other' amino acids (D-amino
720 acids, glutathione, taurine, selenocompounds, etc.), followed by 'protein families metabolism'
721 (various catabolic reactions), xenobiotics degradation (includes cytochrome P_{450} enzymes),
722 amino acid oxidation and breakdown of secondary metabolites. In lipid metabolism, fatty acid
723 and carotenoid oxidation enzymes are common, these are non-fermentable substrates. The

category 'energy metabolism' included enzymes involved in sulfur, methane and nitrogen metabolism as well as enzymes acting in oxidative phosphorylation and carbon fixation in photosynthetic organisms. The lower verticality (higher LGT) of O_2-dependent reactions is not just an effect of accessory genomes, as several functional categories that are typical components of the accessory genome, such as membrane transport and drug resistance (Croll and McDonald, 2012; Jackson et al., 2011; López-Pérez and Rodriguez-Valera, 2016) were devoid of O_2-dependent reactions in the present sample.



KEGG functional category	O2-dependent Protein families	O2-dependent V_{avg}	O2-independent Protein families	O2-independent V_{avg}	Ratio
Metabolism of other amino acids	27	0.065	548	0.622	9.55
Protein families: metabolism	41	0.124	1270	0.873	7.02
Xenobiotics biodegradation and metabolism	122	0.114	627	0.482	4.23
Amino acid metabolism	126	0.269	1804	0.818	3.04
Metabolism of other secondary metabolites	46	0.175	569	0.524	3.00
Lipid metabolism	41	0.155	1039	0.404	2.62
Energy metabolism	127	0.364	1138	0.911	2.50
Metabolism of cofactors and vitamins	75	0.552	1406	0.972	1.76
Carbohydrate metabolism	46	0.399	2564	0.589	1.48
Signal transduction	48	0.414	355	0.578	1.40

Fig 6. Average verticality across the 10 functional categories with the highest frequency of O_2-utilizing protein families. Average verticality values (V_{avg}) for O_2-dependent (blue points) and O_2-independent (red points) protein families by functional category. The last column shows the ratio of

735 average verticality between O_2-independent reactions and O_2-dependent reactions. Values for all
736 functional categories can be found in S6 Table.

737

738 The consistent pattern that O_2-dependent enzymes are more rapidly transferred irrespective
739 of functional category and substrate type (Fig 6) indicates that O_2-dependent enzymes are more
740 frequently retained as the result of transfers, suggesting that they carry some kind of selective
741 advantage. The O_2-dependent reactions are highly exergonic. The mean ΔG' for O_2-dependent
742 reactions in the present sample is –234.3 kJ·mol^{-1}, the mean ΔG' for O_2-independent reactions
743 is on the order of –12.6 kJ·mol^{-1} (see S2 Fig).

744 To further examine the role of LGT in the evolution of O_2-dependent reactions, we plotted
745 the number of gene origins on branches that correspond to the last common ancestor of all
746 phyla possessing a specific O_2-dependent reaction (Fig 7). In this kind of analysis, a gene origin
747 is an occurrence on branches either side of the node. An origin can reflect presence in the
748 common ancestor or lateral transfer among branches spanning the node, which in the case of
749 these enzymes, given their very low verticality, is the default interpretation. Without debating
750 the specific tree topology, of note is the circumstance that cyanobacteria do not have a basal
751 position in Fig 7, nor do they tend to have a basal position in any current prokaryotic phylogeny
752 spanning both bacteria and archaea. That is sensible from a physiological standpoint, because
753 photosynthesis in cyanobacteria requires that cytochromes (heme synthesis), quinones and a
754 fully functional respiratory chain were in place in the ancestors of cyanobacteria, meaning that
755 anaerobic respiration had to precede the origin of chlorophyll based photosynthetic electron
756 transport. Because cyanobacteria produced the O_2 that O_2-dependent enzymes use, any
757 distribution of O_2-dependent enzymes across nodes predating the origin of cyanobacteria are
758 the result of LGT.

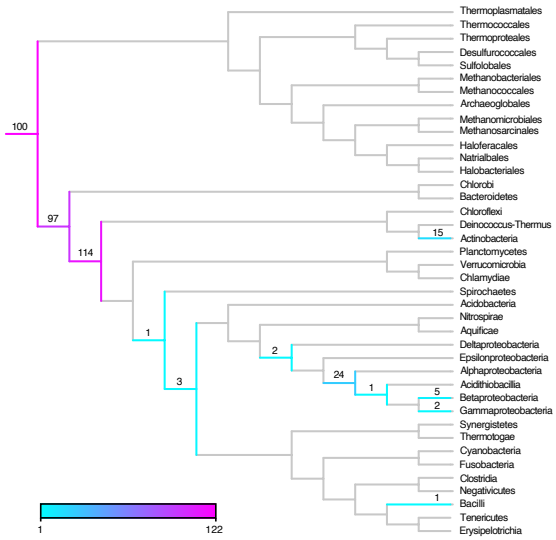


Fig 7. Origins of O_2-dependent reactions across a backbone phylogeny. The origins of 365 O_2-dependent reactions across a backbone phylogeny containing bacterial and archaeal phyla. A node is defined as an origin node when it is reconstructed as the last common ancestor of all phyla possessing a given O_2-dependent reaction. Number of origins per node is indicated by the color scale. Branches and nodes in which no O_2-dependent reactions originated are in grey. See also S7 Table.

As observed in earlier studies using smaller samples (Sousa et al., 2016; Jabłońska and Tawfik, 2019), the genes for O_2-dependent enzymes have been distributed by LGT. In Fig 7, one quarter of the 365 O_2-dependent reactions (100) have their origin in the root node of the tree, that is, they were transferred between bacteria and archaea, for example genes for enzymes involved in the bacterial type respiratory chain of haloarchaea (Nelson-Sathi et al. 2012), while another 97 O_2-dependent reactions are reconstructed to the ancestor of all bacterial species, and 114 O_2-dependent reactions have their origin at the deep bacterial node separating Chlorobi

774 and Bacteroidetes (CFB) from all other bacterial species. Both Chlorobi and Bacteroidetes are
775 primarily anaerobic groups but have acquired O_2-dependent enzymes. Additionally, the origin
776 of 24 O_2-dependent reactions is found at the ancestral node of the clade containing
777 Alphaproteobacteria, Acidithiobacillia, Betaproteobacteria and Gammaproteobacteria. All
778 other branches showing origins in Fig 7 range from one to three origins. Without LGT, this
779 distribution would suggest that roughly one quarter of prokaryotic O_2-dependent reactions were
780 present in the last common ancestor of all cells. Yet 78 out of the 100 origins at the root node
781 are the consequence of the same O_2-dependent reaction being present in bacterial species and
782 halophilic archaea. This is not surprising, because halophilic archaea are derived from
783 methanogens and have acquired genes for respiration via LGT (Nelson-Sathi et al., 2012).
784 Similar bacteria-to-archaea transfers also apply in the case of the Sulfolobales, which are
785 aerobic thermoacidophiles (Leigh et al., 2011) derived from anaerobic ancestors. Fig 7 shows
786 how LGT complicates the evolutionary reconstruction of O_2-dependent reactions. Current
787 phylogenies of archaea place the root of the archaeal tree within methanogens (Mei et al.,
788 2023), which are strict anaerobes. The first cells had to be anaerobes, because O_2 is a product
789 of biochemical evolution.

790

791 Discussion

792 The appearance of environmental O_2 following the GOE confronted prokaryotic life
793 with a few challenges and numerous chemical opportunities. The main challenge was that a
794 very small number of enzymes in anaerobes—sometimes only one enzyme per species, but
795 often in physiologically key positions in metabolism—are poisoned by contact with O_2 or
796 activated forms thereof (Lu and Imlay, 2021). Such enzymes, for example pyruvate:ferredoxin
797 oxidoreductase, a key enzyme in carbon metabolism of anaerobes, or nitrogenase, the

biosphere's N_2 fixing enzyme (Schlesier et al., 2016), typically have either a radical mechanism (Buckel and Golding, 2006) or harbor low potential metal centers, very often FeS centers, that can spontaneously react with O_2 (Imlay, 2006), or both. The O_2-dependent inactivation of essential enzymes can arrest growth until anoxia is restored, which allows the organism to replace the poisoned enzyme by repair or resynthesis (Lu and Imlay, 2021). The opportunities presented by the advent of O_2 resided in the vast amount of chemical energy stored in the O_2 diradical molecule.

The evolutionary significance of O_2 in evolution: energy efficiency?

The evolutionary significance of O_2 is traditionally viewed in terms of energy efficiency, O_2 having enabled improved ATP yield from heterotrophic substrate breakdown. The underlying reasoning often being that "*life with oxygen is better than life without: a given amount of glucose processed in the presence of oxygen produces 18 times as much energy as the same amount of glucose processed without oxygen*" (Rytkönen, 2018). The factor 18 can be questioned. Many organisms stick with fermentations even in the presence of oxygen, some heterotrophic fermenters gain 4 ATP per glucose in the absence of O_2, others gain 2 ATP per glucose in the presence of O_2 (Tielens et al., 2002; Martin et al., 2017), and many microbes do not grow on glucose as their natural substrate. *E. coli* gains 15 ATP per glucose during O_2 respiration under optimal conditions and 4 ATP per glucose from anaerobic fermentation (Szenk et al., 2017), the difference in maximal ATP yield is a factor of 3.8 (not 18).

Studies from *Escherichia coli* indicate that life is actually not better with O_2, because when *E. coli* has the opportunity to obtain roughly 30 ATP per glucose from aerobic respiration as opposed to 4 ATP from acetate fermentation, *E. coli* preferentially expresses forms of respiratory complex I (*ndh*) that do not conserve energy by pumping protons. As a

822 consequence, *E. coli* obtains about 15 rather than 30 ATP per glucose under pure energy
823 respiratory metabolism (Szenk et al., 2017). The majority of *E. coli*'s energy gain in the
824 presence of excess glucose and O_2 is from acetate producing fermentations (Wolfe, 2005;
825 Basan et al., 2015) that typically generate 4 ATP per glucose via substrate level
826 phosphorylation (Szenk et al., 2017). That is, when given the option, *E. coli* actively declines
827 the opportunity for extra energy yield from O_2 (Unden and Bongaerts, 1997). O_2 presence does
828 not always mean more energy, and more energy is not always in line with the physiological
829 needs of the cell. The membrane potential, $\Delta\Psi$, used for substrate import and ATP synthesis
830 under anaerobic growth, –130 mV, is roughly the same as under aerobic growth –140 mV
831 (Tran and Unden, 1998).

832 Yeast exhibits a similar response as *E. coli*, the Crabtree effect, the preference for
833 fermentation in the presence of glucose and oxygen, which results in an increased rate of ATP
834 low efficiency production (Pfeiffer and Morley, 2014), not an increased efficiency of ATP
835 production. Humans running the 100-meter dash rely on ATP synthesis from glycogen
836 fermentation, which generates ATP much faster than mitochondrial respiration, but rapidly
837 leads to lactate accumulation and muscle cramps. Although oxygen dependent reactions have
838 been suggested to offer better adapted alternatives for aerobic and anaerobic conditions
839 (Ouchane et al., 2004), to be more generally efficient than their O_2-independent versions
840 (Raymond and Blankenship, 2005), or to generate higher pathway fluxes and to avoid ATP
841 losses (Jabłońska and Tawfik, 2019), simple observations from *E. coli* and yeast growth run
842 counter to the view that evolutionary significance of O_2-dependent reactions has to do with
843 efficiency.

844 In organisms that respire O_2, there exists a diversity of terminal O_2 reductases for the
845 respiratory chain that exhibit different degrees of energy conservation. Of the six families of
846 O_2-utilizing terminal oxidases (Degli Esposti et al., 2019), the *aa3* type cytochrome *c* oxidases

847 (A1 and A2 families) and the *bo*$_3$ type ubiquinol oxidase pump 4 protons per O_2 reduced, the
848 *cbb*$_3$ type cytochrome *c* oxidases usually pump 2 protons per O_2, while the *bd* type oxidases
849 and the alternative oxidase (AOX) pump 0 protons per O_2 (Han et al., 2011; Degli Esposti et
850 al, 2019), although the *bd* oxidases generate protonmotive force (Borisov et al. ,2011). The
851 subunits of non-pumping *bd* type oxidases are among the most widespread O_2-dependent
852 enzymes in the present genome sample, more widespread than those of the terminal oxidases
853 that pump 2 or 4 protons per O_2 reduced (Table 3).

854 A recent in-depth analysis of the B family of cytochrome *c* oxidases (Murali et al.,
855 2021) revealed that within this family alone there are two subfamilies that pump 4 protons per
856 O_2, 12 subfamilies that pump 2 protons per O_2, and five subfamilies that do not pump protons.
857 As with the Crabtree effect (2 ATP per glucose via cytosolic glycolysis in the presence of O_2),
858 this spectrum of energy conservation within O_2-dependent terminal oxidases from 0 to 4
859 protons pumped per O_2 reduced does not suggest that the main physiological role of O_2 in
860 evolution was improved energy yield. The use of O_2 does not strictly correlate with improved
861 energy yield from heterotrophic substrates. The use of O_2 is, however, linked with redox
862 balance, ensuring that the electrons exit metabolism at exactly the same rate as they enter (Allen
863 and Raven, 1996; Allen, 2005; Sies, 2015), otherwise metabolism comes to an immediate halt
864 for lack of NAD^+ or NADH to mediate substrate oxidations. Hence, even minor alterations of
865 redox balance lead to oxidative stress (Sies et al., 2017).

866

867 The evolutionary significance of O_2 in evolution: why no O_2-
868 dependent SLP?

869 Another observation that does not square with the idea that the main evolutionary role
870 of O_2 was improved energy efficiency concerns substrate level phosphorylation, SLP. Virtually
871 all O_2-dependent reactions with natural organic substrates are sufficiently exergonic to drive
872 SLP, but O_2-dependent SLP reactions are almost unknown. Assuming that ~70 kJ·mol^{-1} is
873 needed for the synthesis of one ATP (Thauer et al., 1977), and given that the average calculated
874 value of ΔG for the reactions that delivered an estimate using eQuilibrator in our sample was
875 –234,3 kJ·mol^{-1}, there is enough energy on average to synthesize roughly 3 ATP per O_2 per
876 reaction, or at least 1 ATP at a poor efficiency. Out of 365 prokaryotic O_2-dependent reactions
877 in KEGG, we could identify only one (0.3 %) that generates a product capable of supporting
878 SLP. The reaction is catalyzed by the H_2O_2-producing (phosphorylating) pyruvate oxidase
879 [EC:1.2.3.3] (Muller and Schulz, 1993), POX, which converts pyruvate and O_2 to CO_2, H_2O_2
880 and acetyl phosphate with a calculated $\Delta G'$ of –158 kJ·mol^{-1}. Acetyl phosphate can generate
881 ATP via SLP in the reaction catalyzed by acetate kinase (Decker et al., 1970). It has been
882 suggested that POX might generate a phosphoryl radical in the reaction mechanism (Tittman
883 et al., 2005), however the function of O_2 in the reaction appears not to be direct involvement
884 in the phosphorylation reaction, but simply the conventional reoxidation of $FADH_2$ (Frey et al.,
885 2006).

886 H_2O_2-producing pyruvate oxidase has a very narrow phylogenetic distribution,
887 occurring in only 50 (1 %) of the 5655 genomes sampled, almost exclusively among members
888 of the *Lactobacilli*, and it was the only enzyme in our sample that used inorganic phosphate,
889 P_i, as a substrate or generated a high energy organophosphate bond. A second more common
890 pyruvate oxidase, the enzyme that occurs in *E. coli* [EC:1.2.5.1], is present in 43 % of the
891 genomes in our sample. It is not O_2-dependent, does not use phosphate, uses UQ as the oxidant
892 and generates acetate and CO_2 (Abdel-Hamid et al., 2001; Cornacchione and Hu, 2020). If O_2
893 were a means to improve energy yield, why would cells squander the ca. –243 kJ per mol of

894 O_2 of free energy associated with >99 % of O_2-reducing reactions? The phosphorylating
895 pyruvate oxidase (POX) reaction demonstrates that it is possible for soluble enzymes to couple
896 O_2 reduction to the synthesis of compounds that can phosphorylate ADP. The enzymatic
897 reactions of O_2 in metabolism provide no hints that the initial role of O_2 in evolution involved
898 improved energy conservation.

899 In the list of 365 O_2-dependent reactions, phosphorylated compounds were generally
900 very rare as substrates, occurring in only six reactions (1 %), whereby only in the pyruvate
901 oxidase reaction was a phosphate bond ultimately affected. Among the other five was the
902 reaction catalyzed by the most abundant enzyme in nature, ribulose-1,5-bisphosphate
903 carboxylase/oxygenase. Its O_2-dependent oxygenase activity generates 3-phosphoglycerate
904 and 2-phosphoglycolate in a highly exergonic reaction with a calculated $\Delta G'$ of –516 kJ·mol^{-1}.
905 The enzymatic inefficiency of RuBisCO as a CO_2 fixing enzyme is well known (Ellis, 1979;
906 Seah et al., 2019), but that it also wastes the energy in O_2 is less broadly appreciated. It is not
907 obvious (to us) why O_2-dependent reactions are not widely coupled to SLP. Perhaps O_2-
908 dependent respiratory chains evolved before any selective pressure to gain energy from O_2-
909 dependent SPL arose.

910

911 The evolutionary significance of O_2 in evolution: eukaryote origin?

912 Another suggestion is that O_2 enabled the expansion of novel biosynthetic pathways so
913 as to foster the origin of complex life: eukaryotes (Margulis, 1970; Raymond and Segrè, 2006).
914 Yet, if O_2 fostered the origin of cellular complexity, avidly O_2 respiring prokaryotes like
915 proteobacteria and cyanobacteria would be as complex as eukaryotes, which is clearly not the
916 case (Lane and Martin, 2010). Instead, proteobacteria and cyanobacteria gave rise to the
917 bioenergetic organelles of eukaryotes. The complexity of eukaryotic cells has more to do with

the origin of mitochondria as a compartment (Gould et al., 2016; Raval et al., 2022) than with the origin of O_2 as a terminal acceptor, and many mitochondria of complex cells do not use O_2 at all, having lost their respiratory chains during ecological specialization to anaerobic niches (Müller et al., 2012). The origin of O_2 preceded the origin of eukaryotes by a billion years, and eukaryotes evolved for another billion years before O_2 rose to present levels (Zimorski et al., 2019; Mills et al., 2022) hence no direct causal connection between eukaryote origin and the emergence of O_2 can be readily drawn. Another proposal is that environmental O_2 availability has impacted the frequency of oxygen atoms in amino acids of proteins (Acquisti et al., 2007; Vieira-Silva et al., 2008), the problem being that in amino acid biosynthesis, none of the O atoms in amino acids actually stem from O_2, they all stem from CO_2 and H_2O (except for one O in the tyrosine salvage pathway).

Oxygen-dependent enzymes spread rapidly in prokaryotic evolution

Lateral gene transfer is germane to the role of O_2 in evolution. Investigations suggesting that oxygen respiration traces to the first cells (Castresana and Moreira, 1999) entail the assumption that terminal oxidases might have been vertically inherited among prokaryotic lineages. Considerable evidence indicates a role for lateral gene transfer (LGT) at least in the distribution of O_2-dependent terminal oxidases across prokaryotic lineages (Borisov et al., 2011; Degli Esposti, 2020; Nelson-Sathi et al., 2012; Osborne and Gennis, 1999). A recent molecular clock study suggested that the origin of O_2 might predate its appearance in the geochemical record (Jabłońska & Tawfik, 2021), but assumed that the genes for the O_2-utilizing enzymes in question had not undergone LGT since their origin over 2 billion years ago. Most genes in prokaryotic genomes have undergone LGT, with 97 % of all genes having

942 undergone at least one case of LGT between bacteria and archaea (Weiss et al., 2016). No gene
943 family present in prokaryotic genomes has been immune to LGT between prokaryotic phyla
944 during evolution (Dagan et al., 2008; Nagies et al., 2020) whereby Arnold et al. (2021) went
945 so far as to surmise that LGT "*is the most conspicuous feature of bacterial evolution*".

946 We found that O_2-dependent enzymes have undergone LGT more than O_2-independent
947 enzymes have (Fig 5) and that this is true across all functional categories (Fig 6). The only
948 plausible reason for the rapid spread of genes for O_2-dependent enzymes is that they conferred
949 a physiological advantage to organisms that acquired and retained them. The nature of that
950 advantage, measured in terms of the frequency of O_2 consuming reactions, was the breakdown
951 of stable bonds in aromatic and nitrogenous compounds, mobilizing substrates. This seems
952 curious at first sight because there are anaerobic pathways that fulfill the same purpose.
953 However, the anaerobic (or O_2-independent) pathways typically involve enzymes with FeS
954 clusters that are inhibited in the presence of O_2. O_2-dependent enzymes can fulfil the substrate
955 mobilizing function in oxic environments, affording the organism access to a new niche.
956 Overall, the data suggest that the first cells to utilize O_2 were in competition for substrates, not
957 for energy.

958

959 Geochemical proposals for the persistence of low O_2 following the 960 GOE

961 A recurring question about O_2 in evolution is: Why did O_2 levels stay low for 2 billion
962 years following the GOE? One proposal posits a steady supply of geochemical reductants from
963 within the Earth, such as Fe^{2+} or S^{2-}, that consumed O_2 either enzymatically or non-
964 enzymatically. The reductants were emitted from the mantle at rates and in amounts that

965 aligned with cyanobacterial O_2 production so as to keep O_2 levels low and constant for 2 billion
966 years without variance or interruption (Poulton et al., 2004b; Canfield et al., 2008). A second
967 proposal has it that anoxygenic phototrophs out-competed oxygenic cyanobacteria for light or
968 for nutrients such as phosphorus (Ozaki et al., 2019). In order to outcompete cyanobacteria for
969 any nutrient, the 'more successful' anoxygenic phototroph would, however, first require more
970 carbon for cell mass than cyanobacteria, meaning a supply of reductant for CO_2 fixation that is
971 more abundant than water.

972 A third and very popular proposal posits that nutrient limitations, in particular
973 molybdenum (Mo), led to limited O_2 production by limiting photosynthetic biomass (Anbar
974 and Knoll, 2002; Moore et al., 2017; Stüecken, 2013). Yet when molybdenum is limiting,
975 cyanobacteria can use either Fe or V as a substitute for Mo in their alternative nitrogenases and
976 resume growth unimpaired (Bothe et al., 2010), such that microbes readily overcome the theory
977 of Mo limitation by well-known routes. A fourth proposal has it that animals affected the
978 degree of mixing between nutrient-rich reservoirs and the photic zone, for example, through
979 animal burrowing activity (Boyle et al., 2014; Canfield and Farquhar, 2009) or by grazing
980 activity of early animals to improve light penetration into the photic zone, increasing O_2
981 production to end the Pasteurian epoch (Butterfield, 2015a). A fifth proposal has it that no
982 mechanism at all is required to account for Earth's stepwise oxygenation, it is an inherent
983 property of biogeochemical cycling as calculated by the model (Alcott et al. 2019). A sixth
984 proposal invokes sudden changes in the magnitude of tides and changes in day length that
985 would somehow impact O_2 accumulation once it set in (Klatt et al., 2021), thereby limiting the
986 planet's oxygenation.

987 Looking at the issue from the standpoint of physiology, one can ask whether the same
988 factor that caused oxygen to appear might have also kept it low for 2 billion years: a
989 cyanobacterial enzyme. The oxygen evolving complex of photosystem II synthesizes O_2, but

990 nitrogenase is inhibited by O_2. The latter can readily limit O_2 levels at a global scale (Allen et
991 al., 2019).

992

993 Nitrogenase inhibition by O_2 kept Earth's O_2 levels low for 2 billion
994 years

995 Nitrogenase feedback inhibition autoregulates atmospheric O_2 levels above cyanobacterial
996 cultures. By dry weight, cyanobacterial cells consist of roughly 50 % carbon and 10 % nitrogen
997 (Fagerbakke et al., 1996). The first cyanobacteria had water as unlimited reductant for CO_2
998 fixation, but, for net growth to occur, N_2 incorporation had to keep pace. Nitrogenase reduces
999 N_2 to NH_3 but the enzyme is inhibited by O_2. The mechanism of nitrogenase inhibition by O_2
1000 is simple: O_2 oxidizes the enzyme's FeS clusters (Fig 8). There are three different FeS clusters
1001 in nitrogenase (Burgess and Lowe, 1996; Hu and Ribbe, 2015). All three are essential for the
1002 protein's activity and all three can be rapidly oxidized by O_2. The nitrogenase from *Azotobacter*
1003 *chroococcum*, like the enzyme from other sources, is rapidly inactivated by O_2. The very O_2-
1004 sensitive 4Fe4S cluster of the Fe protein (FeP) is afforded some degree of protection from O_2
1005 oxidation by binding of the Shethna protein, which has a 2Fe2S cluster that covers the 4Fe4S
1006 cluster of FeP (Rutledge and Akif Tezcan, 2020; Schlesier et al., 2016) under oxidizing
1007 conditions. But that protection is ephemeral. In the presence of air, the 4Fe4S cluster in the Fe
1008 protein (FeP, NifH) of *Azotobacter* nitrogenase has a half-life of about 30-60 seconds while
1009 the FeS clusters in the catalytically active MoFe protein (NifDK) have a half-life of about 5-
1010 10 minutes (Robson 1979). Both components of nitrogenase precipitate immediately if the FeS
1011 clusters are degraded by O_2 (M. Ribbe, pers. comm.).

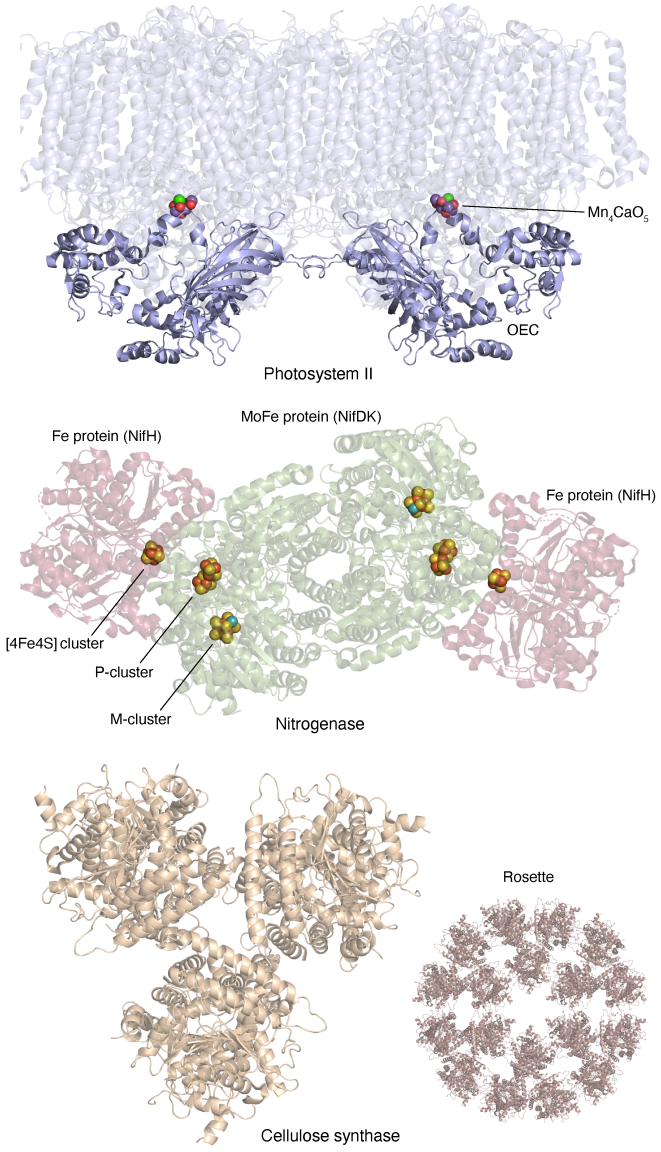


Fig 8. The enzymes that governed atmospheric oxygen concentrations. From top to bottom: cyanobacterial photosystem II (PDB ID: 7D1T) (Kato et al., 2021), nitrogenase (PDB ID: 1G20) (Chiu et al., 2001) and cellulose synthase (PDB ID: 6WLB) (Purushotham et al., 2020). The oxygen evolving complex (OEC) of photosystem II is highlighted, including the Mn_4CaO_5 cluster and the extrinsic proteins PsbO (manganese stabilizing protein), PsbU and PsbV (cytochrome c_{550}) (Shen and Inoue,

1993; Nelson and Yocum ,2006). The nitrogenase MoFe protein (NifDK), an $\alpha_2\beta_2$ tetramer, is shown in green, while the two Fe proteins (NifH), each a γ_2 dimer, are in red. The metal clusters are shown, namely the M-cluster [$MoFe_7S_9$] and the P-cluster [Fe_8S_7] of the MoFe protein, as well as the [Fe_4S_4] cluster of the Fe protein (Jasniewski et al., 2018). The cellulose synthase homotrimer is shown, whereby several of these trimers further organize in the membrane into a rosette – a hexamer of trimers (Purushotham et al., 2020). The structures are symmetrical; therefore, all elements were marked only once for clarity. Photosystem II is viewed from within the membrane plane, while cellulose synthase is viewed from above (in relation to their position in the membrane). The figure was prepared with PyMol (The PyMOL Molecular Graphics System, Version 2.5.4, Schrödinger, LLC).

The atmospheric O_2 levels at which the FeS clusters of nitrogenase are inactivated by O_2 are decisive for regulating the level of O_2 in the Earth's Proterozoic atmosphere. The threshold of atmospheric O_2 concentration above which nitrogen fixation ceases completely is about 10 % O_2, but below 1 % O_2 nitrogenase remains measurably active. This simple biochemical mechanism autoregulates O_2 levels in the air above a cyanobacterial culture, as seen with nitrogen-fixing cyanobacteria such as *Plectonema*. When *Plectonema* cultures are grown under a CO_2-N_2 atmosphere with N_2 as the sole N source, with sufficient CO_2 and light, they grow and accumulate about 0.5 to 1 % O_2 but no more than 2 % O_2 in the gas phase above the culture (Stewart and Lex, 1970; Weare and Benemann, 1974; Rippka and Waterbury, 1977; Rai et al., 1992; Misra, 1999; Berman-Frank et al., 2001, 2003; Staal et al., 2007). This is because air with 1 % O_2 inhibits nitrogenase activity by about 41 % while air with 10 % O_2 inhibits nitrogenase activity completely (Stewart and Lex, 1970).

This autoregulatory circuit (Fig 9) maintains a low O_2 level above the cyanobacterial culture, max. 2 % [v/v] (or 10 % PAL). This O_2 level remains constant during prolonged cultivation. If nitrogenase is inactivated by higher O_2 partial pressure, there is no fixed N to support cell mass accumulation (O_2 production has to balance CO_2 fixation). With less O_2,

nitrogenase activity increases, allowing more CO_2 fixation hence more O_2 production. The resulting circuit maintains O_2 in the atmosphere above the culture below 2 % O_2 or 10 % PAL and is governed by O_2-sensitive FeS clusters in nitrogenase.

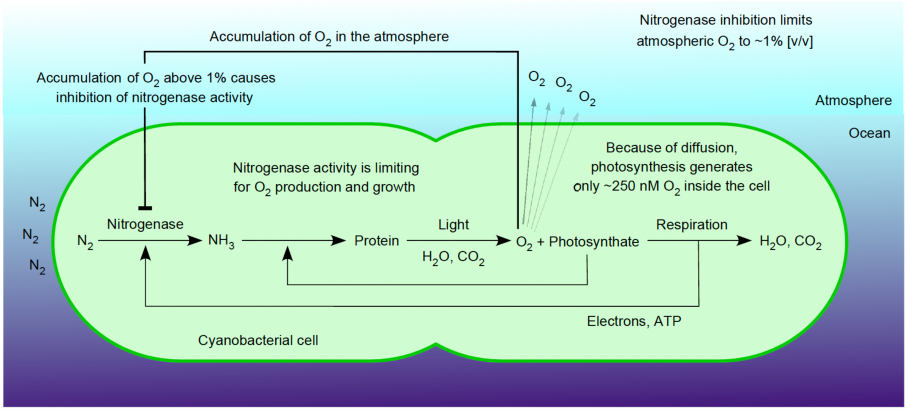


Fig 9. Nitrogenase inhibition by O_2 limits atmospheric O_2 to the Pasteur point during the boring billion (or Pasteurian). See text. Modified from Allen et al. (2019).

In order to generate constant O_2 levels, O_2 production by cyanobacteria had to be counterbalanced by O_2 consumption. Respiration is ubiquitous among all cyanobacterial lineages (Scherer et al., 1988), though its roles are still incompletely understood (Shimakawa et al., 2021). Respiration is essential for N_2 fixation in many (Bergman et al., 1997), if not all, diazotrophic cyanobacteria. *Plectonema* obtains the reducing equivalents and ATP needed for N_2 fixation from respiration (Weare and Benemann, 1974; Misra, 1999). Unicellular N_2 fixing cyanobacteria like diazotrophic *Cyanothece* or *Synechococcus* strains, which have a diel nitrogenase protection mechanism, use respiration of photosynthate as the source of energy and electrons for N_2 fixation (Bandyopadhyay et al., 2013; Rabouille et al., 2014). Respiration is required for N_2 fixation in diazocytes of organisms such as *Trichodesmium* (Sandh et al., 2012),

1061 and in heterocysts, in which respiration provides electrons and ATP for nitrogenase (Torrado
1062 et al., 2019). Of course, before the advent of O_2, the anaerobic ancestors of cyanobacteria,
1063 protocyanobacteria, must have obtained their electrons and energy for N_2 fixation from
1064 anaerobic pathways (Martin et al., 2018). This is still observed in *Synechocystis* strains isolated
1065 from terrestrial hot springs, which obtain the electrons and ATP for N_2 fixation from
1066 fermentations (Steunou et al. 2006). The mechanism of nitrogenase inhibition—spontaneous
1067 (nonenzymatic) oxidation of FeS clusters by O_2—is used by various proteins of bacteria,
1068 including the transcriptional regulator FNR that serves as O_2-sensor (Unden et al., 2002; Barth
1069 et al. 2018), and it is also used by the FeS-containing O_2-sensing protein aconitase in
1070 vertebrates (Gruer et al., 1997; Lushchak et al., 2014; Gunawardena et al., 2016).

1071 The cyanobacterial mechanisms to protect nitrogenase from O_2 include day night (diel)
1072 expression cycles that result in nitrogenase activity in the dark when photosynthetic O_2
1073 production stops (Mitsui et al., 1986), filaments that aggregate to protect N_2-fixing cells called
1074 diazocytes (Bergman et al., 2013), and heterocysts, specialized N_2-fixing cells of filamentous
1075 cyanobacteria that do not produce O_2 (Mitsui et al., 1986; Rippka et al., 1979). The three
1076 different forms of O_2 protection mechanisms to circumvent nitrogenase inhibition by O_2 arose
1077 independently in cyanobacterial evolution (Allen et al., 2019) in a lineage specific manner.
1078 This indicates in turn that cyanobacteria evolved these nitrogenase protection mechanisms in
1079 response to globally rising O_2 levels, not in response to endogenous O_2 production. In support
1080 of that interpretation, endogenous O_2 production in cyanobacteria generates intracellular O_2
1081 concentrations of only 250 nM (Kihara et al., 2014), 1000 times lower than 250 μM, the
1082 modern O_2 concentration in cold, well aerated water and ~100 times lower than the O_2
1083 concentration needed to inhibit nitrogenase; cyanobacteria would have no reason to evolve an
1084 O_2 protection mechanism against endogenous O_2 production. Second, the fossil record
1085 indicates that cyanobacteria evolved nitrogenase O_2 protection mechanisms very late in

evolution, when O_2 rose to approximately modern levels. Cyanobacteria have a relatively good fossil record (Demoulin et al., 2019) but the oldest fossil heterocysts are only 418 million years old (Butterfield 2015b), younger than the oldest land plants. This indicates that cyanobacteria evolved mechanisms of nitrogenase protection against O_2 in response to environmental O_2 levels that were persistently above the 1 % O_2 threshold for nitrogenase inhibition. In the wake of the GOE, nitrogenase inhibition could have kept atmospheric O_2 levels globally low for 2 billion years. How did O_2 levels overcome nitrogenase inhibition? It was the advent of land plants and the activity of one more enzyme: cellulose synthase (Fig 10).

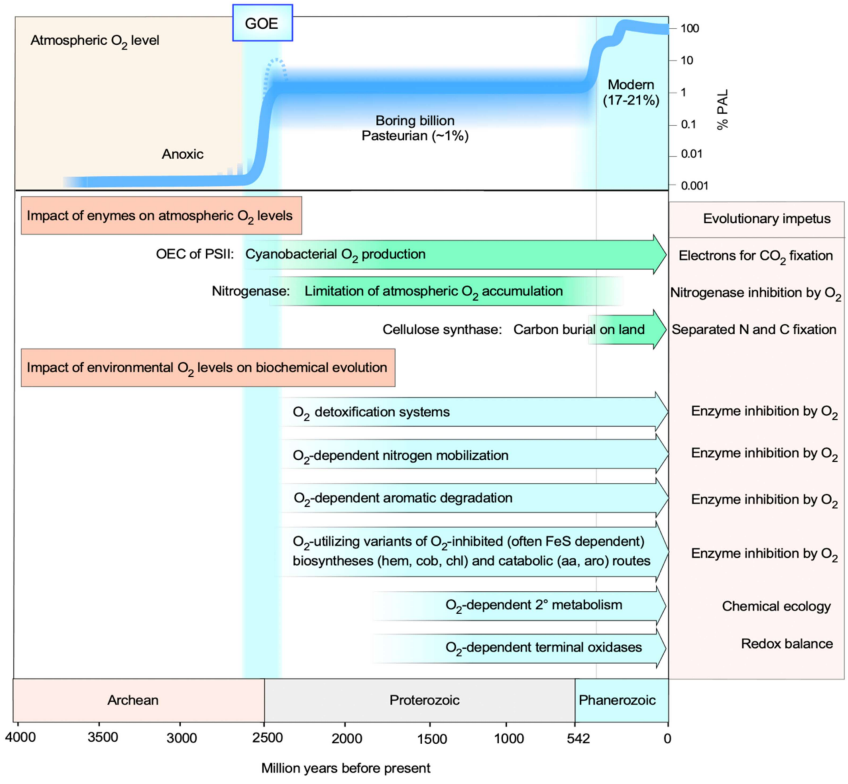


Fig 10. Oxygen levels and enzyme evolution. The three key enzymes are the oxygen evolving complex (OEC) of photosystem II, the nitrogenase and the cellulose synthase. PAL: present atmospheric level. The OEC, once evolved in cyanobacteria, produced oxygen, which accumulated but only to about 1 % PAL, a level that stayed stable for the next 2 billion years, a period called the boring billion to emphasize geochemical stasis or the Pasteurian (Martin et al., 2020) to emphasize low oxygen availability to microbes during that time. Probable cause of low O_2 during the Pasteurian is O_2 inhibition of nitrogenase by oxygen at 1 % PAL (see text). Inhibition of cyanobacterial nitrogenase through inactivation of FeS centers by 1 % O_2 kept O_2 low (Allen et al., 2019). O_2 inhibited other enzymes as well, restricting prokaryotes with O_2-sensitive enzymes to anaerobic niches until the evolution of O_2 detoxification mechanisms and/or O_2-tolerant (often O_2-utilizing) pathways for cofactor biosynthesis, aromatic degradation, nitrogen mobilization, secondary metabolism and respiratory chains that use oxygen as the final electron acceptor. With the evolution of land plants and subsequent large-scale carbon burial due to cellulose synthesis and the physical separation of N_2 fixation in soil from CO_2 fixation (and O_2 production) in leaves, oxygen levels rose towards modern levels (see text).

Cellulose synthase caused O_2 to rise to current levels

Oxygen levels rose as life was starting to emerge on land about 500 million years ago. The exact timing of the final surge in O_2 evolution is still discussed, but newer reports (Stolper and Keller, 2018) suggest that deep-ocean oxygenation occurred only 541 million years ago and perhaps less than 420 million years ago. This is in line with other recent studies (Lenton et al., 2016) that also implicate land plants as the cause of late O_2 rise in Earth history (for a recent review see Mills et al., 2022). Current estimates have it that the first land plants arose roughly 500 MY ago (Morris et al., 2018), which is in line with the evidence from fossil spores that indicate the presence of land plants in the Cambrian (Wellmann and Strother, 2015; Dahl and Arens, 2020).

As oxygen started to rise to modern levels roughly 500 MY ago, nitrogenase inhibition by O_2 no longer limited atmospheric O_2 because land plants (which generate about half of all O_2 today) perform photosynthesis in aerial organs that are physically removed from their source of nitrogen: N_2 fixing microbes in soil. Land plants amplified one simple enzyme activity that is also present in prokaryotes (Römling and Galperin, 2015), cellulose synthase, and optimized it in terms of fiber production in order to generate the main component of land plant cell walls: cellulose. Cellulose is polymeric glucose, $C_6H_{12}O_6$, it is also produced in the cell walls of roots, but it is mainly produced from UDP-glucose in the cell walls of stems and leaves, the tissues where land plant photosynthesis takes place.

Cellulose synthesis altered the global O_2 budget by allowing plants to synthesize massive amounts of nitrogen-free carbon polymers with the help of catalytic amounts of nitrogen in enzymes. The OEC in land plant leaves generates O_2 in the process of supplying the electrons from H_2O for the photosynthetic electron transport chain and CO_2 fixation in the Calvin cycle, which generates phosphorylated sugars. They then form activated glucose monomers that are polymerized upon exiting the cell as cellulose (and other polymers, but mainly cellulose). Land plants cannot turn photosystems off, and they have access to more light than their water-column-dwelling algal ancestors did. While algae can redirect electrons from water splitting to hydrogenases as a safety valve for excess electron flow, land plants lost that ability because they converted their hydrogenases into oxygen sensors that operate much in the same way as aconitase or FNR do, via O_2-dependent FeS cluster oxidation (Gould et al., 2019). As a consequence, land plants have no safety valves for excess electrons, they have to direct electrons to CO_2 and deposit the product outside the cell, typically as cellulose. Roughly 80 % of Earth's current biomass carbon is present in land plants (Bar-On and Milo, 2019), the majority being present in cellulose. At the origin of land plants, the synthesis of cellulose led to an increased rate of carbon burial that was independent of nitrogen availability. In addition,

1146 it took place in aerial organs, separated from N_2 fixation by soil bacteria, which have been
1147 active in the rhizosphere since the origin of land plants (Puginier et al., 2022). The
1148 accumulation of buried carbon in the cell wall of the first land plants is thought to be the factor
1149 that drove the late rise in atmospheric O_2 (Mills et al., 2022; Dahl and Arens, 2020; Krause et
1150 al., 2018).

1151

1152 Cellulase activity on land is the equivalent to marine primary 1153 production

1154 Cellulose can be synthesized in days but in old trees and buildings it can remain stable
1155 over thousands of years because it presents an insoluble, solid phase substrate that is extremely
1156 difficult to metabolize. Cellulose was a keystone component of the terrestrialization process
1157 that witnessed the origin of land plants over 450 MY ago (de Vries et al., 2016). Oxygen started
1158 its Cambrian accumulation at a time in which fossil spores indicate the presence of land plants,
1159 but the fossilized plants themselves are scarce. In the ocean, primary production does not
1160 generate fibers; rather, it mainly generates branched polymers (e.g., alginates) that form
1161 gelatinous sheaths. The first land plants were streptophyte algae like *Chara* (Nishiyama et al.,
1162 2018) that have cellulose cell walls. A consequence of cellulose deposition in terrestrial
1163 habitats is that cellulose became, by weight, the basis of the terrestrial food chain. As a
1164 consequence, enzymes that degrade cellulose—cellulases and cellulosomes (Bayer et al., 1998;
1165 Bayer et al., 2013; Artzi et al., 2017)—became the main source of carbon substrates on land.
1166 Fiber degradation entails two kinds of processes that are performed exclusively by two groups
1167 of microbes. The aerobic process is performed mainly by fungi, in which the end product of
1168 degradation is CO_2 (and fungal cell mass). The anaerobic process is carried out mainly by
1169 bacteria and involves the breakdown of the insoluble (solid-phase) cellulose substrate to

soluble sugars (mono- and oligo-saccharides) that are fermented to intermediates and fermentation end-products. These compounds serve as the food source for intestinal uptake in animals (Mizrahi, 2013; Morais and Mizrahi, 2019) and, collectively, as the basis for microbial communities in the gut, in sediment and in soil, where fiber-degrading bacteria play an essential role. The solubilization of plant fiber on land requires cellulases that can degrade a solid phase substrate, their activity returns stably sequestered carbon (cellulose on land) to the ecosystem in the form of biochemically accessible sugars. Cellulase is thus functionally equivalent to primary production in marine environments, because it returns metabolically inaccessible carbon to the food chain.

Conclusion

Oxygen arose in a world of anaerobes that learned to avoid it, learned to live with it and/or learned to use it. Oxygen-dependent enzymes were subject to LGT more often than O_2-independent enzymes. The O_2 produced by cyanobacteria is clear evidence that these bacteria were constantly present in massive amounts in the 2.4 Ga since the GOE. There is no clear evidence that any other group of microbes was present in similar abundance (Hurley et al., 2021), hence if any group was capable of limiting O_2 synthesis in terms of (bio-)mass action, it was cyanobacteria. Consistent with that interpretation, cyanobacteria are the only microbes that can both synthesize O_2 in significant amounts and concomitantly limit O_2 accumulation—through nitrogenase inhibition by O_2. The late rise in O_2 following the 'boring billion' (the Pasteurian epoch) corresponds to terrestrial cellulose deposition made possible by the physical separation of N_2 fixation (in soil) from CO_2 fixation in aerial organs of the first land plants. The activities of three enzymes were likely crucial for O_2 accumulation during Earth history: the oxygen evolving complex itself, nitrogenase and cellulose synthase. On land, cellulases

1194 mobilize carbon from insoluble cellulose fibers, giving the cellulases of cellulolytic bacteria
1195 and fungi an ecological role similar to CO_2 fixation pathways of marine primary producers.

1196 The main functions of prokaryotic enzymes that use O_2 in central metabolism suggest
1197 that its main benefit was ecological rather than energetic. In secondary metabolism, O_2-
1198 dependent reactions fostered the synthesis of new antimicrobial compounds and pathways of
1199 their degradation (Jabłońska and Tawfik, 2019; Hoffarth et al., 2020). In primary metabolism,
1200 O_2-dependent enzymes rarely, if ever, generated fundamentally new biosynthetic or
1201 assimilatory routes. Instead, they presented alternative, O_2-tolerant routes to preexisting O_2-
1202 independent pathways that involved O_2-sensitive intermediates or O_2-sensitive cofactors, in
1203 particular FeS clusters. This allowed cells to colonize O_2-containing habitats from which they
1204 otherwise were excluded. Depending upon the cell's auxotrophies, several O_2-sensitive
1205 reactions occurring within the same cell could have required O_2-tolerant alternatives
1206 simultaneously for colonization of oxic habitats, and this requirement might have facilitated
1207 the higher fixation rates of genes transferred by LGT that are observed for O_2-dependent
1208 reactions. The colonization of oxic habitats not only required the presence of O_2 detoxification
1209 mechanisms, it required the presence of O_2-tolerant pathways for essential cofactor
1210 biosynthesis and substrate mobilization. Many O_2-dependent enzymes provide alternatives to
1211 the older O_2-sensitive counterparts in cofactor biosynthesis and in the cleavage of stable bonds
1212 that mobilize substrates. Both had to evolve and be in place before O_2 respiration became an
1213 option.

1214 As outlined in Fig 10, the main initial benefit of O_2 was likely to foster the origin of
1215 enzymes that could overcome O_2-toxicity and O_2-inhibition of (often FeS-dependent) reactions
1216 in preexisting anaerobic biosynthetic pathways for essential cofactors such as heme,
1217 cobalamin, chlorophyll (Dailey et al., 2017; Bryant et al., 2020), or the mobilization of organic
1218 substrates. The absence of proton-pumping capabilities in many terminal oxidases (Murali et

al., 2021) and the somewhat surprising absence of O_2-dependent substrate level phosphorylations, despite the highly exergonic nature of reactions between O_2 and organic compounds, suggest that the microbes that learned to harness O_2 in biochemical evolution were not limited in energy efficiency but were instead limited in their ability to grow in oxic habitats because of enzymatic inhibition imposed by O_2. This selective pressure for the origin of O_2-dependent enzymes, suggested for chlorophyll biosynthesis (Gomez Maqueo Chew and Bryant, 2007), appears to apply more generally. Stated another way, what good is improved energetic yield from O_2 if a cell, for lack of essential cofactors or nutrients, cannot grow even in the presence of the terminal acceptor? The evolution and spread of O_2-dependent enzymes offered access to oxygenated environments. Two billion years of low atmospheric O_2 during the Pasteurian (the boring billion) was also the result of O_2-inhibition of FeS centers—in nitrogenase, an essential O_2-sensitive enzyme to which no O_2-dependent alternative has evolved.

Materials and methods

Collection of oxygen-dependent and -independent reactions

Data for 11,804 metabolic reactions were downloaded from the Kyoto Encyclopedia of Genes and Genomes (Kanehisa & Goto, 2000) reaction database (version 10th August, 2022). Additionally, we manually added the reaction linked to superoxide-dismutase (SOD; R00275). In KEGG, SOD was linked to an enzyme commission number, which was in turn linked to the KEGG orthology identifier (KO) for SOD. However, there was no direct link between the SOD reaction and the KO. As there was no discernable reason for this, we added the enzyme SOD and its reaction for the qualitative descriptions of this paper.

The data of the 11,805 was subsequently filtered for reactions involving O_2 (KEGG compound C00007), yielding a set of 1,949 O_2-dependent reactions, most of these being specific to eukaryotes. The reactions were mapped to protein families. These protein families were created using MCL (Enright et al., 2002) as described elsewhere (Nagies et al., 2020). For protein family annotation, all clustered sequences were blasted against the KEGG database using Diamond 2.0.1 (Buchfink et al., 2015; Kanehisa et al., 2017). All best hits with at least 25 % identity and a maximum e-value of $1 \cdot 10^{-10}$ were used for annotation. Based on these hits, one KO (KEGG orthology identifier) annotation was assigned to each cluster based on majority rule (S4 Table). Protein families which contained at least 75 % of unknown sequences without any hits in the KEGG database were not annotated. Reactions were assigned to KOs based on the information provided by the KEGG database. Only reactions that we could link to prokaryotic protein families were retained, yielding 365 O_2-dependent reactions occurring in prokaryotes, referred to as the O_2-dependent reaction set. Of the remaining 9,597 reactions, a set of 3,018 prokaryotic reactions linked to prokaryotic protein families (the O_2-independent reaction set; S8 Table). Verticality values of all protein families of O_2-dependent and O_2-independent reactions and the average verticality of O_2-dependent and O_2-independent reactions were plotted against the number of genomes in a protein family (S1A Fig) and against the number of genomes linked to a reaction (S1B Fig).

Verticality distribution of O_2-dependent and O_2-independent reactions

Each annotated protein family was grouped into O_2-dependent and O_2-independent reactions based on the reactions linked to the annotated KOs. Protein families without a linked reaction were excluded. For the remaining protein families' verticality values (V) from Nagies

et al. (2020) were assigned (S8 Table). The distribution of verticality was generated for both reaction sets. There were more values present in the O_2-independent reaction set (3,018 versus 365 reactions and 8,322 versus 547 protein families with corresponding verticality). T-tests (Welch, 1947) were performed to test whether reaction sets were different in terms of V and corroborated subsequent results by subsampling V from O_2-independent protein families (S2C Table and S2D Table). For this sub-sampling procedure, 100,000 samples of 547 O_2-independent protein families were generated to compare to the 547 protein families of O_2-dependent reactions. The average verticality (V_{avg}) was calculated for each sample. With this, an estimate was possible of how likely it would be to create a sample with a mean verticality as low as is present in the O_2-dependent distribution from values in the O_2-independent distribution. This sub-sampling procedure was repeated, but changed each time to mitigate potential effects of unequal cluster size distributions. For this, four bins of equal size for cluster sizes (number of genomes in a cluster) of the O_2-dependent reaction were defined in the range of 4 genomes (minimum cluster size to calculate phylogenetic trees) to 3,380 genomes. For each bin the number of clusters in the O_2-dependent distribution was counted, and during sampling, bins were filled up with an equal number of clusters from the O_2-independent distribution (S2D Table).

Calculation of Gibbs energy

The change in Gibbs energy $\Delta G'$ was calculated for each reaction of both O_2-dependent and O_2-independent sets using eQuilibrator API (Beber et al., 2021; Flamholz et al., 2012) version 0.4.1 under Python v. 3.6.7. The program bases thermodynamic estimates on the component contribution method (Noor et al., 2013). Calculations were performed for physiological conditions (pH 7, 1 mM concentrations of reactants and products, 25 °C, ionic

strength 250 mM). In each reaction, O_2 was always written as a reactant (C00007 on the left side of the reaction), such that reaction R00009, R02550, R05229 and R00275 were reversed prior to calculation. In total, eQuilibrator yielded $\Delta G'$ for 288 O_2-dependent reactions and 2,139 for O_2-independent reactions (S9 Table). The calculated $\Delta G'$ of the O_2-dependent and O_2-independent sets were compared with a t-test (Welch, 1947) and all values were tested for correlation (Pearson's R^2) with the average verticality of a reaction (arithmetic mean of all V linked to a reaction). Finally, we counted the number of O_2-dependent and O_2-independent reactions in the largest genome of each species in our dataset and applied regression models (see S3 Fig and S2E Table).

Identification of cofactors

Cofactors for each reaction were identified by integrating data from the IUBMB Comments section of the BRENDA database (Chang et al., 2021), the E.C. subclass descriptions if applicable, and literature data associated with each KO entry. In some cases, original literature not listed in KEGG was consulted, the references for which were included in the cofactor table (S5 Table). The BRENDA database was queried via E.C. number, while the KEGG literature data for each enzyme was accessed via the KOs associated with the corresponding reaction. In case of discrepancies, literature data was prioritized.

Only cofactors bound by the protein subunits corresponding to the KO annotation were listed. When there were multiple possibilities for the cofactors of an enzyme/subunit/chain, all were listed. Cytochromes and ferredoxins were listed as cofactors only if they were bound by the enzyme as soluble electron carriers. For cytochrome or ferredoxin enzyme domains the cofactors listed were heme and iron-sulphur clusters, respectively. The reactions sometimes explicitly include an electron donor that provides electrons to the main enzyme indirectly, e.g.,

through an additional reductase component. Such electron donors were not listed, since they are not immediate ligands of the enzyme in question. In some cases, the natural electron donor was unknown, and was therefore not listed. In addition, in cases where the cofactors were substrates or products in the reaction, they were not listed. Co-substrates, such as 2-oxoglutarate in 2-oxoglutarate non-heme iron dependent oxygenases, were also not listed.

Estimation of LGT rates

For the construction of a phylum specific backbone tree, a concatenated alignment of 32 universal and vertical genes was made with MAFFT L-INS-I (version 7.471, Katoh and Standley, 2013). The alignment consists 32 conserved genes present in all 300 genomes belonging to all phyla except of “other_Bacteria” and “other_Archaea”. To reduce any phylogenetic bias, only one genome per species was used and at most 10 genomes per phylum. The phylogenetic tree was calculated with RAxML (version 8.2.8, Stamatakis 2014) using the model ProtCatLG. Monophyly was forced for all 40 phyla. The tree was rooted using the minimal ancestor deviation method (version 2.2, Tria et al., 2017). For further analysis, the monophyletic groups for phyla were collapsed into one branch to which the average branch length of the group was assigned with the help of ETE3 (Huerta-Cepas et al., 2016).

A presence-absence matrix for the distribution of O_2-dependent reactions across 40 bacterial and archaeal phyla was constructed. The reconstruction of origin nodes was done using a prokaryotic backbone phylogeny. An origin node was defined as the last common ancestor of all phyla possessing an O_2-dependent reaction.

Occurrence of functional categories per prokaryotic phylum

The analysis was based on a prokaryotic MCL-clustering (Enright et al., 2002) with an identity-threshold of ≥ 25 % and an e-value cut-off of 10^{-10}. The calculation of protein families included 5,655 genomes (212 archaea and 5,443 bacteria). Oxygen-dependent reactions were assigned to 792 protein families (S4 Table) and each protein family was linked to a KO annotation of the KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa & Goto, 2000; Kanehisa et al., 2017) orthology database. Using the KO identifiers, each protein family was assigned to at least one functional category (B-level) according to the KEGG BRITE classification. The distribution and occurrence of functional categories was examined for each phylum within all analyzed protein families linked to oxygen-dependent reactions. The resulting numbers were sorted by frequency and plotted in a 3D bar chart.

Acknowledgements

This paper is dedicated to the memory of Dan Tawfik.

Author contributions

Conceptualization: WFM, FSPN, JLEW; Methodology: WFM, FSPN, JLEW, NM, MRK, NK; Data Curation: FSPN, JLEW, NM, MRK; Formal Analysis: FSPN, JLEW, NM, MRK, NK, KT, NB, LM; Writing—Original Draft: WFM, FSPN, JLEW, MRK, NM; Writing—Review & Editing: WFM, JFA, FSPN, JLEW, MDE, IM, MRK, NK, KT, NB, NM, LM, REG; Visualization: WFM, FSPN, JLEW, NM, MRK, NB, LM; Supervision: WFM.

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101018894 to WFM). WFM and IM thank the German Israeli project cooperation DIP for funding (1426/23-1 to WFM and 2476/2-1 to IM). This work was also supported by the Deutsche Forschungsgemeinschaft (MA 1426/21-1 to WFM) and Volkswagen Foundation (96742 to WFM).

References

J. Priestley, XIX. Observations on different kinds of air, Philos. Trans. R. Soc. Lond. 62 (1772) 147–264.

C.W. Scheele, Chemische Abhandlung von der Luft und dem Feuer. Verlegt von Magn. Swederus, Buchhändler zu finden bey SL Crusius. 1777.

A.L. Lavoisier, Mémoire sur la formation de l'acide nommé air fixe ou acide crayeux, Mém. Acad. R. Sci. Paris. 2 (1781) 448–458.

L. Pasteur, Note de L. Pasteur - Animalcules infusoires vivant sans gaz oxygène libre et déterminant des fermentations, C. R. T. 52 (1861) 344–347.

C. Mereschkowsky Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis, einer neuen Lehre von der Entstehung der Organismen, Biologisches Centralblatt. 30 (1910) 353–442.

M. Preiner, K. Igarashi, K.B. Muchowska, M. Yu, S.J. Varma, K. Kleinermanns, M.K. Nobu, Y. Kamagata, H. Tüysüz, J. Moran, W.F. Martin, A hydrogen-dependent geochemical analogue of primordial carbon and energy metabolism, Nat. Ecol. Evol. 4 (2020) 534–542.

1382 H.D. Holland, Volcanic gases, black smokers, and the great oxidation event, Geochim.
1383 Cosmochim. Acta. 66 (2002) 3811–3826.
1384 L.R. Kump, The rise of atmospheric oxygenic, Nature. 451 (2008) 277–278.
1385 I. Ślesak, M. Kula, H. Ślesak, Z. Miszalski, K. Strzałka, How to define obligatory anaerobiosis?
1386 An evolutionary view on the antioxidant response system and the early stages of the
1387 evolution of life on Earth, Free Radical Biol. Med. 140 (2019) 61–73.
1388 T.W. Lyons, C. T. Reinhard, N. J. Planavsky, The rise of oxygen in Earth's early ocean and
1389 atmosphere, Nature. 50. (2014) 307–315.
1390 J.F. Allen, B. Thake, W.F. Martin, Nitrogenase inhibition limited oxygenation of Earth's
1391 proterozoic atmosphere, Trends Plant Sci. 24 (2019) 1022–1031.
1392 T.M. Lenton, T.W. Dahl, S.J. Daines, B.J.W. Mills, K. Ozaki, M.R. Saltzman, P. Porada,
1393 Earliest land plants created modern levels of atmospheric oxygen, PNAS. 113 (2016)
1394 9704–9709.
1395 D.A. Stolper, B.C. Keller, A record of deep-ocean dissolved O_2 from the oxidation state of iron
1396 in submarine basalts, Nature. 553 (2018) 323–327.
1397 D.B. Mills, R.A. Boyle, S.J. Daines, E.A. Sperling, D. Pisani, P.C.J. Donoghue, T.M. Lenton,
1398 Eukaryogenesis and oxygen in Earth history, Nat. Ecol. and Evol. 6 (2022) 520–532.
1399 R. Buick, D.J. Des Marais, A.H. Knoll, Stable isotopic compositions of carbonates from the
1400 Mesoproterozoic Bangemall Group, northwestern Australia, Chem. Geol. 123 (1995)
1401 153–171.
1402 I. Mukherjee, R.R. Large, R. Corkrey, L.V. Danyushevsky, The Boring Billion, a slingshot for
1403 complex life on Earth, Sci Rep. 8 (2018) 4432.
1404 W.F. Martin, A.G.M. Tielens, M. Mentel, Mitochondria and anaerobic energy metabolism in
1405 Eukaryotes: Biochemistry and Evolution, De Gruyter, Berlin, Boston, 2020: pp. 1–270.

1406 P.E. Cloud, Atmospheric and hydrospheric evolution on the primitive Earth, Science. 160
1407 (1968) 729–736.
1408 N.J. Planavsky, C.T. Reinhard, X. Wang, D. Thomson, P. McGoldrick, R.H. Rainbird, T.
1409 Johnson, W.W. Fischer, T.W. Lyons, Low mid-proterozoic atmospheric oxygen levels
1410 and the delayed rise of animals, Science. 346 (2014) 635–638.
1411 C.T. Reinhard, N.J. Planavsky, S.L. Olson, T.W. Lyons, D.H. Erwin, Earth's oxygen cycle and
1412 the evolution of animal life, PNAS. 113 (2016) 8933–8938.
1413 L. Margulis, Origin of eukaryotic cells: Evidence and research implications for a theory of the
1414 origin and evolution of microbial, plant and animal cells on the precambrian Earth, Yale
1415 University Press, New Haven, CT, 1970.
1416 J. Raymond, D. Segrè, The effect of oxygen on biochemical networks and the evolution of
1417 complex life, Science. 311 (2006) 1764–1767.
1418 G.E. Budd, The earliest fossil record of the animals and its significance, Philos. Trans. R. Soc.
1419 Lond. B Biol. Sci. 363 (2008) 1425–1434.
1420 V. Zimorski, M. Mentel, A.G.M Tielens, W.F. Martin, Energy metabolism in anaerobic
1421 eukaryotes and Earth's late oxygenation, Free Radical Biol. Med. 140 (2019) 279–294.
1422 M. Müller, M. Mentel, J.J. van Hellemond, K. Henze, C. Woehle, S.B. Gould, R.-Y. Yu, M.
1423 van der Giezen, A.G.M. Tielens, W.F. Martin, Biochemistry and evolution of anaerobic
1424 energy metabolism in eukaryotes, Microbiol. Mol. Biol. Rev. 76 (2012) 444–495.
1425 K.M. Towe, Oxygen collagen priority and early metazoan fossil record, PNAS. 65 (1970) 781–
1426 788.
1427 M.D. Shoulders, R.T.Raines, Collagen structure and stability, Annu. Rev. Biochem. 78 (2009)
1428 929–958.
1429 J.L. Payne, C.R. McClain, A.G. Boyer, J.H. Brown, S. Finnegan, M. Kowaleski, R.A. Krause
1430 Jr, S.K. Lyons, D.W. McShea, P.M. Novack-Gottshall, F.A. Smith, P. Spaeth, J.A.

1431 Stempien, S.C. Wang, The evolutionary consequences of oxygenic photosynthesis: A
1432 body size perspective, Photosynth. Res. 107 (2011) 37–57.
1433 J.F. Harrison, A. Kaiser, J.M. Van den Brooks, Atmospheric oxygen level and the evolution of
1434 insect body size., Proc. R. Soc. B. 277 (2010) 1937–1946.
1435 S.B. Gould, S.G. Garg, M. Handrich, S. Nelson-Sathi, N. Gruenheit, A.G.M. Tielens, W.F.
1436 Martin, Adaptation to life on land at high O_2 via transition from ferredoxin-to NADH-
1437 dependent redox balance, Proc. Biol. Sci. 286 (2019) 20191491.
1438 J.J. Brocks, A.J.M. Jarrett, E. Sirantoine, C. Hallmann, Y. Hoshino, T. Liyanage, The rise of
1439 algae in Cryogenian oceans and the emergence of animals, Nature. 548 (2017) 578–581.
1440 N.T. Arndt, E.G. Nisbet, Processes on the young Earth and the habitats of early life. Annu.
1441 Rev. Earth Planet. Sci. 40 (2012) 521–549.
1442 S.W. Poulton, P.W. Fralick, D.E. Canfield, The transition to a sulphidic ocean similar to ~1.84
1443 billion years ago, Nature. 431 (2004a) 173–177.
1444 L.J. Alcott, B.J. Mills, S.W. Poulton, Stepwise earth oxygenation is an inherent property of
1445 global biogeochemical cycling, Science. 366 (2019) 1333–1337.
1446 J.M. Klatt, A. Chennu, B.K. Arbic, B.A. Biddanda, G.J. Dick, Possible link between Earth's
1447 rotation rate and oxygenation, Nat. Geosci. 7 (2021) 1–7.
1448 T. Tashiro, A. Ishida, M. Hori, M. Igisu, M. Koike, P. Méjean, N. Takahata, Y. Sano, T.
1449 Komiya, Early trace of life from 3.95 Ga sedimentary rocks in Labrador, Canada, Nature.
1450 549 (2017) 516–518.
1451 Y. Ueno, K. Yamada, N. Yoshida, S. Maruyama, Y. Isozaki, Evidence from fluid inclusions
1452 for microbial methanogenesis in the early Archaean era, Nature. 440 (2006) 516–519.
1453 E.G. Nisbet, N.H. Sleep, The habitat and nature of early Life. Nature. 409 (2001) 1083–91.
1454 F. Westall, F. Foucher, B. Cavalazzi, S.T. de Vries, W. Nijman, V. Pearson, J. Watson, A
1455 Verchovsky, I. Wright, J.-N. Rouzaud, D. Machesini, S. Anne, Volcaniclastic habitats

1456 for early life on Earth and Mars: A case study from ~3.5 Ga-old rocks from the Pilbara,
1457 Australia, Planet. Space Sci. 59 (2011) 1093–1106.
1458 J.F. Allen, A proposal for formation of Archaean stromatolites before the advent of oxygenic
1459 photosynthesis, Front. Microbiol. 7 (2016) 1784.
1460 J. Castresana, M. Lübben, M. Saraste, D.G. Higgins, Evolution of cytochrome oxidase, an
1461 enzyme older than atmospheric oxygen, EMBO J. 13 (1994) 2516–2525.
1462 J. Castresana, D. Moreira, Respiratory chain in the last common ancestor of living organisms,
1463 J. Mol. Evol. 49 (1999) 453–460.
1464 C. Brochier-Armanet, E. Talla, S. Gribaldo, The multiple evolutionary history of dioxygen
1465 reductases: Implications for the origin and evolution of aerobic respiration, Mol. Biol.
1466 Evol. 26 (2009) 285–297.
1467 R. Murali, J. Hemp, R.B. Gennis, Evolution of quinol oxidation within the heme-copper
1468 oxidoreductase superfamily, BBA – Bioenergetics. 1863 (2022) 148907.
1469 T. Dagan, Y. Artzy-Randrup, W. Martin, Modular networks and cumulative impact of lateral
1470 transfer in prokaryote genome evolution, PNAS. 105 (2008) 10039–10044.
1471 F.S.P. Nagies, J. Brueckner, F.D.K. Tria, W.F. Martin, A spectrum of verticality across genes,
1472 PLOS Genet. 16 (2020) e1009200.
1473 F. Passardi, N. Bakalovic, F.K. Teixeira, M. Margis-Pinheiro, C. Penel, C. Dunand,
1474 Prokaryotic origins of the non-animal peroxidase superfamily and organelle-mediated
1475 transmission to eukaryotes, Genomics. 89 (2007) 567–579.
1476 F.L. Sousa, S. Nelson-Sathi, W.F. Martin, One step beyond a ribosome: The ancient anaerobic
1477 core, BBA – Bioenergetics. 1857 (2016) 1027–1038.
1478 Z. Lu, J.A. Imlay, When anaerobes encounter oxygen: mechanisms of oxygen toxicity,
1479 tolerance and defense, Nat. Rev. Microbiol. 19 (2021) 774–785.

1480 G. Unden, J. Bongaerts, Alternative respiratory pathways of Escherichia coli: Energetics and
1481 transcriptional regulation in response to electron acceptors, BBA – Bioenergetics. 1320
1482 (1997) 217–234.
1483 A.N. Brown, M.T. Anderson, M.A. Bachman, H.L.T. Mobley, The ArcAB two-component
1484 system: Function in metabolism, redox control, and infection, Microbiol. Mol. Biol. Rev.
1485 86 (2022) e00110-21.
1486 D.R. Nelson, Cytochrome P450 diversity in the tree of life, BBA – Proteins and Proteomics.
1487 1866 (2018) 141–154.
1488 P. M. Wood, The potential diagram for oxygen at pH 7, Biochem J. 253 (1988) 287–289.
1489 K. Schmidt-Rohr, Why combustions are always exothermic, yielding about 418 kJ per mole of
1490 O_2, J. Chem. Educ. 92 (2015) 2094–2099.
1491 W.T. Borden, R. Hoffmann, T. Stuyver, B. Chen, Dioxygen: What makes this triplet diradical
1492 kinetically persistent?, J. Am. Chem. Soc. 139 (2017) 9010–9018.
1493 L. Brewer, The thermodynamic properties of the oxides and their vaporization processes,
1494 Chem. Rev. 52 (1952) 1–75.
1495 F.H. Vaillancourt, J.T. Bolin, L.D. Eltis, The ins and outs of ring-cleaving dioxygenases, Crit.
1496 Rev. Biochem. Mol. Biol. 41 (2006) 241–267.
1497 S.G. Huwiler, C. Löffler, S.E.L. Anselmann, H.J. Stärk, M. von Bergen, J. Flechsler, R.
1498 Rachel, M. Boll, One-megadalton metalloenzyme complex in Geobacter metallireducens
1499 involved in benzene ring reduction beyond the biological redox window, PNAS. 116
1500 (2019) 2259–2264.
1501 H. Gaweska, P.F. Fitzpatrick, Structures and mechanism of the monoamine oxidase family,
1502 Biomol. Concepts. 2 (2011) 365–377.
1503 G. Fuchs, M. Boll, J. Heider, Microbial degradation of aromatic compounds – from one
1504 strategy to four, Nat. Rev. Microbiol. 9 (2011) 803–816.

1505 M.A. Vanoni, Iron-sulfur flavoenzymes: the added value of making the most ancient redox
1506 cofactors and the versatile flavins work together, Open Biol. 11 (2021) 210010.
1507 G. Fuchs, Anaerobic metabolism of aromatic compounds, Ann. N. Y. Acad. Sci. 1125 (2008)
1508 82–99.
1509 A. Gomez Maqueo Chew, D.A. Bryant, Chlorophyll biosynthesis in bacteria: The origins of
1510 structural and functional diversity, Annu. Rev. Microbiol. 61 (2007) 113–129.
1511 W. Buckel, B.T. Golding, Radical enzymes in anaerobes, Annu. Rev. Microbiol. 60 (2006) 27–
1512 49.
1513 W. Buckel, B.T. Golding. 2012 Radical enzymes. In: Chatgilialoglu C, Studer A (eds.).
1514 Encyclopedia of Radicals in Chemistry, Biology and Materials. Chicester, West Sussex,
1515 Hoboken, NJ: John Wiley & Sons, Ltd., 2012, 1501–1546.
1516 B. Meunier, S. P. de Visser, S. Shaik. Mechanism of oxidation reactions catalyzed by
1517 cytochrome P450 enzymes. Chem. Rev. 2004, 104, 3947-3980
1518 S. Nelson-Sathi, T. Dagan, G. Landan, A. Janssen, M. Steel, J.O. McInerney, U. Deppenmeier,
1519 W.F. Martin, Acquisition of 1,000 eubacterial genes physiologically transformed a
1520 methanogen at the origin of Haloarchaea, PNAS. 109 (2012) 20537–20542.
1521 H.S. Tehrani, A.A. Moosavi-Movahedi, Catalase and its mysteries, Prog. Biophys. Mol. Biol.
1522 140 (2018) 5–12.
1523 J.W. Whittaker, Non-heme manganese catalase – the 'other' catalase, Arch. Biochem. Biophys.
1524 525 (2012) 111–120.
1525 L.E. Khmelevtsova, I.S. Sazykin, T.N. Azhogina, M.A. Sazykina, Prokaryotic peroxidases and
1526 their application in biotechnology (Review), Appl. Biochem. Microbiol. 56 (2020) 373–
1527 380.
1528 A. Bafana, S. Dutt, A. Kumar, S. Kumar, P.S. Ahuja, The basic and applied aspects of
1529 superoxide dismutase, J. Mol. Catal. B: Enzym. 68 (2011) 129–138.

1530 O.M. Ighodaro, O.A. Akinloye, First line defence antioxidants-superoxide dismutase (SOD),
1531 catalase (CAT) and glutathione peroxidase (GPX): Their fundamental role in the entire
1532 antioxidant defence grid, Alexandria J. Med. 54 (2018) 287–293.
1533 O.J. Njuma, E.N. Ndontsa, D.C. Goodwin, Catalase in peroxidase clothing: Interdependent
1534 cooperation of two cofactors in the catalytic versatility of KatG, Arch. Biochem.
1535 Biophys. 544 (2014) 27–39.
1536 J.S. Boden, K.O. Konhauser, L.J. Robbins, P. Sánchez-Baracaldo, Timing the evolution of
1537 antioxidant enzymes in cyanobacteria, Nat. Commun. 12 (2021) 4742.
1538 V.B. Borisov, R. Murali, M.L. Verkhovskaya, D.A. Bloch, H. Han, R.B. Gennis, M.I.
1539 Verkhovsky, Aerobic respiratory chain of Escherichia coli is not allowed to work in fully
1540 uncoupled mode, PNAS. 108 (2011) 17320–17324.
1541 Q.H. Tran, G. Unden, Changes in the proton potential and the cellular energetics of Escherichia
1542 coli during growth by aerobic and anaerobic respiration or by fermentation, Eur. J.
1543 Biochem. 251 (1998) 538–543.
1544 G. Unden, S. Achebach, G. Holighaus, H.G. Tran, B. Wackwitz, Y. Zeuner, Control of FNR
1545 function of Escherichia coli by O2 and reducing conditions, J. Mol. Microbiol.
1546 Biotechnol. 4 (2002) 263–268.
1547 C.M. Czekster, J.S. Blanchard, One substrate, five products: reactions catalyzed by the
1548 dihydroneopterin aldolase from Mycobacterium tuberculosis, J. Am. Chem. Soc. 134
1549 (2012) 19758–19771.
1550 F.L. Sousa, R.J. Alves, J.B. Pereira-Leal, M. Teixeira, M.M. Pereira, A bioinformatics
1551 classifier and database for heme-copper oxygen reductases, PLoS One. 6 (2011) e19117.
1552 J.-H. Martens, H. Barg, M.J. Warren, D. Jahn, Microbial production of vitamin B12, Appl.
1553 Microbiol Biotechnol. 58 (2002) 275–285.

1554 T. Mukherjee, J. Hanes, I. Tews, S.E. Ealick, T.P. Begley, Pyridoxal phosphate: Biosynthesis
1555 and catabolism, BBA – Proteins and Proteomics. 1814 (2011) 1585–1596.
1556 F.L. Sousa, L. Shavit-Grievink, J.F. Allen, W.F. Martin, Chlorophyll biosynthesis gene
1557 evolution indicates photosystem gene duplication, not photosystem merger, at the origin
1558 of oxygenic photosynthesis, Genome Biol. Evol. 5 (2013) 200–216.
1559 H.A. Dailey, T.A. Dailey, S. Gerdes, D. Jahn, M. Jahn, M.R. O'Brian, M.J. Warren,
1560 Prokaryotic heme biosynthesis: Multiple pathways to a common essential product,
1561 Microbiol. Mol. Biol. Rev. 81 (2017) 48–16.
1562 D.A. Bryant, C.N. Hunter, M.J. Warren, Biosynthesis of the modified tetrapyrroles—the
1563 pigments of life, J. Biol. Chem. 295 (2020) 6888–6925.
1564 S. Ollagnier-de Choudensa, L. Loiseau, Y. Sanakis, F. Barras, M. Fontecave, Quinolinate
1565 synthetase, an iron–sulfur enzyme in NAD biosynthesis, FEBS Lett. 579 (2005) 3737–
1566 3743.
1567 K. Alexander, I.G. Young, Alternative hydroxylases for the aerobic and anaerobic biosynthesis
1568 of ubiquinone in Escherichia coli, Biochemistry. 17 (1978) 4750–4755.
1569 L. Pelosi, C.D. Vo, S.S. Abby, L. Loiseau, B. Rascalou, M. Hajj Chehade, B. Faivre, M.
1570 Goussé, C. Chenal, N. Touati, L. Binet, D. Cornu, C.D. Fyfe, M. Fontecave, F. Barras,
1571 M. Lombard, F. Pierrel. Ubiquinone biosynthesis over the entire O_2 range:
1572 Characterization of a conserved O_2-Independent pathway. mBio 10 (2019) e01319-19.
1573 R. Leonardi, S.A. Fairhurst, M. Kriek, D.J. Lowe, P.L. Roach, Thiamine biosynthesis in
1574 Escherichia coli: Isolation and initial characterisation of the ThiGH complex, FEBS Lett.
1575 539 (2003) 95–99.
1576 E.C. Settembre, P.C. Dorrestein, J.H. Park, A.H. Augustine, T.P. Begley, S.E. Ealick,
1577 Structural and mechanistic studies on ThiO, a glycine oxidase essential for thiamin
1578 biosynthesis in Bacillus subtilis, Biochemistry. 42 (2003) 2971–2981.

1579 J. M. McCord, I. Fridovich, Superoxide dismutase. An enzymatic function for erythrocuprein
1580 (hemocuprein), J. Biol. Chem. 244 (1969) 6049–6055.
1581 A.L. Brioukhanov, A.I. Netrusov, Aerotolerance of strictly anaerobic microorganisms and
1582 factors of defense against oxidative stress: a review, Appl. Biochem. Microbiol. 43
1583 (2007) 567–82.
1584 J. Jabłońska, D.S. Tawfik, The number and type of oxygen-utilizing enzymes indicates aerobic
1585 vs. anaerobic phenotype, Free Radical Biol. Med. 140 (2019) 84–92.
1586 X. Huang, J.T. Groves, Oxygen activation and radical transformations in heme proteins and
1587 metalloporphyrins, Chem. Rev. 118 (2018) 2491–2553.
1588 L. Zhang, S. Jiang, W. Ma, Z. Zhou, Oxygen reduction reaction on Pt-based electrocatalysts:
1589 Four-electron vs. two-electron pathway, Chin. J. Catal. 43 (2022) 1433–1443.
1590 B.A. Palfey, D.P. Ballou, V. Massey, Oxygen activation by flavins and pterins, Active oxygen
1591 in biochemistry (1995) 37–83.
1592 E. Romero, J.R. Gómez Castellanos, G. Gadda, M.W. Fraaije, A. Mattevi, Same substrate,
1593 many reactions: Oxygen activation in flavoenzymes, Chem Rev. 118 (2018) 1742–1769.
1594 J.-A. Losman, P. Koivunen, W.G. Kaelin Jr., 2-Oxoglutarate-dependent dioxygenases in
1595 cancer, Nat. Rev. Cancer. 20 (2020) 710–726.
1596 C.Q. Herr, R.P. Hausinger, Amazing diversity in biochemical roles of Fe(II)/2-oxoglutarate
1597 oxygenases, Trends Biochem Sci. 43 (2018) 517–532.
1598 J.P. Klinman, Life as aerobes: are there simple rules for activation of dioxygen by enzymes?,
1599 J. Biol. Inorg. Chem. 6 (2001) 1–13.
1600 T. Wongnate, P. Surawatanawong, S. Visitsatthawong, J. Sucharitakul, N.S. Scrutton, P.
1601 Chaiven, Proton-coupled electron transfer and adduct configuration are important for
1602 C4a-hydroperoxyflavin formation and stabilization in a flavoenzyme, J. Am. Chem. Soc.
1603 136 (2014) 241–253.

1604 Y. Wang, J. Li, A. Liu, Oxygen activation by mononuclear nonheme iron dioxygenases
1605 involved in the degradation of aromatics, J. Biol. Inorg. Chem. 22 (2017) 395–405.
1606 S.M. Barry, G.L. Challis, Mechanism and catalytic diversity of Rieske non-heme iron-
1607 dependent oxygenases, ACS catal. 3 (2013) 2362–2370.
1608 J.M. Bujnicki, Y. Oudjama, M. Roovers, S. Owczarek, J. Caillet, L. Droogmans, Identification
1609 of a bifunctional enzyme MnmC involved in the biosynthesis of a hypermodified uridine
1610 in the wobble position of tRNA, RNA. 10 (2004) 1236–1242.
1611 J. Kim, S.C. Almo, Structural basis for hypermodification of the wobble uridine in tRNA by
1612 bifunctional enzyme MnmC, BMC Struct. Biol. 13 (2013) 1–13.
1613 P.F. Widboom, E.N. Fielding, Y. Liu, S.D. Bruner, Structural basis for cofactor-independent
1614 dioxygenation in vancomycin biosynthesis, Nature. 447 (2007) 342–345.
1615 G. Sciara, S.G. Kendrew, A.E. Miele, N.G. Marsh, L. Federici, F. Malatesta, G. Schimperna,
1616 C. Savino, B. Vallone, The structure of ActVA-Orf6, a novel type of monooxygenase
1617 involved in actinorhodin biosynthesis, EMBO J. 22 (2003) 205–125.
1618 U. Frerichs-Deeken, K. Ranguelova, R. Kappl, J. Hüttermann, S. Fetzner, Dioxygenases
1619 without requirement for cofactors and their chemical model reaction: compulsory order
1620 ternary complex mechanism of 1 H-3-hydroxy-4-oxoquinaldine 2, 4-dioxygenase
1621 involving general base catalysis by histidine 251 and single-electron oxidation of the
1622 substrate dianion, Biochemistry. 43 (2004) 14485–14499.
1623 T. Grocholski, H. Koskiniemi, Y. Lindqvist, P. Mantsala, J. Niemi, G. Schneider, Crystal
1624 structure of the cofactor-independent monooxygenase SnoaB from Streptomyces
1625 nogalater: implications for the reaction mechanism, Biochemistry. 49 (2010) 934–944.
1626 B.J. Baas, H. Poddar, E.M. Geertsema, H.J. Rozeboom, M.P. de Vries, H.P. Permentier,
1627 A.M.W.H. Thunnissen, G.J. Poelarends, Functional and structural characterization of an
1628 unusual cofactor-independent oxygenase, Biochemistry. 54 (2015) 1219–1232.

1629 N. Colloc'h, M.E. Hajji, B. Bachet, G. l'Hermite, M. Schiltz, T. Prangé, C. Castro, J.P. Mornon,
1630 Crystal structure of the protein drug urate oxidase-inhibitor complex at 2.05 Å resolution,
1631 Nat. Struct. Biol. 4 (1997) 947–952.
1632 C. Bathellier, L.J. Yu, G.D. Farquhar, M.L. Coote, G.H. Lorimer, G. Tcherkez, Ribulose 1, 5-
1633 bisphosphate carboxylase/oxygenase activates O2 by electron transfer, PNAS. 117
1634 (2020) 24234–24242.
1635 F.R. Tabita, T.E. Hanson, S. Satagopan, B.H. Witte, N.E. Kreel, Phylogenetic and evolutionary
1636 relationships of RubisCO and the RubisCO-like proteins and the functional lessons
1637 provided by diverse molecular forms, Philos. Trans. R. Soc. Lond. B Biol. Sci. 363
1638 (2008) 2629–2640.
1639 G. Tcherkez, The mechanism of rubisco-catalysed oxygenation, Plant Cell Environ. 39 (2015)
1640 983–997.
1641 Y.M. Bar-On, R. Milo, The global mass and average rate of rubisco, PNAS. 116 (2019) 4738–
1642 4743.
1643 J.A. Imlay, Cellular defenses against superoxide and hydrogen peroxide, Annu. Rev. Biochem.
1644 77 (2008) 755–776.
1645 J.A. Imlay, Iron-sulphur clusters and the problem with oxygen, Mol. Microbiol. 59 (2006)
1646 1073–1082.
1647 L.C. Seaver, J.A. Imlay, Are respiratory enzymes the primary sources of intracellular hydrogen
1648 peroxide?, J. Biol. Chem. 279 (2004) 48742–48750.
1649 J.A. Imlay, The molecular mechanisms and physiological consequences of oxidative stress:
1650 lessons from a model bacterium, Nat. Rev. Microbiol. 11 (2013) 443–454.
1651 D. Croll, B.A. McDonald, The accessory genome as a cradle for adaptive evolution in
1652 pathogens, PLOS Pathogens. 8 (2012) e1002608.

1653 R.W. Jackson, B. Vinatzer, D.L. Arnold, S. Dorus, J. Murillo, The influence of the accessory
1654 genome on bacterial pathogen evolution. Mob. Genet. Elements. 1 (2011) 55–65.
1655 M. López-Pérez, F. Rodriguez-Valera, Pangenome evolution in the marine bacterium
1656 Alteromonas, Genome Biol. Evol. 8 (2016) 1556–1570.
1657 M. Degli Esposti, M. Mentel, W. Martin, F.L. Sousa, Oxygen reductases in
1658 alphaproteobacterial genomes: Physiological evolution from low to high oxygen
1659 environments, Front. Microbiol. 10 (2019) 499.
1660 H. Han, J. Hemp, L.A. Pace, H. Ouyang, K. Ganesan, J.H. Roh, F. Daldal, S.R. Blanke, R.B.
1661 Gennis, Adaptation of aerobic respiration to low O_2 environments, PNAS. 108 (2011)
1662 14109–14114.
1663 V.B. Borisov, R.B. Gennis, J. Hemp, M.I. Verkhovsky, The cytochrome bd respiratory oxygen
1664 reductases, BBA – Bioenergetics. 1807 (2011) 1398–1413.
1665 J.A. Leigh, S.V. Albers, H. Atomi, T. Allers, Model organisms for genetics in the domain
1666 Archaea: methanogens, halophiles, Thermococcales and Sulfolobales, FEMS Microbiol.
1667 Rev. 35 (2011) 577–608.
1668 R. Mei, M. Kaneko, H. Imachi, M.K. Nobu, The origin and evolution of methanogenesis and
1669 Archaea are intertwined, PNAS nexus. 2 (2023) pgad023.
1670 J. Schlesier, M. Rohde, S. Gerhardt, O. Einsle, A conformational switch triggers nitrogenase
1671 protection from oxygen damage by Shethna protein II (FeSII), J. Am. Chem. Soc. 138
1672 (2016) 239–247.
1673 K.T. Rytkönen, Evolution: Oxygen and early animals, eLife. 7 (2018) e34756.
1674 A.G. Tielens, C. Rotte, J.J. van Hellemond, W.F. Martin, Mitochondria as we don't know them,
1675 Trends Biochem. Sci. 27 (2002) 564–572.

1676 W.F. Martin, A.G.M. Tielens, M. Mentel, S.G. Garg, S.B. Gould, The Physiology of
1677 Phagocytosis in the Context of Mitochondrial Origin, Microbiol. Mol. Biol. Rev. 81
1678 (2017) e00008-17.
1679 M. Szenk, K.A. Dill, A.M.R. de Graff, Why do fast-growing bacteria enter overflow
1680 metabolism? Testing the membrane real estate hypothesis, Cell Syst. 5 (2017) 95–104.
1681 A.J. Wolfe, The acetate switch, Microbiol. Mol. Biol. Rev. 69 (2005) 12–50.
1682 M. Basan, S. Hui, H. Okano, Z. Zhang, Y. Shen, J.R. Williamson, T. Hwa, Overflow
1683 metabolism in Escherichia coli results from efficient proteome allocation, Nature. 528
1684 (2015) 99–104.
1685 T. Pfeiffer, A. Morley, An evolutionary perspective on the Crabtree effect, Front. Mol. Biosci.
1686 1 (2014) 00017.
1687 R. Murali, R.B. Gennis, J. Hemp, J. Evolution of the cytochrome bd oxygen reductase
1688 superfamily and the function of CydAA' in Archaea, ISME J. 15 (2021) 3534–3548.
1689 J.F. Allen, J.A. Raven, Free-radical-induced mutation vs redox regulation: costs and benefits
1690 of genes in organelles, J. Mol. Evol. 42 (1996) 482–492.
1691 J.F. Allen, A redox switch hypothesis for the origin of two light reactions in photosynthesis,
1692 FEBS Lett. 579 (2005) 963–968.
1693 H. Sies, Oxidative stress: a concept in redox biology and medicine, Redox Biol. 4 (2015) 180–
1694 183.
1695 H. Sies, C. Berndt, D.P. Jones, Oxidative stress, Annu. Rev. Biochem. 86 (2017) 715–748.
1696 R.K. Thauer, K. Jungermann, K. Decker, Energy conservation in chemotrophic anaerobic
1697 bacteria, Bacteriol. Rev. 41 (1977) 100–180.
1698 Y.A. Muller, G.E. Schulz, Structure of the thiamine- and flavin-dependent enzyme pyruvate
1699 oxidase, Science. 259 (1993) 965–967.

1700 K. Decker, K. Jungermann, R.K. Thauer, Energy production in Anaerobic Organisms, Angew.
1701 Chem. Int. Ed. Engl. 9 (1970) 138–158.
1702 K. Tittmann, G. Wille, R. Golbik, A. Weidner, S. Ghisla, G. Hübner, Radical phosphate
1703 transfer mechanism for the thiamin diphosphate- and FAD-dependent pyruvate oxidase
1704 from Lactobacillus plantarum. Kinetic coupling of intercofactor electron transfer with
1705 phosphate transfer to acetyl-thiamin diphosphate via a transient FAD
1706 semiquinone/hydroxyethyl-ThDP radical pair, Biochemistry. 44 (2005) 13291–13303.
1707 P.A. Frey, A.D. Hegeman, G.H. Reed, Free radical mechanisms in enzymology, Chem. Rev.
1708 106 (2006) 3302–3316.
1709 A. Abdel-Hamid, M.M. Attwood, J.R. Guest, Pyruvate oxidase contributes to the aerobic
1710 growth efficiency of Escherichia coli, Microbiology (Reading). 147 (2001) 1483–1498.
1711 L.P. Cornacchione, L.T. Hu, Hydrogen peroxide-producing pyruvate oxidase from
1712 Lactobacillus delbrueckii is catalytically activated by phosphotidylethanolamine, BMC
1713 Microbiol. 20 (2020) 128.
1714 R.J. Ellis, The most abundant protein in the world, Trends Biochem. Sci. 4 (1997) 241–244.
1715 B.K.B. Seah, C.P. Antony B. Huettel, J. Zarzycki, L.S. von Borzyskowski, T.J. Erb, A. Kouris,
1716 M. Kleiner, M. Liebeke, N. Dubilier, H.R. Gruber-Vodicka, Sulfur-oxidizing symbionts
1717 without canonical genes for autotrophic CO_2 fixation, MBio. 10 (2019) e01112–19.
1718 N. Lane, W.F. Martin, The energetics of genome complexity, Nature. 467 (2010) 929–934.
1719 S.B. Gould, S.G. Garg, W.F. Martin, Bacterial vesicle secretion and the evolutionary origin of
1720 the eukaryotic endomembrane system, Trends Microbiol. 24 (2016) 525–534.
1721 P.K. Raval, S.G. Garg, S.B. Gould, Endosymbiotic selective pressure at the origin of eukaryotic
1722 cell biology, eLife. 11 (2022) e81033.
1723 C. Acquisti, J. Kleffe, S. Collins, Oxygen content of transmembrane proteins over
1724 macroevolutionary time scales, Nature. 445 (2007) 47–52.

1725 S. Vieira-Silva, E.P.C. Rocha, An assessment of the impacts of molecular oxygen on the
1726 evolution of proteomes, Mol. Biol. Evol. 25 (2008) 1931–1942.
1727 M. Degli Esposti, On the evolution of cytochrome oxidases consuming oxygen, BBA –
1728 Bioenergetics. 1861 (2020) 148304.
1729 J.P. Osborne, R.B. Gennis, Sequence analysis of cytochrome bd oxidase suggests a revised
1730 topology for subunit I, BBA – Bioenergetics. 1410 (1999) 32–50.
1731 J. Jabłońska, D.S. Tawfik, The evolution of oxygen-utilizing enzymes suggests early
1732 biosphere oxygenation, Nat. Ecol. Evol. 5 (2021) 442–448.
1733 M.C. Weiss, F.L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, W.F.
1734 Martin, The physiology and habitat of the last universal common ancestor, Nat.
1735 Microbiol. 1 (2016) 1–8.
1736 B.J. Arnold, I.T. Huang, W.P. Hanage, Horizontal gene transfer and adaptive evolution in
1737 bacteria, Nat. Rev. Microbiol. 20 (2021) 206–218.
1738 S.W. Poulton, M.D. Krom, R. Raiswell, A revised scheme for the reactivity of iron
1739 (oxyhydr)oxide minerals towards dissolved sulfide, Geochim. Cosmochim. Acta. 68
1740 (2004b) 3703–3715.
1741 D.E. Canfield, S.W. Poulton, A.H. Knoll, G.M. Narbonne, G. Ross, T. Goldberg, H. Strauss,
1742 Ferruginous conditions dominated later neoproterozoic deep-water chemistry, Science.
1743 321 (2008) 949–952.
1744 K. Ozaki, K.J. Thompson, R.L. Simister, S.A. Crow, C.T. Reinhard, Anoxygenic
1745 photosynthesis and the delayed oxygenation of Earth's atmosphere, Nat. Commun. 10
1746 (2019) 3026.
1747 A.D. Anbar, A.H. Knoll, Proterozoic ocean chemistry and evolution: A bioinorganic bridge?,
1748 Science. 297 (2002) 1137–1142.

1749 E.K. Moore, B.I. Jelen, D. Giovannelli, H. Raanan, P.G. Falkowski, Metal availability and the
1750 expanding network of microbial metabolisms in the Archaean eon, Nat. Geosci. 10
1751 (2017) 629–636.
1752 E.E. Stüecken, A test of the nitrogen-limitation hypothesis for retarded eukaryote radiation:
1753 Nitrogen isotopes across a Mesoproterozoic basinal profile, Geochim. Cosmochim. Acta.
1754 120 (2013) 121–139.
1755 H. Bothe, O. Schmitz, M.G. Yates, W.E. Newton, Nitrogen fixation and hydrogen metabolism
1756 in cyanobacteria, Microbiol Mol Biol Rev. 74 (2010) 529–551.
1757 R.A. Boyle, T.W. Dahl, G.A. Shields-Zhou, M. Zhu, M.D. Brasier, D.E. Canfield, T.M.
1758 Lenton, Stabilization of the coupled oxygen and phosphorus cycles by the evolution of
1759 bioturbation, Nat. Geosci. 7 (2014) 671–676.
1760 D.E. Canfield, J. Farquhar, Animal evolution, bioturbation, and the sulfate concentration of the
1761 oceans, PNAS. 106 (2009) 8123–8127.
1762 N.J. Butterfield, Early evolution of the Eukaryota, Palaeontology. 58 (2015a) 5–17.
1763 K.M. Fagerbakke, M. Heldal, S. Norland, Content of carbon, nitrogen, oxygen, sulfur and
1764 phosphorus in native aquatic and cultured bacteria, Aquat. Microb. Ecol. 10 (1996) 15–
1765 27.
1766 B.K. Burgess, D.J. Lowe, Mechanism of molybdenum nitrogenase, Chem. Rev. 96 (1996)
1767 2983–3011.
1768 Y. Hu, M.W. Ribbe, Nitrogenase and homologs, J. Biol. Inorg. Chem. 20 (2015) 435–445.
1769 H.L. Rutledge, F. Akif Tezcan, Electron transfer in nitrogenase, Chem. Rev. 120 (2020) 5158–
1770 5193.
1771 R.L. Robson, Characterization of an oxygen-stable nitrogenase complex isolated from
1772 Azotobacter chroococcum, Biochem. J. 181 (1979) 569–575.

1773 W.D. Stewart, M. Lex, Nitrogenase activity in the blue-green alga Plectonema boryanum strain
1774 594, Arch. Microbiol. 73 (1970) 250–260.
1775 N.M. Weare, J.R. Benemann, Nitrogenase activity and photosynthesis in Plectonema
1776 boryanum, J. Bacteriol. 119 (1974) 258–265.
1777 R. Rippka, J.B. Waterbury, Synthesis of nitrogenase by non-heterocystous cyanobacteria,
1778 FEMS Microbiol. Lett. 2 (1977) 83–86.
1779 A.N. Rai, M.B. Syiem, B. Bergmann, Nitrogenase derepression, its regulation and metabolic
1780 changes associated with diazotrophy in the nonheterocystous cyanobacterium
1781 Plectonema boryanum PCC-73110, J. Gen. Microbiol. 138 (1992) 481–491.
1782 H.S. Misra, Oxygen implication in the diazotrophic growth of Plectonema boryanum in dark-
1783 light cycles, Plant Sci. 143 (1999) 135–142.
1784 I. Berman-Frank, P. Lundgren, Y.B. Chen, H. Kupper, Z. Kolber, B. Bergman, P. Falkowski,
1785 Segregation of nitrogen fixation and oxygenic photosynthesis in the marine
1786 cyanobacterium Trichodesmium, Science. 294 (2001) 1534–1537.
1787 I. Berman-Frank, P. Lundgren, P. Falkowski, Nitrogen fixation and photosynthetic oxygen
1788 evolution in cyanobacteria, Res. Microbiol. 154 (2003) 157–164.
1789 M. Staal, S. Rabouille, L.J. Stal, On the role of oxygen for nitrogen fixation in the marine
1790 cyanobacterium Trichodesmium sp, Environ. Microbiol. 9 (2007) 727–736.
1791 T.M. Lenton, S.J. Daines, B.J.W. Mills, COPSE reloaded: An improved model of
1792 biogeochemical cycling over Phanerozoic time, Earth Sci. Rev. 178 (2018) 1–28.
1793 S. Scherer, H. Almon, P. Böger, Interaction of photosynthesis, respiration and nitrogen fixation
1794 in cyanobacteria, Photosynth. Res. 15 (1988) 95–114.
1795 G. Shimakawa, A. Kohara, C. Miyake, Characterization of light-enhanced respiration in
1796 cyanobacteria, Int. J. Mol. Sci. 22 (2021) 342.

1797 B. Bergman, J.R. Gallon, A.N. Rai, L.J. Stal, N2 fixation by non-heterocystous cyanobacteria,
1798 FEMS Microbiol. Rev. 19 (1997) 139–185.
1799 A. Bandyopadhyay, T. Elvitigala, M. Liberton, H.B. Pakrasi, Variations in the rhythms of
1800 respiration and nitrogen fixation in members of the unicellular diazotrophic
1801 cyanobacterial genus Cyanothece, Plant Physiol. 161 (2013) 1334–1346.
1802 S. Rabouille, D.B. Van de Waal, H.C.P. Matthijs, J. Huisman, Nitrogen fixation and respiratory
1803 electron transport in the cyanobacterium Cyanothece under different light/dark cycles,
1804 FEMS Microbiol. Ecol. 87 (2014) 630–638.
1805 G. Sandh, L.H. Xu, B. Bergman, Diazocyte development in the marine diazotrophic
1806 cyanobacterium Trichodesmium, Microbiology (Reading). 158 (2012) 345–352.
1807 A. Torrado, C. Ramirez-Moncayo, J.A. Navarro, V. Mariscal, F.P. Molina-Heredia,
1808 Cytochrome c(6) is the main respiratory and photosynthetic soluble electron donor in
1809 heterocysts of the cyanobacterium Anabaena sp. PCC 7120, BBA – Bioenergetics. 1860
1810 (2019) 60–68.
1811 W.F. Martin, D.A. Bryant, J.T. Beatty, A physiological perspective on the origin and evolution
1812 of photosynthesis, FEMS Microbiol. Rev. 42 (2018) 201–231.
1813 A.S. Steunou, D. Bhaya, M.M. Bateson, M.C. Melendrez, D.M. Ward, E. Brecht, J.W. Peters,
1814 M. Kühl, A.R. Grossman, In situ analysis of nitrogen fixation and metabolic switching
1815 in unicellular thermophilic cyanobacteria inhabiting hot spring microbial mats, PNAS.
1816 103 (2006) 2398–2403.
1817 C. Barth, M.C. Weiss, M. Roettger, W.F. Martin, G. Unden, Origin and phylogenetic
1818 relationships of [4Fe–4S]-containing O2 sensors of bacteria, Environ. Microbiol. 20
1819 (2018) 4567–4586.
1820 M.J. Gruer, P.J. Artymiuk, J.R. Guest, The aconitase family: three structural variations on a
1821 common theme, Trends Biochem. Sci. 22 (1997) 3–6.

1822 O. v. Lushchak, M. Piroddi, F. Galli, V. Lushchak, Aconitase post-translational modification
1823 as a key in linkage between Krebs cycle, iron homeostasis, redox signaling, and
1824 metabolism of reactive oxygen species, Redox Rep. 19 (2014) 8–15.
1825 N.D. Gunawardena, V. Schrott, C. Richardson, T.N. Cole, C.G. Corey, Y. Wang, S.S. Shiva
1826 G.C. Bullock, Aconitase: an iron sensing regulator of mitochondrial oxidative
1827 metabolism and erythropoiesis, Blood. 128 (2016) 74.
1828 A. Mitsui, S. Kumazawa, A. Takahashi, H. Ikemoto, S. Cao, T. Arai, Strategy by which
1829 nitrogen-fixing unicellular cyanobacteria grow photoautotrophically, Nature. 323 (1986)
1830 720–722.
1831 B. Bergman, G. Sandh, S. Lin, J. Larsson, E.J. Carpenter, Trichodesmium– a widespread
1832 marine cyanobacterium with unusual nitrogen fixation properties, FEMS Microbiol. Rev.
1833 37 (2013) 286–302.
1834 R. Rippka, J. Deruelles, J.B. Waterbury, Generic assignments, strain histories and properties
1835 of pure cultures of cyanobacteria, J. Gen. Microbiol. 111 (1979) 1–61.
1836 S. Kihara, D.A. Hartzler, S. Savikhin, Oxygen concentration inside a functioning
1837 photosynthetic cell, Biophys. J. 106 (2014) 1882–1889.
1838 C.F. Demoulin, Y.J. Lara, L. Cornet, C. François, D. Baurain, A. Wilmotte, E.J. Javaux,
1839 Cyanobacteria evolution: Insight from the fossil record, Free Radical Biol. Med. 140
1840 (2019) 206–223.
1841 N.J. Butterfield, Proterozoic photosynthesis – a critical review, Palaeontology. 58 (2015b)
1842 953–972.
1843 K. Kato, N. Miyazaki, T. Hamaguchi, Y. Nakajima, F. Akita, K. Yonekura, J.R. Shen, High-
1844 resolution cryo-EM structure of photosystem II reveals damage from high-dose electron
1845 beams, Comm. Biol. 4 (2021) 382.

1846 H.J. Chiu, J.W. Peters, W.N. Lanzilotta, M.J. Ryle, L.C. Seefeldt, J.B. Howard, D.C. Rees,
1847 MgATP-bound and nucleotide-free structures of a nitrogenase protein complex between
1848 the Leu 127Δ-Fe-protein and the MoFe-protein, Biochemistry. 40 (2001) 641–650.
1849 P. Purushotham, R. Ho, J. Zimmer, Architecture of a catalytic active homotrimeric plant
1850 cellulose synthase complex, Science. 369 (2020) 1089–1094.
1851 J.R. Shen, Y. Inoue, Binding and functional properties of two new extrinsic components,
1852 cytochrome c-550 and a 12-kDa protein, in cyanobacterial photosystem II, Biochemistry.
1853 32 (1993) 1825–1832.
1854 N. Nelson, C.F. Yocum, Structure and function of photosystems I and II, Annu. Rev. Plant
1855 Biol. 57 (2006) 521–565.
1856 A.J. Jasniewski, N.S. Sickerman, Y. Hu, M.W. Ribbe, The Fe protein: An unsung hero of
1857 nitrogenase. Inorganics. 6 (2018) 25.
1858 J.L. Morris, M.N. Puttick, J.W. Clark, D. Edwards, P. Kenrick, S. Pressel, C.H. Wellman, Z.
1859 Yang, H. Schneider, C.P.J. Donoghue, The timescale of early land plant evolution,
1860 PNAS.115 (2018) E2274–E2283.
1861 C.H. Wellman, P.K. Strother, The terrestrialbiota prior to the origin of land plants
1862 (embryophytes):Terrestrialization in the Ordoviciana review of the evidence,
1863 Palaeontology. 58 (2015) 601–627.
1864 T.W. Dahl, S.K. Arens, The impacts of land plant evolution on Earth's climate and oxygenation
1865 state–An interdisciplinary review, Chem. Geol. 547 (2020) 119665.
1866 U. Römling, M.Y. Galperin, Bacterial cellulose biosynthesis: diversity of operons, subunits,
1867 products, and functions, Trends Microbiol. 23 (2015) 545–557.
1868 C. Puginier, J. Keller, P.M. Delaux, Plant-microbe interactions that have impacted plant
1869 terrestrializations, Plant Physiol. 190 (2022) 72–84.

1870 A.J. Krause, B.J. Mills, S. Zhang, N.J. Planavsky, T.M. Lenton, S.W. Poulton, Stepwise
1871 oxygenation of the Paleozoic atmosphere, Nat. Commun. 9 (2018) 4081.
1872 J. de Vries, A. Stanton, J.M. Archibald, S.B. Gould, Streptophyte terrestrialization in light of
1873 plastid evolution. Trends Plant Sci. 21 (2016) 467–476.
1874 E.A. Bayer, H. Chanzy, R. Lamed, Y. Shoham, Cellulose, cellulases and cellulosomes. Curr.
1875 Opin. Struct. Biol. 8 (1998) 548–557.
1876 E.A. Bayer, Y. Shoham, R. Lamed, Lignocellulose-decomposing bacteria and their enzyme
1877 systems, in: E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt, F. Thompson (Eds.),
1878 The Prokaryotes, Springer, Berlin, Heidelberg, 2013: pp. 215–266.
1879 L. Artzi, E.A. Bayer, S. Moraïs, Cellulosomes: bacterial nanomachines for dismantling plant
1880 polysaccharides, Nat. Rev. Microbiol. 15 (2017) 83–95.
1881 I. Mizrahi, Rumen Symbioses, in: E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt, F.
1882 Thompson (Eds.), The Prokaryotes, Springer, Berlin, Heidelberg, 2013: pp. 533–544.
1883 S. Morais, I. Mizrahi, Islands in the stream: From individual to communal fiber degradation in
1884 the rumen ecosystem, FEMS Microbiol. Rev. 43 (2019) 362–379.
1885 T. Nishiyama, S. Hidetoshi, J. de Vries, H. Buschmann, D. Saint-Marcoux, K.K. Ullrich, F.B.
1886 Haas, L. Vanderstraeten, D. Becker, D. Lang, et al., The Chara genome: Secondary
1887 complexity and implications for plant terrestrialization, Cell. 174 (2018) 448–464.
1888 S.J. Hurley, B.A. Wing, C.E. Jasper, N.C. Hill, J.C. Cameron, Carbon isotope evidence for the
1889 global physiology of Proterozoic cyanobacteria, Sci. Adv. 7 (2021) abc8998.
1890 E.R. Hoffarth, K.W. Rothchild, K.S. Ryan, Emergence of oxygen- and pyridoxal phosphate-
1891 dependent reactions. FEBS J. 287 (2020) 1403–1428.
1892 S. Ouchane, A.-S. Steunou, M. Picaud, C. Astier, Aerobic and anaerobic Mg-protoporphyrin
1893 monomethyl ester cyclases in Purple Bacteria: A strategy adopted to bypass the
1894 repressive oxygen control system. J. Biol. Chem. 279 (2004) 6385–6394.

1895 J. Raymond, R.E. Blankenship, Biosynthetic pathways, gene replacement and the antiquity of
1896 life. Geobiology 2 (2005) 199-203.
1897 M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids
1898 Res. 28 (2000) 27–30.
1899 A.J. Enright, S. van Dongen, C.A. Ouzounis, An efficient algorithm for large-scale detection
1900 of protein families, Nucleic Acids Res. 30 (2002) 1575–1584.
1901 B. Buchfink, C. Xie, D. Huson, Fast and sensitive protein alignment using DIAMOND, Nat.
1902 Methods. 12 (2015) 59–60.
1903 M. Kanehisa, M. Furumichi, M. Tanabe, M. Sato, K. Morishima, KEGG: New perspectives on
1904 genomes, pathways, diseases and drugs, Nucleic Acids Res. 45 (2017) D353-D361
1905 B.L. Welch, The generalization of "Student's" problem when several different population
1906 variances are involved, Biometrika. 34 (1947) 28–35.
1907 M.E. Beber, M.G. Gollub, D. Mozaffari, K.M. Shebek, E. Noor, eQuilibrator 3.0 – a platform
1908 for the estimation of thermodynamic constants, Nucleic Acids Res. 50 (2021) D603–
1909 D609.
1910 A. Flamholz, E. Noor, A. Bar-Even, R. Milo, eQuilibrator—the biochemical thermodynamics
1911 calculator, Nucleic Acids Res. 40 (2012) D770–D775.
1912 M.M. Noor, A.P. Wandel, T. Yusaf, Design and development of MILD combustion burner, J.
1913 Mech. Eng. Sci. 5 (2013) 662–676.
1914 A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitz, I. Schomburg, M. Neumann-Schaal,
1915 D. Jahn, D. Schomburg, BRENDA, the ELIXIR core data resource in 2021: new
1916 developments and updates, Nucleic Acids Res. 49(D1) (2021) D498–D508.
1917 K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7:
1918 Improvements in performance and usability, Mol. Biol. Evol. 30 (2013) 772–780.

1919 A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
1920 phylogenies, Bioinformatics. 30 (2014) 1312–1313.
1921 F.D.K. Tria, G. Landan, T. Dagan, Phylogenetic rooting using minimal ancestor deviation, Nat.
1922 Ecol. Evol. 1 (2017) 193.
1923 J. Huerta-Cepas, F. Serra, P. Bork, ETE3: Reconstruction, analysis, and visualization of
1924 phylogenomic data, Mol. Biol. Evol. 33 (2016) 1635–1638.
1925

1926 Supporting information

1927 **S1 Fig. Correlation of protein family verticality and frequency among prokaryotic**
1928 **genomes.** For each protein family the distribution of verticality in relation to the frequency
1929 among 5,655 prokaryotic genomes is plotted **A.** for each protein family and **B.** for each
1930 reaction, combining multiple protein families associated with a reaction. O_2-dependent protein
1931 families are colored in blue and O_2-independent protein families in orange.

1932 **S2 Fig. Distribution of Gibbs energy Δ*G*' for oxygen-dependent and -independent**
1933 **reactions.** The two boxplots are showing the distribution of the Gibbs energy for 288 O_2-
1934 dependent reactions on the left (blue) and 2,139 O_2-independent reactions on the right (orange).
1935 Both reaction types are on average exergonic (mean ΔG' O_2-dependent = -229.6 kJ·mol^{-1},
1936 mean ΔG' O_2-independent = -12.6 kJ·mol^{-1}), but the mean Gibbs energy is overall significantly
1937 lower than in O_2-dependent reactions (Welch's t-test *p*-value = $9.348 \cdot 10^{-52}$, also see S2F
1938 Table).

1939 **S3 Fig. Correlation of reaction frequency versus genome size.** For each species in the
1940 dataset, only the largest strain was used. Panel **A.** shows the correlation for O_2-dependent
1941 reactions and **B.** for O_2-independent reactions (see also Fig 2).

S1 Table. Information on the 5,655 prokaryotic genomes used in this study.

S2 Table. Statistical tests with relevant parameters. A. T-test number of O_2-dependent reactions in aerobes and anaerobes and genomes size of aerobes and anaerobes **B.** T-test distribution of O_2-dependent and O_2-independent verticalities. **C.** T-test distribution of O_2-dependent and O_2-independent verticalities, excluding verticalities greater 1. **D.** Average verticality V_{avg} and standard deviation of all samples taken. **E.** Models of number of reactions in a genome against genome size. **F.** Correlation of $\Delta G'$ and verticality within O_2-dependent and O_2-independent reactions.

S3 Table. Enzyme commission numbers and reaction type for 362 out of the 365 O_2-dependent reactions.

S4 Table. All protein families with respective KEGG Orthology identifier (KO), corresponding reactions, protein family size and verticality *V*. A. List of all 792 protein families linked to O_2-dependent reactions. **B.** List of all 260,191 protein families linked to O_2-independent reactions or without a linked reaction.

S5 Table. Cofactors for 365 O_2-dependent reactions.

S6 Table. Functional categories, their respective average verticality and number of clusters in that category.

S7 Table. Ancestral state reconstruction for 365 O_2-dependent reactions. The origin node for each reaction and the underlying phylogenetic tree in newick format are given.

S8 Table. All O_2-dependent and -independent reactions from KEGG that were linked to protein families. A. List of 365 O_2-dependent reactions with linked protein families and average verticality V_{avg}. **B.** List of 3,018 O_2-independent reactions with linked protein families and average verticality V_{avg}.

S9 Table. Gibbs energy Δ*G*' for O_2-dependent and -independent reactions. A. Gibbs energy Δ*G*' of 288 O_2-dependent reactions. **B.** Gibbs energy Δ*G*' of 2,139 O_2-independent reactions.

Publication 4:

To what inanimate matter are we most closely related and does the origin of life harbor meaning?

Authors: Martin, W. F., Nagies, F. S. P., & do Nascimento Vieira, A.

Published: 2021 in *Philosophies, 6(2). 33-52.*

Contribution of Falk Sascha Per Nagies:

I conceptualized with the other authors the first version of the manuscript. For this, I performed literature research that would among other things lead to the idea of the basic steps for origin of life hypotheses. Then I contributed to the writing, editing and reviewing of the final manuscript.

philosophies

MDPI

Article

To What Inanimate Matter Are We Most Closely Related and Does the Origin of Life Harbor Meaning?

William F. Martin *, Falk S. P. Nagies and Andrey do Nascimento Vieira

Department of Biology, Institute for Molecular Evolution, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany; falk.nagies@hhu.de (F.S.P.N.); nascima@uni-duesseldorf.de (A.d.N.V.)
* Correspondence: bill@hhu.de

Abstract: The question concerning the meaning of life is important, but it immediately confronts the present authors with insurmountable obstacles from a philosophical standpoint, as it would require us to define not only what we hold to be life, but what we hold to be meaning in addition, requiring us to do both in a properly researched context. We unconditionally surrender to that challenge. Instead, we offer a vernacular, armchair approach to life's origin and meaning, with some layman's thoughts on the meaning of origins as viewed from the biologist's standpoint. One can observe that biologists generally approach the concept of biological meaning in the context of evolution. This is the basis for the broad resonance behind Dobzhansky's appraisal that "Nothing in biology makes sense except in the light of evolution". Biologists try to understand living things in the historical context of how they arose, without giving much thought to the definition of what life or living things are, which for a biologist is usually not an interesting question in the practical context of daily dealings with organisms. Do humans generally understand life's meaning in the context of history? If we consider the problem of life's origin, the question of what constitutes a living thing becomes somewhat more acute for the biologist, though not more answerable, because it is inescapable that there was a time when there were no organisms on Earth, followed by a time when there were, the latter time having persisted in continuity to the present. This raises the question of where, in that transition, chemicals on Earth became alive, requiring, in turn, a set of premises for how life arose in order to conceptualize the problem in relation to organisms we know today, including ourselves, which brings us to the point of this paper: In the same way that cultural narratives for origins always start with a setting, scientific narratives for origins also always start with a setting, a place on Earth or elsewhere where we can imagine what happened for the sake of structuring both the problem and the narrative for its solution. This raises the question of whether scientific origins settings convey meaning to humans in that they suggest to us from what kind of place and what kinds of chemicals we are descended, that is, to which inanimate things we are most closely related.

Keywords: origins of life; epistemology; hydrothermal vents; warm little pond; site of life's origin

Citation: Martin, W.F.; Nagies, F.S.P.; do Nascimento Vieira, A. To What Inanimate Matter Are We Most Closely Related and Does the Origin of Life Harbor Meaning? *Philosophies* **2021**, *6*, 33. https://doi.org/10.3390/philosophies6020033

Academic Editor: Rainer Zimmermann

Received: 4 March 2021
Accepted: 12 April 2021
Published: 15 April 2021

1. Introduction

The question of how life emerged is older than science. Historically, various cultures have found satisfactory explanations for how life emerged. Such explanations have been and continue to be of great importance both to society as a whole and to individual members thereof. We will make no attempt to present the origins narratives of different cultures here; every reader will know some examples against which to vet the merit of our claim. The explanations that different cultures have handed down over generations differ in their principles and narrative, though all fulfil a common goal of providing an account of how living things in general and humans in particular fit into the continuum of time. Scientists also have explanations for how life emerged. The explanations that different scientists offer also differ in their principles and narrative, although it is unclear whether scientific origins narratives generally provide an account of how humans fit into

the continuum of time, because most scientific origins narratives do not directly connect to specific kinds of living cells that we can observe today that could serve as a starting point of biological evolution. That is, most scientific origins narratives either offer an account of an abstract or idealized cell or, more often, strive to explain some specific aspect or individual component of cells, rather than addressing the origin of the whole cell. For example, many origins narratives present an account for where RNA came from and what it can do if one assumes the existence of an unlimited supply of activated RNA precursors on the face of the early Earth [1,2]. That is a much simpler task than presenting a narrative for where a whole living cell came from under the premise that there was no unlimited supply of activated RNA precursors anywhere on the early Earth [3,4]. RNA is a molecule: a nucleic acid with three simple components—phosphate, sugar, and bases—that are iterated in a linear polymer that can assume different conformations via folding. It is easier to explain the origin of a molecule than it is to explain the origin of a prokaryotic cell: a small volume of space roughly one micrometer on a side, containing about ten thousand ribosomes, each consisting of about 50% RNA by weight, plus five million individual protein molecules that catalyze on the order of 2000 different chemical reactions that harness environmentally available energy and nutrients to make an imperfect copy of itself in a period of time that might last 20 min or 2000 years. It is also unclear whether scientific origins narratives, whether for molecules or cells, even qualify as science, because the object of investigation lies irretrievably buried in the past, such that even if we were to recreate life from chemicals in a laboratory, we would have no way of ascertaining whether the first cells, our ancestors, arose that way.

Should we then just give up on studying origins? There are good reasons to do so, but at the same time, humans apparently have an innate curiosity about where we came from and how living things came to be the way that we observe them to be in nature. It is part of our human condition to seek and find answers to the question of where we came from. This might relate to the circumstance that humans seem to generally fear things that they cannot explain and tend to find comfort in explanations of otherwise inexplicable phenomena. Scientific narratives for origins do not provide comfort, though they do strive to soothe our curiosity about the living past. The interest of the general public in origins would, of itself, seem to justify origins research.

During the Enlightenment, European science sought independence from religious doctrines, but its scientific revolution had no immediate impact on scientific concepts about origins because that first generation of people whom we would tend to call scientists by today's measures, while taking a pragmatic rather than dogmatic approach to natural phenomena, had no grasp whatsoever of what makes living things alive and what they are made of, let alone how such might have arisen. Priestly's discovery of oxygen [5] was a step towards identifying chemical requirements for life, but offered little help towards understanding what living things are or where they come from.

Though today's libraries can boast much progress in organic chemistry, microbiology, geoscience, astrobiology, and computational biology, the question of how life began remains perennially near the top of the ten most wanted list of major (and underdetermined) problems or questions facing science. Traditional approaches to the problem start from the simple and work forward to the complex. If we place the origins problem in the context of chemistry that was known around 1900, scientists knew that atoms form molecules and that cells contain molecules, though most of the molecules were too large and complex to identify. Scientists knew that cells transform substrates into other molecules in some manner, but the nature of such transformations was obscure, as was the question of whether any chemical transformations were universal among cells. Otto Warburg helped to unravel a good many of the chemical transformations germane to the reactions that keep cells alive [6,7]. Today, we know a great deal about how cells transform molecules during the process of growth (physiology, the chemical reactions of growth) and how the information that directs the synthesis of a new cell is stored and retrieved, but the origins problem of how such reactions started remains, although some newer findings do harbor hints of

progress in that they identify a distinct chemical connection between geochemical reactions and what might have been the first biochemical reactions [8,9].

Tracking down the first chemical reactions that led to metabolic pathways would allow us to pin a starting point on the map of physiological reactions. If we had a robust clue about the starting point of metabolism, we could probe the concept further by experiment. However, if we consult the literature about the myriad ways in which cells make a living physiology, the undertaking of assigning a starting point might seem like a task without chance of success. There are hundreds of biochemical pathways used by countless prokaryotic lineages, many with unique traits, and most microbes that we know to exist (because we can see them) are still vastly under characterized because they do not (yet) grow in pure culture. But is the search for a starting point of physiology really futile? Among all the pathways we know, whereby hundreds are used for harvesting energy alone [10], common sense has it that they cannot all be equally old. Some pathways must be older and some must be younger. They cannot all have arisen at once. There must have been a temporal progression in metabolic evolution [11]. Finding a suitable starting point for the evolution of biochemical reactions would help constrain an otherwise unwieldy problem, because it would identify the chemical reactions at the very base of the life process as it arose.

Given the daunting diversity of microbial metabolism, finding its starting point appears as a typical needle in a haystack problem: hopeless. However, let us suppose that the needle really *is* in the haystack, that is, that among the thousands of pathways we know, one really is the oldest. To find it, all we have to do is use a really strong magnet, and suddenly, finding that needle is easy, a day's work for one person—with that we can identify the starting point of metabolism, the origin of the chemical reactions that produce cells. Where can we obtain such a magnet? We will return to that at the end of this paper.

1.1. Narratives Are Set in Places

In recent years, rapid progress in the field of genome sequencing and analytical biogeochemistry has enabled the creation of robust molecular libraries that allow a far more precise categorical classification of life forms than anything we had previously. In the past however, many of these categorizations were performed via generalization derived from complex and systematic observation of macroscopic animals and their traits over many generations in a populational context. In most cases, evolutionary experiments were not feasible. In the 1800s, the rise of evolution as a theory changed everything about the way scientists thought about the history of life, and brought with it the possibility of a single origin from which all else might have emerged via time and change. Though no one in Darwin's time could have provided a satisfactory answer to the fundamental problem of origins, his speculations on the first organism's environment, the "warm little pond" in his correspondence with Hooker, paved the way to the proposal of the first concrete experiments in the field many decades later [12]. Darwin left a hint about the conundrum of origins in a time where geochemical data for the deep past were scarce, if not completely unavailable. Still today, the answers to origins are buried irretrievably deep in the past. In Darwin's day, no experimental approach was robust enough to spark progress on the issue, thus leaving most hypotheses to the sheer taste of personal belief systems. To some extent, that is still true today.

Darwin set a rarely recognized tradition in origins research; he started the origins narrative exactly the same way as traditional cultural and religious narratives usually have, namely he started with a place and a setting, a warm little pond: "But if (& oh what a big if) we could conceive in some warm little pond with all sorts of ammonia & phosphoric salts,—light, heat, electricity &c. present, that a protein compound was chemically formed, ready to undergo still more complex changes . . . " [13]. Curiously, origins research still works that way today in that scientific origins narratives usually start with a place.

Of course, it would have helped Darwin to know more about the chemical reactions of cells. In his day, the most severe impediment to approaching origins in concrete language was a lack of knowledge about what cells are made of in detail and how they work as a

chemical reaction. In Darwin's day and well into the 20th century, biologists relied on the concept of protoplasm to explain the seemingly inexplicable properties of life [14]. The term protoplasm traces to the middle of the 1800s and the Czech physiologist Jan E. Purkinje and the German physiologist Hugo von Mohl [14,15]. At the heart of the protoplasm concept was the notion that a special vital force, a vis vitalis, is associated with living substances but is lacking in non-living substances, creating, in essence, two different kinds of matter. Strong proponents of protoplasm were called vitalists, their opponents were called mechanists [16]. In Darwin's time and thereafter, biologists had no chemical understanding of the life process within cells. Vitalists held that protoplasm represented a special kind or organization of matter that bestows the property of life and distinguishes living from non-living things. In his book on protoplasm, Drysdale [17] characterized protoplasm as follows " ... the elements are in a state of combination not to be called chemical at all in the ordinary sense, but one which is utterly sui generis. That, in fact, no albumin, fibrin, myosin, protagon, or fats exist at all in the living matter, but that the sum of the elements of all these is united into a compound, for which we have no chemical name, and the complex mode in which the atoms are combined we can form no idea; and it is only at the moment of death that those chemical compounds, with which we are familiar, take their origin. [...] Vitality is thus a property inherent in each particle of the living matter, and all the parts of a complex organism differ in function, each part has a specific kind of vitality peculiar to itself." Clearly, if one held that life was chemically distinct from other forms of matter, then the key to understanding the origin of life was understanding the origin of protoplasm, a substance immune to direct investigation, but whose properties, in theory, remained stable enough over the eons of life's history to distinguish major lineages in the living world [18]. Protoplasm might be seen as a kind of dialectic capitulation before the severity of the origins problem—its nature is too complex, hence unknowable.

Despite a lack of understanding of what cells are and how they function, and despite the absence of an empirically supported concept of the conditions on early Earth, Mereschkowsky [18] inferred that the first cells arose as the young Earth was still hot and covered in boiling water. Like Darwin, he had a setting, but it was much harsher than a warm little pond, because he thought the first cells must have been extremophiles. In his view, they were extremely small, they could thrive at temperatures of 100 °C, they were anaerobes, they had the ability to synthesize proteins and carbohydrates from inorganic substances without the help of chlorophyll and they were resilient against alkaline solutions, concentrated salt solutions, sulfur compounds, and diverse toxins [18]. Those extreme conditions sound much like those of modern theories for an autotrophic origin of life in hydrothermal vents, theories that we can find in modern college textbooks [19]. Haeckel [20] espoused similar but much less detailed thoughts about the nature of the first cells. However, even today, the thought of origins in a dark, deep, hot and oxygen-free abyss, from gasses that react all by themselves in the presence of catalysts [9], conjures a hellish notion that has almost a demonic character for proponents of the warm little pond [21].

By the time of Oparin's book [12], and by Miller's [22] experiment at the latest, biological and microbiological renderings of origins gave way to chemical renderings of origins; beginnings from CO_2 gave way to beginnings in some form of pond. For the majority of the last century, Darwin's hypothesis, enriched by Oparin's [12] lengthy book and Haldane's [23] tersely argued narrative of a prebiotic broth, or organic soup, inspired many researchers to embrace a simple explanation to origins: The action of sunlight on carbon in the ocean could, in principle, generate all sorts of organics, which could somehow assemble themselves in solution into something more complex. This approach assumed that at the early stages of planetary evolution, Earth's atmosphere was poor in oxygen (O_2) but rich in reducing gasses such as ammonia (NH_3), methane (CH_4), and hydrogen (H_2), based on studies of other planetary bodies like the outer Jovian planets in our solar system. This kind of atmosphere, which was used in Miller's experiment, generated amino acids and other organics using energy supplied by an electric spark, simulating lightning.

Many subsequent hypotheses built upon this foundation, adhering to the premise that, in essence, the origins problem consisted of two components: an initial process or phase of molecular synthesis to obtain the basic building blocks of life (the origin of soup) in a particular setting, followed by a subsequent process or phase of molecular organization that arranged pre-existing components into more highly structured state (self-organization). In the 1970s, a concept emerged that RNA molecules could compete with one another for activated RNA monomers (resources), such that the fastest replicating molecules became the most fit in a Darwinian sense [24] by bringing forth the most progeny. This idea was exceptionally well suited to empirical endeavor and experimental tests. It gave rise to over five decades of productive research on a concept and a field that have become widely known as the RNA world. In many modern papers, one can read about the RNA world as if it were an established "fact", a known whose properties merely require further characterization [1]. In other papers, the term RNA world is used more or less synonymously with the origins of life [25]. There is now much evidence underpinning the view that RNA molecules can multiply if they are provided with a steady stream of biochemically pure precursors, so much evidence in fact that some have begun to ponder the "ecology" of RNA molecules in such a world [1] as if it, the RNA world, were an observation in nature as opposed to a premise.

Fascination with RNA has, however, distracted from the more important and still unanswered question of whether an RNA world ever existed and if it existed [26,27], whether it had anything to do with the origin of things that are actually alive—microbial cells [28]—notwithstanding the sobering observation that RNA, given inorganic substrates, is clearly no more alive than an isolated protein, a fat droplet, or a grain of starch. A critic will immediately interject that we are constraining the issue too much by imposing the criterion of living from "inorganic substrates", but if we accept the evidence indicating that the moon forming impact converted the Earth to a ball of boiling magma, converting all carbon to CO_2, then CO_2 was the initial form of carbon from which life emerged [29]. Another critic will complain that after the magma oceans cooled, there was a late heavy bombardment that brought a veneer of new organics from space, countering the CO_2 dictum [30]. We would counter that the late heavy bombardment probably never even occurred; it is more likely an artefact of (mis)interpreting lunar craters [31]. Another critic might complain that we are countering philosophical critique with appeals to evidence rather than logic. Yes, as we said at the outset, this is an informal essay about origins and meaning.

1.2. Molecules or Cells?

For the biologist, it is sometimes more useful to discuss the origin of "microbial cells" than to discuss the origin of "life" because if one debates the origin of life, one can debate in very open terms and with a long literature the question of what life is, leading to philosophically fertile but biologically barren fields of discourse. If we constrain the issue to concern the origin of things that are obviously alive, microbial cells, and without whose existence there would be no creatures that debate, then we get closer to the issue, the dimension of which causes many researchers to give up: how do we get from the early Earth to a fully fledged free living microbial cell whose main function it is to first convert environmental energy into expendable chemical currency for survival and then, if resources permit, to grow from inorganic compounds all by itself. The hard part of seeing the origin of life as the origin of microbes is that one first has to learn a lot of biology, the nuts and bolts of what cells are and how they work, so as to be able to verbalize specific processes underpinning the origins of the components and the whole. That is why it is much more convenient to reduce the origins problem to the origin of RNA, which is a very simple explanandum compared to a cell, or to detour into definitions of life where there are almost no constraints in observation to guide our reasoning. If we insist on defining the problem as the origin of microbes, we can get straight to work on that problem without having to debate the existence of an RNA world or define life before starting. Solving the problem

of the origin of microbes would then be left up to biologists, where it arguably belongs (and where it began), because (with *pro domo* immodesty) nobody knows the individual chemical reactions that compose living things and the ~2000 enzymes that catalyze those reactions in a given microbial cell better than biologists.

We also face the problem that the origin of microbes was a singular event by all logic, because of the universality of central metabolism [32] and the universality of the genetic code [33], and that origin event occurred roughly 4 billion years ago according to isotope data [34]. Again, even if we performed an experiment in which microbial cells demonstrably arose de novo from inorganic compounds in a laboratory experiment, we would still have no evidence in hand that life (our unicellular microbial ancestors) actually arose that way, we would just have a narrative richer than our current ones on how it might have occurred. A priori, we have no access to a more systematic approach to natural problems than of hypothesis, experimentation, observation and interpretation. It seems clear that we can use the scientific method to explore aspects related to the origins problem, but the problem might be without solution. No final answers to be had at origins? That would be an honest admission, but it would not satisfy the curiosity of scientists and the public when it comes to wanting to know where we come from. It is part of our human nature to want to know about the past. All human cultures have a natural interest in the question of where living things came from, a question that, in contrast to agriculture or medicine, has no obvious practical importance unless, of course, concepts of origins help provide meaning, and meaning simultaneously has practical importance, a question that is well beyond the scope of this paper.

If we want to probe problems rooted deep in the past, such as the emergence of the first prokaryotic cells and the diversification of primordial prokaryotic lineages some 4 billion years ago, we have to make some assumptions for the sake of moving forward, and we have to state the assumptions explicitly. Darwin started in a pond, but he never justified why he chose to start in a pond. There might be several ways to recreate life in the laboratory, but the universal laws of thermodynamics that govern the chemical reaction that we call biology apply across the tree of all life. All cells require proteins that are made of amino acids; cells are about 50% protein by dry weight. Genetic material is made of nucleotides; cells are about 20% RNA and 3% DNA by dry weight [35]. Cells require a constant far from the equilibrium system from which to harness energy; cells always synthesize much more ATP than they need, often three times more [36,37], because ATP is the main energy currency of the cell and the converse would violate the 2nd law of thermodynamics. Armed with a few observations of this type, we can discuss the origins problem in the comfort that we are just thinking about it, not trying to solve it.

1.3. Teleology and the Notion of Epistemological Obstacles

Epistemology (the theory of knowledge, the methods to obtain knowledge, and the scope of knowledge) is generally traced to ancient Greece with the works of Plato and Aristotle, who pondered what we know and the distinction between what exists and what does not. In the seventeenth century, John Locke's philosophic categorization and understanding of knowledge became a branch discipline of philosophy per se. Gaston Bachelard [38] contextualized the fields of scientific knowledge and its intrinsic robustness and is regarded by some to be one of the founding fathers of modern epistemology [39–41]. Bachelard [38] proposed that the scientific mind must develop against human nature and that epistemological obstacles are natural stressing points to scientific knowledge that hinders scientific progress. He proposed obstacle categories that have affected science. These include the first experience, or how the first object of research can establish a bias for further experiments, leading to mischaracterization of the subject of study. This can snowball to generate a premature generalization of the problem, another obstacle. The limitations of natural language, the broad use of scientific terms or analogies and the restrictive nature of explanatory words can further impede understanding as can animism as an explanation to natural problems or indiscriminate use of knowledge with its often-

overlooked teleological conception. Few scientific questions lend themselves more readily to animism and hidden teleology (the explanation of phenomena in terms of the purpose they serve rather than of the cause by which they arise) than the issue of how inorganic material became living cells.

According to Mayr [42], no other ideology influenced biology as much as teleology. Mayr categorizes the teleological obstacle into five subsections: (i) Teleonomic processes, referring to phenomena with a goal-directed purpose guided by an implicit program. (ii) Teleomatic processes that are not guided by a pre-established program but instead follow passively a directed path of events as a result of the action of natural laws [43]. (iii) Cosmic teleology refers to processes guided by a supreme force. (iv) Adaptive programs, or processes in which there is a direction of events towards a posteriori outcomes such as those observed in non-Darwinian evolutionary theories. Lastly, there is (v) purposeful behavior in which the subject of change is directing its behavior towards a certain need. Mayr proposes that the use of teleological arguments must be avoided at all costs as no phenomenon in nature is innately teleological. In Masatoshi Nei's mutation driven theory of evolution [44], teleological components are forbidden by mechanistic means, the primary process-driving vectors of evolution being mutation, which all things being equal can be (somewhat safely) assumed to be blind. As such, the use of teleology in the field of origins fuels the discussion of whether there is a role to dogma in science and if such concepts hinder the acquisition and transmission of knowledge by (falsely, we assume) giving purpose to the natural phenomena permeating chemical and biological evolution.

1.4. Could Nei's Conjecture Be True?

Yet before we look into stumbling blocks of how we think about evolution, let us consider something that has bothered the senior author of the present paper for some time. Is Nei right? Does mutation really set a vector in evolution [44], or is mutation unlimited in scope with selection doing the work of bringing forth new forms from the set of all possible? Worse, is there any way we can even approach an avenue towards obtaining an answer to the problem? We know more about our world than in Plato's day, maybe we can get an estimate. We start with the question of how many cells have ever lived and the estimate that roughly 10^{30} microbial cells are alive today [45]. Most of these cells are living in the subsurface or marine sediment where they are growing very slowly, if at all, some with doubling times estimated as hundreds or even thousands of years [46]. We generously give them fast growth and an average doubling time of ten generations per year. Ten generations (doublings) per year would mean a 1024-fold increase of 10^{30} microbes or 10^{33} new microbes per year, but global biomass cannot increase 1000-fold every year, as nutrients are limiting (fortunately, for us all). However, we are still generous and say that each microbe nonetheless manages 10 generations per year (accompanied by many microbes that simultaneously die). For 10^{30}, microbes that gives us 10^{33} individual doublings per year, or, summing up across 4 billion years, we have about $4{\cdot}10^{42}$ generations on the books, which we conveniently round to 10^{42} new microbes in history because for most of earth's history, fewer microbes existed per year than today, owing to the lack of oxygen. How many mutations have there been in those ~10^{42} cells? Let us say that a microbe has a mutation rate similar to that in *Escherichia coli*, on the order of 10^{-3} per generation per genome [47,48]. At that rate, for every 1000 cells that undergo one cell division, one new mutation will accrue. Over 4 billion years, that means that every 1000th cell gets a new mutation, or about 10^{39} new mutations in all of evolution. That sounds like a big number. Yet maybe we underestimated the number of cells badly, for example that we should be using much faster doubling times, so we throw in a factor of 10^9 for good measure, giving a rather generous estimate of 10^{48} mutations that have occurred in evolution—a big number, close to the number of water molecules in the oceans, but smaller than the approximate number of protons in the universe, ~10^{80}. How does the number of roughly 10^{48} mutations in evolution stack up against the number of possible mutational states for microbial genomes? Are they roughly equal?

For that, we have to estimate how many mutational states are possible. If each microbe has one genome of only 1000 genes with avg. 1000 bp each, then there are 10^6 bp in the genome, with four possible bases per position, or $4^{1,000,000}$ possible sequences, which is about $10^{602,060}$, or close enough to $10^{600,000}$ to round it for our purposes. We can also count gene transfers into the mutation category, which does not generate sequence variants beyond the $10^{600,000}$, although it could increase the number of sites in the genome, raising the number of possible mutational states. The number of possible sequences that could be realized for small genomes during evolution, ~$10^{600,000}$, is hundreds of thousands of orders of magnitude greater than the number of mutations that took place, generously estimated at ~10^{48}. One might interject that many sequences cannot be realized because of the nature of the genetic code, the size of proteins, the occurrence of stop codons, etc. To conservatively take that into account, let us say that the structure of genes, reading frames, proteins and the genetic code constrains evolution so tightly that only one single base per gene is allowed to vary by mutation. That is an extreme exaggeration of reality, but we are just trying to get an estimate. Allowing only one base per gene to mutate per genome, keeping all other bases constant throughout evolution, reduces the number of possible sequences from $4^{1,000,000}$ to 4^{1000} possible sequences because we have assumed 1000 genes per genome, and that still translates to 10^{603} possible sequences, such that even then, the number of possible sequences, over-conservatively estimated, that could be realized for small genomes during evolution, is still 10^{500} times greater than the number of mutations that took place, overgenerously estimated, recalling that the number of protons in the universe is roughly 10^{80} for comparison.

Regardless of how we cut the cake of possibilities, mutation rules, it would seem. Does Darwin's natural selection even figure into this? Yes, in the big picture, selection is important, obviously, as it weeds out unviable sequence variants in specific environments. However, the point here is that mutation never even had a ghost of a chance to explore what combinations are possible with genomes in four billion years. Nei once said in a lecture to 1000 evolutionary biologists in Puerto Rico "Natural selection is overrated". The audience gasped, some quietly scoffed. The person sitting next to one of us (WM) asked "Did he really just say that?", "Just listen and pay attention" was the reply.

The foregoing back of the envelope calculation shows that the vectors of evolutionary lineages from origins to the present were mechanistically limited by mutation (and driven forward in time by thermodynamics, we know without the calculation). There was never a world on this planet where life explored all possibilities, with selection pruning all viable states from the set of all possible. If mutation is blind, which we generally hold to be true, then the course of evolution is only one of (for practical purposes) an infinite number of possibilities. The path that mutations, not selection, took brought us to where we are. That is possibly irrelevant to the issue of how we think about origins, early evolution, and meaning, but possibly not. Stephen J. Gould famously asked whether if we could replay the tape of evolution, a similar result would unfold. Some argue yes [49]. The foregoing indicates that the answer is a clear no, although life still has to obey the 2nd law, meaning that the finite number of chemical reactions on Earth (there are only 92 natural elements and they have $\leq$8 oxidation states each) that can be harnessed to support life sets constraints. Is natural selection overrated? It very possibly could be, and we are hardly the first to suggest that it is so.

1.5. The Concept of Epistemological Obstacles in Hypothesis Pervading Origins

We can briefly consider the role of epistemological obstacles and how they influence the progress of a heavily polarized field such as origins. Given the influence of teleology in evolutionary biology, we can ask how the categories proposed by Mayr might perpetuate a dogmatic line of thought that was deeply rooted in evolutionary biology. Although abundant in the humanities, scientific case studies of epistemological obstacles in the fields of origins and early evolution are rare by comparison [50–54]. We contextualize the concept of epistemological obstacles in origins and early evolution by identifying

teleological arguments in origins and estimating their impacts. Being biologists by training, our philosophical scope is narrow.

The year 2020 witnessed continued progress in the chemistry of origins. The gap between opinion-based to experiment-driven hypothesis laid a concrete environment for new theoretical and experimental work. However, the origins field is divided into schools with conflicting, mutually incompatible viewpoints and heated debates as to how life started. This is a sign that the problem is underdetermined, like the origin of eukaryotes [55], suggesting a role for philosophy of science to further progress. In simplistic terms, experimental work on the topic should have an explicit theoretical basis that generates predictions, that is based on robust methods, that involves the careful observation of data that lends itself to meaningful interpretation. In practice, there is no way to perform such work without personal bias. We are the product of our experience, never genuinely objective. The literature is vast and nobody knows all of it. How to even begin forming hypotheses under such suboptimal conditions? Even in the absence of knowledge, humans are blessed with curiosity. Curiosity can lead to good experimental questions, regardless of whether we have any idea of how curiosity works. Plain curiosity can work in favor of progress. In our favor also, good scientific work ethics has it that we want to do sound work that will stand the test of time, that will make a difference as science moves forward. We all know papers that had an impact on our own view of the world, papers that made a difference in how we approach scientific work. Often, we want our papers to be like those positive examples that, together with curiosity, led us down the path that we have taken. No objectivity there. Curiosity leads investigations like mutation leads evolution.

A problem is how to reduce human bias in a field like origins. Given the multitude of theories on origins, it is only natural that scientists rely on intuition, which might be driven by curiosity. Although highly subjective, it is intuition that tells us what clues to follow and which directions to pursue further. Curiosity sets a course, intuition decides among alternative paths forward. Of course, intuition is also influenced by beliefs. With the ever-growing information pool in the context of schools following mutually incompatible hypotheses, it is only human to follow the information that meshes most harmoniously with what we "think" we already "know", leading different schools in different directions, which is perhaps the best thing that can happen to a field, for if everybody in a field is following the same idea, there is little opportunity for discovery.

1.6. Settings in Origins Theories: Where to Start?

There are currently a number of competing hypotheses about origins that, despite their differences, have a few things in common. For example, all theories have to assume a constant supply of carbon, nitrogen, sulfur, phosphorus and trace elements as all life is made of these; hence, we can presume that the first cells were as well, such that a constant element supply was needed to replenish reserves and to permit growth. In addition, these elements had to be supplied in such a way that they could react to form covalent bonds and simple primordial molecules, which needed to become concentrated enough to react further and form more complex molecules. All the above can be achieved in a variety of ways under certain assumed primordial earth conditions, such that the assumptions concerning the early Earth and the specifics of geochemical conditions at the site of origins constitute a main criterion by which the alternative hypothesis differs. Concerning the transition from simple chemicals to systems of molecular self-organization, few hypotheses have clear concepts, help coming from network theories and concepts of autocatalysis [56,57]. The flux into and out of a steady state pool of reproducibly formed molecules (a metabolic "identity") has to be or become stable enough for them to form increasingly complex structures, ultimately seeding a process that we would describe as evolution today. This leads to roughly seven phases in origin of life hypotheses.

1. The initial setting and medium including soluble materials and catalysts;
2. Generation of organic molecules—substrates and energy;

3. Concentration of organics;
4. Increased molecular system complexity;
5. Stable but far from equilibrium environment fostering the newly formed system;
6. Emergence of the first free living cells;
7. The lifestyle of those first cells.

The setting that one assumes for origins bears upon all other aspects of the molecular process that leads to life or components of life or cells or microbes, so it is worthwhile to look at settings and what they bring to the origins issue.

The pond. As mentioned in the introduction, the first hints of chemically driven origins came from Darwin himself. He proposed that life started as a set of molecules that changed over time to what we consider life today. This gave rise to Oparin's [12] and Haldane's [23] primordial soup concept. This hypothesis assumed that as there were no primary consumers for the first organic molecules, over time and the action of continuous synthesis, these molecules would accumulate to high concentrations in the primordial oceans while also being free to interact with each other until life emerged. A few decades later, Miller [22] experimentally demonstrated the synthesis of some basic building blocks of life by applying electric spark to a gaseous mixture in order to simulate the early atmosphere. The demonstrable synthesis of organic molecules in the Miller experiment had broad impact because it left no doubt that basic molecules of life are nothing special (in the sense of protoplasm, for example)—they can arise through simple reactions of inorganic compounds under imaginable prebiotic conditions. There was also no specific mechanism yet that could bridge the gap between simple molecules and a living, replicating and evolving system. A possible scenario was that by chance, the first gene emerged and then life began [58]. Yet, this would make the emergence so improbable that even Oparin [59] criticized the idea [60]. Nevertheless, that did not stop researchers from building upon these ideas and realizing that, at some point, simple organic molecules were probably created from gases and that over time, these molecules had to assemble into cell-like states [61].

Clay. A possible role of clay minerals in origins was proposed in 1951 by Bernal [62], according to which life could have possibly emerged from a clay matrix [60]. In this hypothesis, organics could be adsorbed into minerals, ultimately concentrating them. A first step into how the necessary concentrations for life's reactions might have been formed apart from a primordial soup. Cairns-Smith [60] then changed this idea to what can be boiled down to self-replicating crystal systems evolving by virtue of their structural changes, which is guided by the specific organic compounds they adsorb. Eventually, these clays would undergo "a genetic metamorphosis" to purely organic units of heritage and thus, life. The incorporation of clay or mineral matrices into the theory has several advantages that would be reused in several later origin of life hypotheses. Nevertheless, another problem with several assumptions on this and other hypotheses slowly became apparent. The primordial soup was deemed unrealistic if not synthesized under increasingly specific circumstances. For example, due to the equilibrium of reactions, it would be impossible for organics to accumulate in the concentrations often assumed [63,64]. This constraint prompted the postulation of drying out phases. That is, the pond is a site for chemical synthesis and then dries up to concentrate and products of synthesis so that they might react further. One round of drying is clearly not sufficient to drive the process forward, leading to the introduction of wet–dry cycles at the site of origins [65]. Several experiments showed potential to polymerize activated RNA monomers using clay matrices as condensing agents [66].

Space. Regardless of how they polymerized, life's basic molecules had to come from somewhere [67]. Besides electricity and UV light, a popular scenario began with the discovery of amino acids (albeit mostly non-proteinogenic) in meteorites [68]. This seeded the panspermia hypothesis, a relatively old idea [69], that life could travel between planets by virtue of interstellar radiation [70] or hitchhike by meteorites [71]. This, of course, did not solve the problem of how life or its basic molecules were created in the first place.

However, it succeeded in motivating some to change the setting for origins to Mars, for example [72], using a reasoning that a few particular chemical reactions relevant to life (polymerizations) would work better in the absence of water, such that a planet with less water might be a more likely site of origin. Cells are about 90% water by weight; cells have no problem with water, they have a problem when there is not enough water.

Black smokers. With the first reports of deep-sea hydrothermal vents in [73], there was immediate discussion of the possibility that these structures played a major role in the origin of life [74,75]. The initial proposal was promptly criticized as the temperatures at the black smoker types of vents, the first ones discovered, were around 400 °C, too hot to sustain life [76,77]. Life is able to survive and even thrive at temperatures of 110–120 °C [78,79], but not 400 °C, which left this particular hypothesis initially with few supporters. Vents, however, impacted the origins issue deeply in that a different setting for the origins problem gave rise to different paths of origins narrative development [80]. The initial focus on temperature at black smokers was unfortunate in that it distracted from the more important observation that they presented a continuously far from equilibrium chemical setting with many metals that play a role as catalysts in microbial metabolism and with many sources of chemical energy to fuel physiological reactions [75].

Volcanic iron and sulfur. In the wake of debates about black smokers, volcanic settings for origins were proposed, where minerals once again played a central role, but not in the form of clays that could absorb chemicals, rather in the form of iron sulfur minerals that performed catalytic activity [81]. As the start of metabolism, Wächtershäuser proposed a citric acid cycle that was energetically driven forward by pyrite synthesis [82]. This harkened back to the conditions proposed by Mereschkowsky [18], and a start of metabolism from CO_2, or autotrophic origins. Volcanos offered an alternative to ponds as a possible site where one could imagine origins, and it led to a different narrative, one that started with surface metabolism, or protobiological reactions in two dimensions before moving to three-dimensional cells [83]. As with the temperature aspect of black smokers, critics honed in on one specific aspect of the theory that seemed particularly vulnerable to criticism—the concept of two-dimensional life [84].

A nuclear geyser. Another setting was suggested by Ebisuzaki and Maruyama [85], a nuclear geyser in which natural radioactive elements provided a great deal of energy in the form of heat and radiation. In the nuclear geyser hypothesis, such sites could provide so much ionizing and thermal energy that organic molecules would easily be generated. This scenario emphasized the energy requirements of early life and that the combination of wet–dry cycles should allow molecular concentration and synthesis of more complex organics. Yet, they make no specific proposals for how LUCA emerged from this new setting. In addition, the damaging effects of radiation from intense radioactive decay for biological molecules are well known. Along the same lines, the minimotif synthesis hypothesis [86] does not try to claim how LUCA emerged nor how the primordial soup was created; rather, this hypothesis tries to bridge the gap between the primordial soup and the first evolving RNA world system by creating a scenario of continuously interacting and thus, evolving minimotifs of the first macromolecules.

An RNA world, somewhere. Minimotifs are examples of addressing the origin of one component of a cell. The paradigm for this principle is the RNA world. In the 1960s, many hypotheses involving a seemingly central role of RNA as a simple, possibly self-reproducing component were emerging [87,88]. With the recognition that RNA molecules had catalytic activity, the primordial soup hypothesis morphed into an RNA world hypothesis [89,90]. The term "RNA world" itself was coined by Gilbert [91] in the context of Cech's discovery of self-splicing introns [92], with RNA molecules envisaged as recombining in a phase of early evolution of RNA genes preceding the advent of DNA genes. The notion of an "RNA world" quickly gained a much broader meaning in the context of chemical origins however, by incorporating aspects of White's suggestion that coenzymes preceded proteins [93] and the enzyme catalyzed in vitro RNA replication experiments of Eigen from the 1970s [94]. This gave rise to the concept of natural selection among exponentially

replicating molecules before the advent of natural selection among cells. It is an enticing idea, but it might be wrong [95]. Proponents of the idea that RNA was the starting point of evolution have gone so far as to propose the existence of specific mountainside settings on the early Earth with individual meteorite impacts and a variety of atmosphere-derived cyanide compounds and temperature ranges across hundreds of degrees as Archaean incarnations of chemical reaction conditions that lead to highly specific RNA monomer synthesis in the laboratory [2]. That is, a set of (geochemically quite questionable) Earth conditions are proposed based on the need for their existence in order to accommodate RNA synthesis for the RNA world, creating a curious kind of cart (RNA) and horse (Earth) problem. The early Earth conditions that geochemists infer are not in the slightest conducive to the specific synthesis of RNA bases [29,96] requiring a different early Earth than the one that geochemists have to offer for the RNA world to work. Some researchers question the need for an independent replicator at any point in prebiotic evolution [97,98]. Others build entire theories upon the premise of its existence [1], but the question of what natural early Earth setting would accommodate the specific synthesis and operation of a replicator is usually unanswered. Nisbet's proposal of hydrothermal springs to harbor an RNA world [99] with organic synthesis starting from methane and ammonia found little resonance among RNA world proponents. Hydrolysis of RNA has traditionally been seen as an insurmountable problem for hydrothermal vents, and it still is in some circles [21]. Yet it is only a problem if one believes in an RNA world that exists in free aqueous solution, as in typical laboratory experiments, because alkaline hydrothermal vents harbor many local environments of low water activity [8].

H_2-producing hydrothermal vents. Deep sea hydrothermal vents that emerge from serpentinizing systems have a chemistry that is fundamentally different from that of black smokers in that their effluent is alkaline and ~100 °C rather than acidic and ~400 °C. The existence of such alkaline vents on the early Earth was inferred from studies of metal ore deposits [99,100]. The first modern deep sea alkaline hydrothermal vent system discovered was the Lost City hydrothermal field [101]. The term alkaline is important because the alkalinity in vents is generated in the process of serpentinization: an unfamiliar word for a philosophy journal but an important one because it produces H_2, a well-known currency of chemical energy, in the crust. H_2 exits the vent in the effluent; the more alkaline the water is, the more H_2 (chemical energy) it contains [102]. The effluent of modern serpentinizing vents contains about four orders of magnitude more H_2 than modern H_2-dependent microorganisms need to grow. The overall chemistry at alkaline vents shares long overlooked similarity with the main energy harnessing reactions of microbes that live from H_2, which can react with CO_2 to generate organic compounds spontaneously because the reaction releases energy (it is an exergonic reaction) as recently demonstrated in the laboratory [9]. Alkaline vents continuously synthesize systems of inorganic microcompartments that can concentrate the products of organic synthesis where they are made, offering a means to concentrate the products so that they can react further. The continuous supply of H_2 interfacing with inexhaustible reserves of CO_2 on the early Earth provides a continuous source of chemical energy to drive the system forward towards higher complexity. In continuity of this reaction, the first cells to emerge from the vent have a carbon and energy metabolism that is based on the exergonic reaction of H_2 and CO_2, in which the cells synthesize acetate (acetogenic bacteria) or methane (methanogenic archaea) [4,103,104] as the main product of energy metabolism. In that view, the first free living cells were acetogens and methanogens, the starting point from which further physiological evolution took place [11]. This narrative addresses the seven criteria outlined above in a level of explicit detail that others do not; it is compatible with the LUCA inferred from genomic reconstructions [105], while at the same time, connecting the origin of life narrative with real microbial cells that still grow in such environments today.

Variations on a vent. There are variants of the H_2-dependent vent narrative for origins. Some replace the chemical energy of the vent with ultraviolet light [106], but the theories do not connect to modern cells because UV light kills cells and no cells can harness energy

or live from UV light. This generates a hybrid between a pond and a vent, relying on UV light as an energy source like Oparin and Haldane did at a time before anyone knew how cells conserve energy or grow. The hot spring hypothesis [107–109] goes back to the warm little pond, with the settings changed in that the pond is fed by hydrothermal springs (providing thermal energy and wet–dry cycles with first molecules possibly coming from meteorites). There is no chemical role for the vent in the hot spring narrative, it just supplies water and salts, nor is there a connection to the lifestyle of the first cells. Other variants replace the chemical energy of the reaction between H_2 and CO_2 with chemical energy of the reaction of methane with nitrogen oxides [110], the problem being that in that narrative, the vent is no longer needed because the source of nitrogen oxides is lightning [111], effectively translocating the setting of the vent to the surface, as in the UV model [106] or the hot spring pond. While the H_2 + CO_2 reaction of the H_2 theory produces essential compounds of central metabolism using only hydrothermal catalysts like awaruite or magnetite overnight [9], no similar laboratory reaction is available to which the methane nitrogen oxide narrative could appeal. Other settings include tidal cycles that provided wet–dry cycles [112] for cyclic DNA (or RNA) synthesis. A different hypothesis that relies heavily on the RNA world is the hydrogel hypothesis [113], where the authors explain the stability that was needed for the system to evolve into primordial life by relying on the physicochemical features of a hydrogel. These structures would be similar to today's cytoplasm in retaining its form until cell-like structures with lipid encapsulation could form. This hypothesis proposes that LUCA possibly emerged from a biofilm-like hydrogel formation [114].

2. Conclusions

The foregoing provides some examples to underpin the claim in this paper that scientific origins narratives tend to start in a setting, a place that we can imagine. How do those settings connect to the chemical reactions of life? In an earlier passage, we said that trying to find the starting point in the evolution of metabolic pathways is like trying to find a needle in a haystack, but that the problem becomes manageable for one person in one day (given an average sized haystack) if we have a very powerful magnet. One of us has been claiming that we have found that starting point. If so, what was the magnet? The magnet would appear to be physiology: the reactions that cells use to make a living are the reactions that make ATP, which is in turn the molecule whose hydrolysis drives all else in the cell forward (thermodynamics). For very few organisms, the reactions that synthesize ATP are the same reactions that supply carbon to metabolism from CO_2: acetogens and methanogens [11,102–104]. If we look for simple forms of energy metabolism, we find that the simplest ones (anaerobic, H_2-dependent, CO_2 reducing, using linear pathways instead of cycles, many transition metals as catalysts, lacking cytochromes and quinones hence, older than heme) are similar to naturally occurring exergonic geochemical reactions. When we react H_2 with CO_2 in the laboratory in water in the presence of a simple hydrothermal mineral as a catalyst—either the pure nickel metal alloy awaruite, the iron oxide magnetite, or the iron sulfide greigite—the backbone of carbon and energy metabolism unfolds all by itself: formate, acetate and pyruvate accumulate in physiologically relevant amounts [9]. All coincidence, possibly, but possibly not.

There is no shortage of settings for origin of life hypotheses and new proposals are constantly emerging. At the most basic level, these fall into two categories with regard to setting: surface or subsurface, with energy coming from the heavens or energy coming from below. This might or might not have a bearing on our unspoken preferences for scientific origins narratives because it would bear rather directly on the possible meaning of origins, facing us with a fairly distinctive choice for what we would rather call our closest relatives among objects of inanimate matter—pointing to the heavens above us or pointing to the Earth beneath us. Genomic reconstructions of LUCA, that is, inferences of the habitat and lifestyle of the last universal ancestor of all cells based on the evidence for

early life preserved in genomes, recover the physiology of a cell that lived from H_2, CO_2, N_2, and H_2S in a hot metal-rich environment [105].

In the beginning of our solar system, the Earth accreted from remnants of a supernova, so says the planetary narrative. In that sense, our closest relative is stardust from the heavens. We live on the surface. In that sense, our closest relative is the surface. We are buried in the Earth, and when the microbes have had the last word, our carbon becomes CO_2 that becomes life of future generations. That was true for our ancestors, in the sense that we come from below. Looking way back, before there was photosynthesis, all life on this planet depended on H_2 from serpentinization [115,116]. In terms of energy for life, H_2 (chemical energy) was the precursor of light (electromagnetic energy). In that sense, our closest inanimate relative would be the reaction of rocks with water in the Earth's crust. Were that so, does it have meaning? Worse, does it have purpose? These are questions that, as a rule, biologists cannot readily answer. However, we can pose them.

In the abstract, we surrendered to the challenge of defining either life or meaning. Greater courage was displayed by Cleland and Chyba [117], who delved into both definitions, concluding that there is no easy definition of life and that the issue regarding the definition of meaning is no simpler, although exploration of both can enrich the way we approach the problems to which these definitions pertain, whereby caution is warranted because many things we now hold to be true as a reference system for definitions might turn out to be false. Cleland [118] later concluded that current approaches to deriving definitions of life are all "deeply flawed". Though not well versed in the long literature of life's definitions, we tend to concur. We furthermore contend that one can study processes relevant to the origin of things that are obviously alive (microbes) and even obtain useful insights into the problem without the strict need for definitions of the process (life) whose origin is under study.

By all accounts, life arose when the Earth was still young. Carbon isotopes indicate the existence of autotrophs by 3.95 [34] or even 4.1 [119] billion years ago, although such studies also point out the possibility that the kind of carbon isotope fractionation found in such ancient samples might be the result of geochemical H_2-dependent CO_2 reduction rather than the biological process. For proponents of H_2-dependent autotrophic origins under hydrothermal conditions, the difference between geochemical and biological CO_2 reduction via the most ancient pathway [9,103] boils down to a matter of grade concerning the sophistication of the catalysts.

There is an issue at origins concerning what the Earth "really" looked like >3.5 billion years ago when life was already up and running, having already evolved a sulfur-based energy metabolism [120,121]. In the least human-friendly versions, the oceans were 10 km deep because the roughly one (or more) ocean volume of water that is sequestered in the crust and mantle today was still in the ocean [122], meaning that there was no land, a severe problem for some theories about the site of origin. The continents did not form until about 3 billion years ago, long after there was life and they were made of rocks that are very low in silicate content (mafic) meaning that they were particularly prone to generate H_2 via serpentinization [123,124]. The atmosphere was like that of Venus, mainly CO_2 and N_2 [29], meaning that the interface of H_2-producing vents with a CO_2-containing atmosphere and ocean would have generated sites with a good fit to the H_2-producing hydrothermal vent idea.

There is also an issue of whether life's origin (less so its meaning) might have something to do with dissipative structures of the kind that Prigogine [125] described. Dissipative structures are discussed in the context of the origin of physicochemical order, as in Zhabotinsky reactions, and in the context of life [126]. Though skeptical that life could be seen as belonging to a category that would fall under the label of dissipative structures, we can say that living cells might have some properties in common dissipative structures, as the latter only arise in far from equilibrium systems, whereby both living things and H_2-producing hydrothermal vents are far from equilibrium systems. Dissipative structure or not, what is the source of life's order? It of course depends on who one asks and where

one thinks how life might have arisen. If life arose from the reaction of H_2 and CO_2 in hydrothermal vents [103,125], as we contend, then its structure emerges from the geometry of carbon bonds as carbon oxides undergo reduction with electrons and hydrides from H_2, reacting to generate longer and more diverse carbon chains [9] while reacting with other elements like nitrogen and sulfur to produce the amino acids, nucleosides, and cofactors that comprise life [127,128], through overall reactions that are thermodynamically favored [129] under the conditions of hydrothermal vents.

In that way, the form and function of molecules of life emerge from the geometry of orbitals in carbon, and the order in cells emerges from the geometry and properties of its molecules as they are synthesized with the help of microbial energy metabolism. That is how life arises today. This order does not conflict with entropy. Measurements of entropy change during growth have repeatedly shown that the entropy change in cells is always zero or close to zero because, "cells are assembled in a spontaneous process" [130]. That is, if a cell has what it needs to grow, it organizes environmentally available components into more of itself as an effortless by-product of the exergonic growth process [103]. The circumstance that central compounds of life arise from H_2 and CO_2 without enzymes [9] might be taken to mean that that life has a natural tendency to emerge under suitable conditions—but which conditions? We have listed a few of the proposals here. Microbes say: H_2-producing hydrothermal vents. Humans have more diverse views. Identifying inanimate matter to which we are most closely related would help us anticipate what to expect, chemically, during the search for life elsewhere—even if it finds us first.

Author Contributions: Conceptualization, A.d.N.V., F.S.P.N., W.F.M.; writing—original draft preparation, final manuscript preparation, A.d.N.V., F.S.P.N., W.F.M.; writing—review and editing, A.d.N.V., F.S.P.N., W.F.M.; revision, W.F.M.; supervision, W.F.M.; funding acquisition, W.F.M. All authors have read and agreed to the published version of the manuscript.

Funding: The authors thank the DFG (Ma 1426/21-1), ERC (666,053), and VW foundation (93,046 and 96,742) for funding.

Acknowledgments: We thank Rainer Zimmermann for the kind invitation to submit a paper for this issue and Dan Graur for many stimulating discussions on mutation in the evolutionary process.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kudella, P.W.; Tkachenko, A.V.; Salditt, A.; Maslov, S.; Braun, D. Structured sequences emerge from random pool when replicated by templated ligation. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2018030118. [CrossRef]
2. Patel, B.H.; Percivalle, C.; Ritson, D.J.; Duffy, C.D.; Sutherland, J.D. Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nat. Chem.* **2015**, *7*, 301–307. [CrossRef]
3. Martin, W.; Russell, M.J. On the origins of cells: A hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2003**, *358*, 59–85. [CrossRef] [PubMed]
4. Martin, W.; Russell, M.J. On the origin of biochemistry at an alkaline hydrothermal vent. *Philos. Trans. R. Soc. B Biol. Sci.* **2007**, *362*, 1887–1926. [CrossRef] [PubMed]
5. Priestley, J.; Hey, W. XIX. Observations on different kinds of air. *Philos. Trans. R. Soc. Lond.* **1772**, *62*, 147–264. [CrossRef]
6. Krebs, H.A. Otto Heinrich Warburg, 1883-1970. *Biogr. Mem. Fellows R. Soc.* **1972**, *18*, 628–699.
7. Höxtermann, E. A comment on Warburg's early understanding of biocatalysis. *Photosynth. Res.* **2007**, *92*, 121–127. [CrossRef] [PubMed]
8. Vieira, A.D.N.; Kleinermanns, K.; Martin, W.F.; Preiner, M. The ambivalent role of water at the origins of life. *FEBS Lett.* **2020**, *594*, 2717–2733. [CrossRef] [PubMed]
9. Preiner, M.; Igarashi, K.; Muchowska, K.B.; Yu, M.; Varma, S.J.; Kleinermanns, K.; Nobu, M.K.; Kamagata, Y.; Tüysüz, H.; Moran, J.; et al. A hydrogen-dependent geochemical analogue of primordial carbon and energy metabolism. *Nat. Ecol. Evol.* **2020**, *4*, 534–542. [CrossRef]
10. Amend, J.P.; Shock, E.L. Energetics of overall metabolic reactions of thermophilic and hyperthermophilic Archaea and Bacteria. *Fems Microbiol. Rev.* **2001**, *25*, 175–243. [CrossRef]
11. Martin, W.F. Physiology, phylogeny, and the energetic roots of life. *Period. Biol.* **2017**, *118*, 343. [CrossRef]
12. Oparin, A.I. *The Origin of Life*; Moscow Worker Publisher: Moscow, Russia, 1924.

13. Darwin, C.R. Darwin Correspondence Project. Available online: https://www.darwinproject.ac.uk/letter/DCP-LETT-7471.xml (accessed on 3 February 2021).
14. Liu, D. The cell and protoplasm as container, object, and substance, 1835–1861. *J. Hist. Biol.* **2016**, *50*, 889–925. [CrossRef]
15. Hall, T.S. Ideas of life and matter. *Philos. Sci.* **1969**, *39*, 101–102.
16. Geison, G.L. The protoplasmic theory of life and the vitalist-mechanist debate. *Isis* **1969**, *60*, 273–292. [CrossRef]
17. Drysdale, G.J.S. *The Protoplasmic Theory of Life*, 1st ed.; Bailliere, Tindall & Cox: London, UK, 1874.
18. Kowallik, K.V.; Martin, W.F. The origin of symbiogenesis: An annotated English translation of Mereschkowsky's 1910 paper on the theory of two plasma lineages. *Biosystems* **2021**, *199*, 104281. [CrossRef]
19. Madigan, M.T.; Bender, K.S.; Buckley, D.H.; Sattley, M.; Stahl, D.A. *Brock Biology of Microorganisms*, 15th ed.; Pearson Global Edition: New York, NY, USA, 2019.
20. Haeckel, E. *Natürliche Schöpfungs-Geschichte. Gemeinverständliche Wissenschaftliche Vorträge Über Die Entwickelungslehre. Zehnte Verbesserte Auflage. Zweiter Theil: Allgemeine Stammesgeschichte*; Georg Reimer: Berlin, Germany, 1902.
21. Marshall, M. How the first life on Earth survived its biggest threat—water. *Nature* **2020**, *588*, 210–213. [CrossRef] [PubMed]
22. Miller, S.L. A production of amino acids under possible primitive earth conditions. *Science* **1953**, *117*, 528–529. [CrossRef] [PubMed]
23. Haldane, J.B.S. The origin of life. *Ration. Annu.* **1929**, *148*, 3–10.
24. Eigen, M.; Gardiner, W.; Schuster, P.; Winkler-Oswatitsch, R. The origin of genetic information. *Sci. Am.* **1981**, *244*, 88–118. [CrossRef]
25. Powner, M.W.; Gerland, B.; Sutherland, J.D. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nat. Cell Biol.* **2009**, *459*, 239–242. [CrossRef]
26. Shapiro, R. Prebiotic cytosine synthesis: A critical analysis and implications for the origin of life. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 4396–4401. [CrossRef]
27. Shapiro, R. Small Molecule Interactions were Central to the Origin of Life. *Q. Rev. Biol.* **2006**, *81*, 105–126. [CrossRef] [PubMed]
28. Martin, W.; Baross, J.; Kelley, D.; Russell, M.J. Hydrothermal vents and the origin of life. *Nat. Rev. Genet.* **2008**, *6*, 805–814. [CrossRef]
29. Sossi, P.A.; Burnham, A.D.; Badro, J.; Lanzirotti, A.; Newville, M.; O'Neill, H.S. Redox state of Earth's magma ocean and its Venus-like early atmosphere. *Sci. Adv.* **2020**, *6*, eabd1387. [CrossRef]
30. Zahnle, K.J.; Lupu, R.; Catling, D.C.; Wogan, N. Creation and evolution of impact-generated reduced atmospheres of early Earth. *Planet. Sci. J.* **2020**, *1*, 11. [CrossRef]
31. Mann, A. Cataclysm's end. A popular theory about the early solar system comes under fire. *Nature* **2018**, *553*, 393–395. [CrossRef] [PubMed]
32. Smith, E.; Morowitz, H.J. Universality in intermediary metabolism. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 13168–13173. [CrossRef]
33. Abel, P. Evidence for the universality of the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* **1964**, *29*, 185–187. [CrossRef]
34. Tashiro, T.; Ishida, A.; Hori, M.; Igisu, M.; Koike, M.; Méjean, P.; Takahata, N.; Sano, Y.; Komiya, T. Early trace of life from 3.95 Ga sedimentary rocks in Labrador, Canada. *Nat. Cell Biol.* **2017**, *549*, 516–518. [CrossRef] [PubMed]
35. Schönheit, P.; Buckel, W.; Martin, W.F. On the origin of heterotrophy. *Trends Microbiol.* **2016**, *24*, 12–25. [CrossRef]
36. Tempest, D.W.; Neijssel, O.M. The status of YATP and maintenance energy as biologically interpretable phenomena. *Annu. Rev. Microbiol.* **1984**, *38*, 459–486. [CrossRef]
37. Russell, J.B. The energy spilling reactions of bacteria and other organisms. *J. Mol. Microbiol. Biotechnol.* **2007**, *13*, 1–11. [CrossRef] [PubMed]
38. Bachelard, G. *Formation de l'Esprit Scientifique*, 5th ed.; Librairie Philosophique J. VRIN: Paris, France, 1934.
39. Herscovics, N.; Linchevski, L. A cognitive gap between arithmetic and algebra. *Educ. Stud. Math.* **1994**, *27*, 59–78. [CrossRef]
40. Brousseau, G.; Balacheff, N. *Theory of Didactical Situations in Mathematics: Didactique des Mathématiques, 1970–1990*, 1st ed.; Springer: Dordrecht, The Netherlands, 1997.
41. Cornu, B. Limits. In *Advanced Mathematical Thinking*; J.B. Metzler: Dordrecht, The Netherlands, 2002; pp. 153–166.
42. Mayr, E. *What Makes Biology Unique? Considerations on the Autonomy of a Scientific Discipline*, 1st ed.; Cambridge University Press: Cambridge, MA, USA, 2005.
43. Mayr, E. The Idea of Teleology. *J. Hist. Ideas* **1992**, *53*, 117. [CrossRef]
44. Nei, M. *Mutation-Driven Evolution*, 1st ed.; Oxford University Press: Oxford, UK, 2013.
45. Whitman, W.B.; Coleman, D.C.; Wiebe, W.J. Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 6578–6583. [CrossRef]
46. Hoehler, T.M.; Jørgensen, B.B. Microbial life under extreme energy limitation. *Nat. Rev. Genet.* **2013**, *11*, 83–94. [CrossRef]
47. Foster, P.L.; Lee, H.; Popodi, E.; Townes, J.P.; Tang, H. Determinants of spontaneous mutation in the bacterium Escherichia coli as revealed by whole-genome sequencing. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E5990–E5999. [CrossRef]
48. Sprouffske, K.; Aguilar-Rodríguez, J.; Sniegowski, P.; Wagner, A. High mutation rates limit evolutionary adaptation in *Escherichia coli*. *PLoS Genet.* **2018**, *14*, e1007324. [CrossRef]
49. Orgogozo, V. Replaying the tape of life in the twenty-first century. *Interface Focus* **2015**, *5*, 20150057. [CrossRef]
50. Bedau, M. Can biological teleology be naturalized? *J. Philos.* **1991**, *88*, 647–655. [CrossRef]
51. Wächtershäuser, G. The origin of life and its methodological challenge. *J. Theor. Biol.* **1997**, *187*, 483–494. [CrossRef] [PubMed]

52. Galli, L.M.G.; Meinardi, E.N. the role of teleological thinking in learning the Darwinian model of evolution. *Evol. Educ. Outreach* **2010**, *4*, 145–152. [CrossRef]
53. Toepfer, G. Teleology and its constitutive role for biology as the science of organized systems in nature. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **2012**, *43*, 113–119. [CrossRef] [PubMed]
54. Ribeiro, M.G.L.; Larentis, A.L.; Caldas, L.A.; Garcia, T.C.; Terra, L.L.; Herbst, M.H.; Almeida, R. On the debate about teleology in biology: The notion of "teleological obstacle". *História Ciências Saúde-Manguinhos* **2015**, *22*, 1321–1333. [CrossRef]
55. Bonnin, T. Monist and pluralist approaches on underdetermination: A case study in evolutionary microbiology. *J. Gen. Philos. Sci.* **2020**, 1–21. [CrossRef]
56. Hordijk, W.; Steel, M.; Kauffman, S.A. The structure of autocatalytic sets: Evolvability, enablement, and emergence. *Acta Biotheor.* **2012**, *60*, 379–392. [CrossRef]
57. Xavier, J.C.; Hordijk, W.; Kauffman, S.; Steel, M.; Martin, W.F. Autocatalytic chemical networks at the origin of metabolism. *Proc. R. Soc. B Biol. Sci.* **2020**, *287*, 20192377. [CrossRef]
58. Muller, H.J. Pilgrim trust lecture—The gene. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **1947**, *134*, 1–37.
59. Oparin, A.I. *The Origin of Life on the Earth*, 3rd ed.; Academic Press Inc.: New York, NY, USA, 1957.
60. Cairns-Smith, A. The origin of life and the nature of the primitive gene. *J. Theor. Biol.* **1966**, *10*, 53–88. [CrossRef]
61. Fox, S.W. A Theory of macromolecular and cellular origins. *Nat. Cell Biol.* **1965**, *205*, 328–340. [CrossRef]
62. Bernal, J.D. *The Physical Basis of Life*, 1st ed.; Routledge and Paul: London, UK, 1951.
63. Sillen, L. Oxidation state of Earths ocean and atmosphere. I. A model calculation on earlier states. Myth of probiotic soup. *Ark. Kemi* **1965**, *24*, 431.
64. Hulett, H. Limitations on prebiological synthesis. *J. Theor. Biol.* **1969**, *24*, 56–72. [CrossRef]
65. Ross, D.S.; Deamer, D. Dry/wet cycling and the thermodynamics and kinetics of prebiotic polymer synthesis. *Life* **2016**, *6*, 28. [CrossRef]
66. Ponnamperuma, C.; Shimoyama, A.; Friebele, E. Clay and the origin of life. *Orig. Life Evol. Biosph.* **1982**, *12*, 9–40. [CrossRef] [PubMed]
67. Miller, S.L.; Urey, H.C.; Oró, J. Origin of organic compounds on the primitive earth and in meteorites. *J. Mol. Evol.* **1976**, *9*, 59–72. [CrossRef]
68. Kvenvolden, K.A.; Lawless, J.G.; Ponnamperuma, C. Nonprotein amino acids in the Murchison meteorite. *Proc. Natl. Acad. Sci. USA* **1971**, *68*, 486–490. [CrossRef]
69. Arrhenius, S. Die Verbreitung des Lebens im Weltenraum. *Die Umsch.* **1903**, *7*, 481–485.
70. Weber, P.; Greenberg, J.M. Can spores survive in interstellar space? *Nature* **1985**, *316*, 403–407. [CrossRef]
71. Melosh, H.J. The rocky road to panspermia. *Nature* **1988**, *332*, 687–688. [CrossRef]
72. Benner, S.A.; Kim, H.-J. The case for a Martian origin for Earth life. In Proceedings of the Instruments, Methods, and Missions for Astrobiology XVII, San Diego, CA, USA, 9–13 August 2015; p. 96060C.
73. Corliss, J.B.; Dymond, J.; Gordon, L.I.; Edmond, J.M.; Von Herzen, R.P.; Ballard, R.D.; Green, K.; Williams, D.; Bainbridge, A.; Crane, K.; et al. Submarine thermal springs on the Galápagos Rift. *Science* **1979**, *203*, 1073–1083. [CrossRef]
74. Corliss, J.B.; Baross, J.; Hoffman, S. An hypothesis concerning the relationship between submarine hot springs and the origin of life on Earth. *Oceanol. Acta* **1981**, *SP*, 59–69.
75. Baross, J.A.; Hoffman, S.E. Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Orig. Life Evol. Biosph.* **1985**, *15*, 327–345. [CrossRef]
76. Joyce, G. Hydrothermal vents too hot? *Nature* **1988**, *334*, 564. [CrossRef]
77. Nisbet, E.G. Origin of life. *Nature* **1989**, *337*, 23. [CrossRef]
78. Brock, T.D.; Isaksen, M.F.; Jannasch, H.W. Life at high temperatures. *Science* **1985**, *230*, 132–138. [CrossRef]
79. Kurr, M.; Huber, R.; Jannasch, H.W.; Fricke, H.; Trincone, A.; Kristjansson, J.K.; Stetter, K.O. *Methanopyrus kandleri*, gen. and sp. nov. represents a novel group of hyperthermophilic methanogens, growing at 110 °C. *Arch. Microbiol.* **1991**, *156*, 239–247. [CrossRef]
80. Maden, B.H. No soup for starters? Autotrophy and the origins of metabolism. *Trends Biochem. Sci.* **1995**, *20*, 337–341. [CrossRef]
81. Wächtershäuser, G. Groundworks for an evolutionary biochemistry: The iron-sulphur world. *Prog. Biophys. Mol. Biol.* **1992**, *58*, 85–201. [CrossRef]
82. Wächtershäuser, G. Pyrite formation, the first energy source for life: A hypothesis. *Syst. Appl. Microbiol.* **1988**, *10*, 207–210. [CrossRef]
83. Wächtershäuser, G. Before enzymes and templates: Theory of surface metabolism. *Microbiol. Rev.* **1988**, *52*, 452–484. [CrossRef]
84. De Duve, C.; Miller, S.L. Two-dimensional life? *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 10014–10017. [CrossRef]
85. Ebisuzaki, T.; Maruyama, S. Nuclear geyser model of the origin of life: Driving force to promote the synthesis of building blocks of life. *Geosci. Front.* **2017**, *8*, 275–298. [CrossRef]
86. Schiller, M.R. The minimotif synthesis hypothesis for the origin of life. *J. Transl. Sci.* **2016**, *2*, 289–296. [CrossRef] [PubMed]
87. Crick, F. The origin of the genetic code. *J. Mol. Biol.* **1968**, *38*, 367–379. [CrossRef]
88. Orgel, L. Evolution of the genetic apparatus. *J. Mol. Biol.* **1968**, *38*, 381–393. [CrossRef]
89. Kruger, K.; Grabowski, P.J.; Zaug, A.J.; Sands, J.; Gottschling, D.E.; Cech, T.R. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell* **1982**, *31*, 147–157. [CrossRef]

90. Pace, N.R.; Marsh, T.L. Rna catalysis and the origin of life. *Orig. Life Evol. Biosph.* **1985**, *16*, 97–116. [CrossRef] [PubMed]
91. Gilbert, W. Origin of life: The RNA world. *Nature* **1986**, *319*, 618. [CrossRef]
92. Cech, T.R.; Zaug, A.J.; Grabowski, P.J. In vitro splicing of the ribosomal RNA precursor of tetrahymena: Involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* **1981**, *27*, 487–496. [CrossRef]
93. White, H.B. Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.* **1976**, *7*, 101–104. [CrossRef] [PubMed]
94. Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **1971**, *58*, 465–523. [CrossRef]
95. Baross, J.A.; Martin, W.F. The ribofilm as a concept for life's origins. *Cell* **2015**, *162*, 13–15. [CrossRef] [PubMed]
96. Zahnle, K.; Arndt, N.; Cockell, C.; Halliday, A.; Nisbet, E.; Selsis, F.; Sleep, N.H. Emergence of a habitable planet. *Space Sci. Rev.* **2007**, *129*, 35–78. [CrossRef]
97. Shapiro, R. A replicator was not involved in the origin of life. *Iubmb Life* **2000**, *49*, 173–176. [CrossRef]
98. Nisbet, E.G. Origin of life: RNA and hot-water springs. *Nature* **1986**, *322*, 206. [CrossRef]
99. Russell, M.J.; Daniel, R.M.; Hall, A.J.; Sherringham, J.A. A hydrothermally precipitated catalytic iron sulphide membrane as a first step toward life. *J. Mol. Evol.* **1994**, *39*, 231–243. [CrossRef]
100. Russell, M.J.; Hall, A.J. The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. *J. Geol. Soc.* **1997**, *154*, 377–402. [CrossRef] [PubMed]
101. Kelley, D.S.; Karson, J.A.; Früh-Green, G.L.; Yoerger, D.R.; Shank, T.M.; Butterfield, D.A.; Hayes, J.M.; Schrenk, M.O.; Olson, E.J.; Proskurowski, G.; et al. A serpentinite-hosted ecosystem: The Lost City hydrothermal field. *Science* **2005**, *307*, 1428–1434. [CrossRef] [PubMed]
102. Preiner, M.; Xavier, J.C.; Sousa, F.L.; Zimorski, V.; Neubeck, A.; Lang, S.Q.; Greenwell, H.C.; Kleinermanns, K.; Tüysüz, H.; McCollom, T.M.; et al. Serpentinization: Connecting geochemistry, ancient metabolism and industrial hydrogenation. *Life* **2018**, *8*, 41. [CrossRef]
103. Martin, W.F. Older than genes: The acetyl CoA pathway and origins. *Front. Microbiol.* **2020**, *11*, 817. [CrossRef]
104. Martin, W. On the ancestral state of microbial physiology. In *Life Strategies of Microorganisms in the Environment and in Host Organisms*; Amann, R., Goebel, W., Schink, B., Widdel, F., Eds.; Wissenschaftliche: Darmstadt, Germany, 2008; Volume 96, pp. 53–60.
105. Weiss, M.C.; Sousa, F.L.; Mrnjavac, N.; Neukirchen, S.; Roettger, M.; Nelson-Sathi, S.; Martin, W.F. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **2016**, *1*, 16116. [CrossRef]
106. Mulkidjanian, A.Y.; Bychkov, A.Y.; Dibrova, D.V.; Galperin, M.Y.; Koonin, E.V. Origin of first cells at terrestrial, anoxic geothermal fields. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E821–E830. [CrossRef]
107. Damer, B.; Deamer, D. Coupled phases and combinatorial selection in fluctuating hydrothermal pools: A scenario to guide experimental approaches to the origin of cellular life. *Life* **2015**, *5*, 872–887. [CrossRef] [PubMed]
108. Damer, B.; Deamer, D. The hot spring hypothesis for an origin of life. *Astrobiology* **2020**, *20*, 429–452. [CrossRef] [PubMed]
109. Longo, A.; Damer, B. Factoring origin of life hypotheses into the search for life in the solar system and beyond. *Life* **2020**, *10*, 52. [CrossRef] [PubMed]
110. Nitschke, W.; Russell, M.J. Beating the acetyl coenzyme A-pathway to the origin of life. *Philos. Trans. R. Soc. B Biol. Sci.* **2013**, *368*, 20120258. [CrossRef] [PubMed]
111. Ducluzeau, A.-L.; Van Lis, R.; Duval, S.; Schoepp-Cothenet, B.; Russell, M.J.; Nitschke, W. Was nitric oxide the first deep electron sink? *Trends Biochem. Sci.* **2009**, *34*, 9–15. [CrossRef]
112. Lathe, R. Fast tidal cycling and the origin of life. *Icarus* **2004**, *168*, 18–22. [CrossRef]
113. Trevors, J.T.; Pollack, G.H. Hypothesis: The origin of life in a hydrogel environment. *Prog. Biophys. Mol. Biol.* **2005**, *89*, 1–8. [CrossRef]
114. Trevors, J.T. Hypothesized origin of microbial life in a prebiotic gel and the transition to a living biofilm and microbial mats. *C. R. Biol.* **2011**, *334*, 269–272. [CrossRef]
115. Martin, W.F.; Bryant, D.A.; Beatty, J.T. A physiological perspective on the origin and evolution of photosynthesis. *FEMS Microbiol. Rev.* **2018**, *42*, 205–231. [CrossRef] [PubMed]
116. McMahon, S.; Parnell, J. The deep history of Earth's biomass. *J. Geol. Soc.* **2018**, *175*, 716–720. [CrossRef]
117. Cleland, C.E.; Chyba, C.F. Does "life" have a definition? In *Planets and Life: The Emerging Science of Astrobiology*; Sullivan, W.T., Baross, J.A., Eds.; Cambridge University Press: Cambridge, UK, 2007; pp. 119–131.
118. Cleland, C.E. Life without definitions. *Synthese* **2011**, *185*, 125–144. [CrossRef]
119. Bell, E.A.; Boehnke, P.; Harrison, T.M.; Mao, W.L. Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 14518–14521. [CrossRef]
120. Wacey, D.; Kilburn, M.R.; Saunders, M.; Cliff, J.; Brasier, M.D. Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nat. Geosci.* **2011**, *4*, 698–702. [CrossRef]
121. Morrison, P.R.; Mojzsis, S.J. Tracing the early emergence of microbial sulfur metabolisms. *Geomicrobiol. J.* **2020**, *38*, 66–86. [CrossRef]
122. Korenaga, J.; Planavsky, N.J.; Evans, D.A.D. Global water cycle and the coevolution of the Earth's interior and sur-face environment. *Philos. Trans. R. Soc. A* **2017**, *375*, 20150393. [CrossRef]

123. Dhuime, B.; Wuestefeld, A.; Hawkesworth, C.J. Emergence of modern continental crust about 3 billion years ago. *Nat. Geosci.* **2015**, *8*, 552–555. [CrossRef]
124. Tang, M.; Chen, K.; Rudnick, R.L. Archean upper crust transition from mafic to felsic marks the onset of plate tecton-ics. *Science* **2016**, *351*, 372–375. [CrossRef]
125. Prigogine, I. Time, Structure, and Fluctuations. *Science* **1978**, *201*, 777–785. [CrossRef]
126. Kondepudi, D.K.; De Bari, B.; Dixon, J.A. Dissipative structures, organisms and evolution. *Entropy* **2020**, *22*, 1305. [CrossRef] [PubMed]
127. Martin, W.F. Hydrogen, metals, bifurcating electrons, and proton gradients: The early evolution of biological energy conservation. *FEBS Lett.* **2012**, *586*, 485–493. [CrossRef] [PubMed]
128. Wimmer, J.L.E.; Vieira, A.d.N.; Xavier, J.C.; Kleinermanns, K.; Martin, W.F.; Preiner, M. The autotrophic core: An ancient network of 404 reactions converts H_2, CO_2, and NH_3 into amino acids, bases, and cofactors. *Microorganisms* **2021**, *9*, 458. [CrossRef] [PubMed]
129. Amend, J.P.; McCollom, T.M. Energetics of biomolecule synthesis on early Earth. In *ACS Symposium Series*; American Chemical Society (ACS): Washington, DC, USA, 2010; pp. 63–94.
130. Hansen, L.D.; Criddle, R.S.; Battley, E.H. Biological calorimetry and the thermodynamics of the origination and evo-lution of life. *Pure Appl. Chem.* **2009**, *81*, 1843–1855. [CrossRef]

Concluding remarks

The path from abiotic molecules to life itself is difficult to describe. Over the last century, several hypotheses have been proposed to explain how life was created. And these hypotheses are often opposed to each other. But each hypothesis still has a specific starting point and circumstances that define it. It is difficult to discern which of these starting points is the "right" one, so biases can easily affect the scientific investigation. This is often the case with problems of ancient evolution, which was explored with the origin of life in the third publication of this thesis (Martin et al. 2021). In the origin of life research, many opposing hypotheses still follow the same general principles because life has several basic traits that must be accounted for. In the works of this thesis, these basic traits have been used to describe seven basic steps for each origin of life hypothesis following a generalized framework.

In this framework, life first has to start with water. Each origin of life hypothesis can vastly differ in how they interpret this water source. For example, in hydrothermal vent hypotheses, the vent environment defines the proponents of the surroundings. Like in the hydrothermal vent hypothesis, this setting must also explain the second step: the energy source that will ultimately create organic components from simple molecules. After these steps, the accumulation or concentration of these molecules usually follows in the proposed system (the third step), which is then used as a basis for a complex molecular system (the fourth step). Once such a system exists, there has to be an explanation as to why it persists and does not fall back into equilibrium, but continues to evolve, the fifth step. If all of this is given, the first living systems can be created (step six), which emerges as LUCA with a lifestyle defined by how the previous steps are laid out. These basic steps show that complex problems can still be accumulated into a common framework. And depending on what evidence is found within this framework, certain hypotheses can be favored.

Another case with vastly opposing hypotheses is the question of how eukaryotes evolved. In this case, opposition exists regarding whether eukaryotes evolved before or due to the endosymbiosis event. For this, one expects to see specific patterns in eukaryotic genomes depending on when the endosymbiosis happened in relation to eukaryogenesis. The second publication of this thesis looks into this problem by explicitly investigating gene duplication patterns in eukaryotic genomes (Tria et al. 2021). For this analysis, 150 genomes were used as a basis for a protein sequence clustering from which over 160,000 duplication events could be identified. Most of these duplications occurred late in the evolution of eukaryotes. 713 duplications were possibly traced to the Last Eukaryotic Common Ancestor (LECA). These duplications were significantly more often from the bacterial origin and included more mitochondrial functions. Additionally, no lineage-specific gene acquisitions could be found, so vertical evolution is the norm. Both trace to support mitochondria-early instead of mitochondria-late hypotheses for eukaryogenesis.

Nevertheless, vertical inheritance is not the norm for the ancestral symbionts of eukaryotes: bacteria and archaea. Both domains are vastly affected by horizontal gene transfer events, but this tendency is unequal among all genes (Nagies et al. 2020). Reliable recovery of taxonomic phylogenies using specific gene sets shows the tendency for some genes to undergo more vertical evolution (Hansmann and Martin 2000). This allows for conceptualizing a measure that can describe horizontal gene transfer by instead specifically describing the verticality of genes. This measure has a simple basis: more vertical genes should show more monophyletic taxonomic groups. Verticality was the theme of the first publication of this thesis. The measure was applied to a prokaryotic sequence clustering of 5,655 prokaryotic genomes, in which 40 taxonomic groups were defined. For over 100,000 clusters, the verticality measure was calculated, which showed that universal genes have higher verticality. These vertical genes were often also those genes used in phylogenetic studies, for example, ribosomal genes (Hansmann and Martin 2000).

This also showed that most genes were transferred around, which means that all taxonomic groups were also affected by HGT but differed in the amount of HGT. This was the primary source

of genetic variability in prokaryotes as opposed to gene duplications that are much more common in eukaryotes (Tria and Martin 2021). Comparing the average verticality of taxonomic groups showed that very ecologically diverse groups and those known for HGT had lower verticality. In contrast, specialized groups tended to be more vertical. However, this generalization does not extend to all members of a group. For example, the Gammaproteobacteria were one of the least vertical groups but also included an intracellular symbiont species that showed the highest average verticality. Since all groups of prokaryotes are affected by gene transfers, this also extends to the phylogenetic patterns visible in eukaryotes. Some clusters among the 100,000 clusters with verticality value were linked to eukaryotic-prokaryotic clusters, calculated beforehand, while other clusters were directly related to genes of the mitochondrial genome of *Reclinomonas americana*, a protist in whose mitochondrial genome more genes than usual remain. These eukaryotic linked clusters showed a broad spectrum of verticality values. The ancestral symbionts that formed eukaryotes thus had genomes made up of genes that were likely already coming from different genomic sources that were also transferred more since eukaryogenesis (Martin 1999). The result is that only the most vertical genes linked to eukaryotes reliably recover Alphaproteobacteria and Cyanobacteria as the sister groups of those genes related to the mitochondrial or plastid ancestor, respectively. Due to the transfer events occurring, sister groups can be very variable for most other genes, so all groups in the dataset were recovered as sister groups to eukaryotes at least once (Nagies et al. 2020).

Further comparisons of verticality were possible on functional annotations. In the first paper of this thesis, a KEGG release of 2017 was used to compare the average verticality of functional annotations (Kaneshisa et al. 2016). This revealed that translation was the most vertical function. Other functional annotations linked to important cellular processes, e.g., nucleotide metabolism and replication, also showed higher verticality. On the other hand, some functions had relatively low verticality. These usually included functions known to be transferred among prokaryotes, for example, general metabolism genes, transporters, xenobiotic breakdown, and secondary metabolism. So, this shows that the function of a gene can influence its transfer rates. So, it could also be that specific enzymes have different transfer rates depending on what reactions they enable and what molecules participate in them. The dataset of clusters and associated verticality values allows testing this in the case of enzymes dependent or independent of oxygen.

Oxygen has a special place in Earth's history since life evolved initially in an oxygen-free environment (Martin and Sousa 2016, Sousa et al. 2016). However, the oxygen concentration rose suddenly during the Great Oxygenation Event after Cyanobacteria evolved oxygenic photosynthesis (Lyons et al. 2014). Since this event was 2.4 billion years ago, it is difficult to conceptualize the reasons and consequences of this sudden rise. But similarly to genomic patterns in eukaryotes reflecting their unique history, the GOE should leave traces in the evolution of prokaryotic genes. The verticality values were used in manuscript four to find these traces.

The clustering was first reannotated with a more recent version of KEGG from 2022 to attain a more thorough comparability of functional groups. This allowed to specifically assign clusters with one KEGG Orthology number (KO) each. These KO stand for a specific enzyme, which also had one to several reactions linked in KEGG. Then, based on the linked reactions, clusters could be sorted into enzymes that utilize oxygen and those that do not. There were, in total, 364 reactions that involved oxygen linked to the prokaryotic sequence clustering. Finally, the two sets of oxygen-dependent and oxygen-independent enzymes were compared in their verticality. This showed that oxygen-dependent enzymes had significantly lower verticality. Hence, there were more lineage breaks in those enzymes prompted by gene transfer events. One of the reasons for this could be that oxygen is a very high-energy molecule (Schmidt-Rohr 2015), so the higher energy release in reactions with oxygen could create a selective advantage. But surprisingly, verticality did not correlate with Gibbs free energy, so this cannot be assumed to be the reason for more gene transfers in an enzyme. Instead, it seems to be due to oxygen itself and the changes it allowed during and after the Great Oxygenation Event.

Oxygen is a strong oxidant (Schmidt-Rohr 2015). That means that this molecule makes it easier to break down compounds like aromatics, giving a selective advantage once oxygen-dependent enzymes are incorporated. In fact, many bacteria like Escherichia coli do not utilize the full energy potential of oxygen but instead use pathways of lower energy conversion (Unden and Bongaerts 1997). Rather, the main advantage lies in mobilizing nutrients bound in aromatic and nitrogenous compounds. This is reflected in the main compounds that oxygen occurs within the 364 oxygen-utilizing reactions, often aromatics es educts, and ammonia as a product. But another compound that is present in 80 reactions of this set is hydrogen peroxide. Hydrogen peroxide is usually seen as a waste product indirectly created in oxygen utilizing metabolism. But the proportion of reactions that produce hydrogen peroxide as a side product gave a need for the enzyme catalase, which metabolizes hydrogen peroxide and explains this enzyme's comparable low verticality. Thus, hydrogen peroxide is very ancient in metabolism but a consequence of metabolism utilizing oxygen.

However, oxygen also puts on constraint on metabolism, especially on the group that produced oxygen as a waste product: Cyanobacteria. Nitrogenase, needed for the primary production of Cyanobacteria, is inhibited by oxygen concentrations over 1% PAL (Allen et al. 2019) so that a feedback loop of primary production of Cyanobacteria creating oxygen molecules and oxygen, ultimately hindering further primary production is created. This put a limit on Earth's atmospheric oxygen concentration initially at around 1% PAL, creating the Boring Billion. This also showed additionally the value of oxygen mobilizing nitrogen compounds. Only later, with the evolution of land plants, was it possible that cellulase production increased so much that, with the increased rate of carbon burial, the atmospheric oxygen concentration was able to rise again until it reached present atmospheric levels. Thus, the oxygen concentration was not neatly controlled by geological factors but by three enzymes: the oxygen-evolving complex of cyanobacterial photosystem II, nitrogenase, and cellulose synthase.

Bibliography

Acar Kirit H, Lagator M, Bollback JP. 2020. Experimental determination of evolutionary barriers to horizontal gene transfer. BMC Microbiology 20: pmid:33115402.

Allen JF, Thake B, Martin WF. 2019. Nitrogenase Inhibition Limited Oxygenation of Earth's Proterozoic Atmosphere. Trends in Plant Science 24: 1022–1031.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

Andreani NA, Hesse E, Vos M. 2017. Prokaryote genome fluidity is dependent on effective population size. The ISME Journal 11: 1719–1721.

Arnold BJ, Huang IT, Hanage WP. 2021. Horizontal gene transfer and adaptive evolution in bacteria. Nature Reviews Microbiology 20: 206–218.

Bahir I, Fromer M, Prat Y, Linial M. 2009. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Molecular Systems Biology 5: 311-325.

Baquero F. 2015. Causes and interventions: need of a multiparametric analysis of microbial ecobiology. Environmental microbiology reports 7: 13–14.

Baquero F, Coque TM, Galán JC, Martinez JL. 2021. The Origin of Niches and Species in the Bacterial World. Frontiers in Microbiology 12: 657986.

Barraclough TG, Birky CW, Burt A. 2003. Diversification in sexual and asexual organisms. Evolution 57: 2166–2172.

Bartke K, Huseby DL, Brandis G, Hughes D. 2022. Evolution of Bacterial Interspecies Hybrids with Enlarged Chromosomes. Genome Biology and Evolution 14: evac135.

Beavan AJS, McInerney JO. 2022. Gene essentiality evolves across a pangenome. Nature Microbiology 7: 1510–1511.

Bendall ML, Stevens SLR, Chan LK, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, Froula J, Kang D, Tringe SG, Bertilsson S, Moran MA, Shade A, Newton RJ, McMahon KD, Malmstrom RR. 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J 10: 1589–601.

Berg JM. 1990. Zinc Finger Domains: Hypotheses and Current Knowledge. Annu. Rev. Biophys. Biophys. Chern 19: 405–426.

Bernardes JS, Vieira FRJ, Costa LMM, Zaverucha G. 2015. Evaluation and improvements of clustering algorithms for detecting remote homologous protein families. BMC Bioinformatics 16: 1–14.

Blomberg SP, Garland T. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. Journal of Evolutionary Biology 15: 899–910.

Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. 2019. The Ecology and Evolution of Pangenomes. Current Biology 29: R1094–R1103.

Brown JR. 2003. Ancient horizontal gene transfer. Nature Reviews Genetics 4: 121–132.

Brückner J, Martin WF. 2020. Bacterial Genes Outnumber Archaeal Genes in Eukaryotic Genomes. Genome Biology and Evolution 12: 282–292.

Brückner J. 2021. Analysis of early evolutionary events during the transition from prokaryotes to eukaryotes. Ph.D. dissertation. Heinrich Heine University, Düsseldorf.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10: 1–9.

Campbell N, Cain M, Minorsky P, Reece J, Urry L, Wasserman S. 2017. Biology. 11th ed. Pearson Education, Inc, London.

Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, Rooney AP, Yi H, Xu XW, de Meyer S, Trujillo ME. 2018. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. International Journal of Systematic and Evolutionary Microbiology. 68: 461–466.

Cohan FM. 2002. What are bacterial species? Annual Review of Microbiology. 56: 457–487.

Cohen O, Gophna U, Pupko T. 2011. The complexity hypothesis revisited: Connectivity Rather Than function constitutes a barrier to horizontal gene transfer. Molecular Biology and Evolution 28: 1481–1489.
Collins MD, East AK. 1998. Phylogeny and taxonomy of the food-borne pathogen Clostridium botulinum and its neurotoxins. Journal of applied microbiology 84: 5–17.
Conrad RE, Viver T, Gago JF, Hatt JK, Venter SN, Rossello-Mora R, Konstantinidis KT. 2021. Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations. The ISME Journal 2021 16:5 16: 1222–1234.
Dagan T, Roettger M, Bryant D, Martin W. 2010. Genome Networks Root the Tree of Life between Prokaryotic Domains. Genome Biology and Evolution 2: 379–392.
Darwin C. 1859. On the Origin of Species. 1st ed. John Murra: England, London.
Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure (ed. Dayhoff, M. O.) National Biomedical Research Foundation 345–352.
van Dongen S. 2000. Graph Clustering by Flow Simulation. Ph.D. Thesis. Utrecht University, Amsterdam.
Dykhuizen DE, Green L. 1991. Recombination in Escherichia coli and the definition of biological species. Journal of Bacteriology 173: 7257–7268.
Enright AJ, van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30: 1575–1584.
Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Systematic Biology 61: 717–726.
Fitch WM. 1970. Distinguishing Homologous from Analogous Proteins. Systematic Biology 19: 99–113.
Fox GE, Wisotzkey JD, Jurtshuk P. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. International Journal of Systematic Bacteriology. 42: 166–170.
Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. Science 323: 741–746.
Freudenstein J v., Broe MB, Folk RA, Sinn BT. 2017. Biodiversity and the Species Concept—Lineages are not Enough. Systematic Biology 66: 644–656.
Fuchsman CA, Collins RE, Rocap G, Brazelton WJ. 2017. Effect of the environment on horizontal gene transfer between bacteria and archaea. PeerJ 5: e3865.
Funahashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, Kitano H. 2008. CellDesigner 3.5: A versatile modeling tool for biochemical networks. Proceedings of the IEEE 96: 1254–1265.
Gao N, Lu G, Lercher MJ, Chen WH. 2017. Selection for energy efficiency drives strand-biased gene distribution in prokaryotes. Scientific Reports 7: 1–10.
Gehring WJ, Ikeo K. 1999. Pax 6: mastering eye morphogenesis and eye evolution. Trends in Genetics 15: 371–377.
Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, van de Peer Y, Vandamme P, Thompson FL, Swings J. 2005. Re-evaluating prokaryotic species. Nature Reviews Microbiology 3: 733–739.
Haigh J. 1978. The accumulation of deleterious genes in a population—Muller's Ratchet. Theoretical Population Biology 14: 251–267.
Hanage WP, Fraser C, Spratt BG. 2005. Fuzzy species among recombinogenic bacteria. BMC Biology 3: 1–7.
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. International Journal of Systematic and Evolutionary Microbiology 50: 1655–1663.
Hedlund BP, Chuvochina M, Hugenholtz P, Konstantinidis KT, Murray AE, Palmer M, Parks DH, Probst AJ, Reysenbach AL, Rodriguez-R LM, Rossello-Mora R, Sutcliffe IC, Venter SN, Whitman WB. 2022. SeqCode: a nomenclatural code for prokaryotes described from sequence data. Nature Microbiology 7: 1702–1708.
Hernández-Arriaga AM, Chan WT, Espinosa M, Díaz-Orejas R. 2014. Conditional activation of toxin-antitoxin systems: postsegregational killing and beyond. Pages 175-192 in. Plasmids: Biology and Impact in Biotechnology and Discovery, ASM Press: USA, Washington, DC.
Hey J. 2001. The mind of the species problem. Trends in Ecology & Evolution 16: 326–329.
Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications 9: 1–8.
Jiao X, Yang Z. 2021. Defining Species When There is Gene Flow. Systematic Biology 70: 108–119
Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research 44: D457–D462.
Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of Gene Sharing among 329 Proteobacterial Genomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths. Molecular Biology and Evolution 28: 1057–1074.
Kluge AG, Wolf AJ. 1993. Cladistics: What's in a word? Cladistics 9: 183–199.
Konstantinidis KT, Tiedje JM. 2005a. Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences of the United States of America 102: 2567–2572.
Konstantinidis KT, Tiedje JM. 2005b. Towards a genome-based taxonomy for prokaryotes. Journal of Bacteriology 187: 6258–6264.
Koonin E v, Wolf YI. 2009. Is evolution Darwinian or/and Lamarckian? Biology Direct 4: 1–14.
Kosoy M, Hayman DTS, Chan KS. 2012. Bartonella bacteria in nature: Where does population variability end and a species start? Infection, Genetics and Evolution 12: 894–904.
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.
Lang AS, Zhaxybayeva O, Beatty JT. 2012. Gene transfer agents: phage-like elements of genetic exchange. Nature Reviews Microbiology 10: 472–482.
Lercher MJ, Pál C. 2008. Integration of Horizontally Transferred Genes into Regulatory Interaction Networks Takes Many Million Years. Molecular Biology and Evolution 25: 559–567.
Lyons TW, Reinhard CT, Planavsky NJ. 2014. The rise of oxygen in Earth's early ocean and atmosphere. Nature 506: 307–315.
Macqueen J. 1967. Some methods for classification and analysis of multivariate observations. 5th Berkeley Symp. Math. Statist. Probability. Los Angeles LA USA: University of California.
Malmgren BA, Berggren WA, Lohmann GP. 1983. Evidence for punctuated gradualism in the Late Neogene Globorotalia tumida lineage of planktonic foraminifera. Paleobiology 9: 377–389.
Martin WF. 1999. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. Bioessays 21: 99–104.
Martin WF, Nagies FSP, do Nascimento Vieira A. 2021. To what inanimate matter are we most closely related and does the origin of life harbor meaning? Philosophies 6: 33-52.
Martin WF, Sousa FL. 2016. Early Microbial Evolution: The Age of Anaerobes. Cold Spring Harbor Perspectives in Biology 8: a018127.
Mayr E. 1996. What is a species, and what is not? Philosophy of Science 63: 262–277.

Mayr E. 1999. Systematics and the origin of species, from the viewpoint of a zoologist. Harvard University Press: USA, Cambridge.
McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes. Nature Microbiology 2: 1–5.
Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Current Opinion in Genetics & Development 15: 589–594.
Moreno-Hagelsieb G, Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24: 319–324.
Nagies FSP, Brueckner J, Tria FDK, Martin WF. 2020. A spectrum of verticality across genes. PLOS Genetics 16: e1009200.
Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 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. Proceedings of the National Academy of Sciences of the United States of America 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.
Niehus R, Mitri S, Fletcher AG, Foster KR. 2015. Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nature Communications 6: 1–9.
Nouioui I, Sangal V. 2022. Advanced prokaryotic systematics: the modern face of an ancient science. New Microbes and New Infections 49: 101036.
Ochman H, Lawrence JG, Grolsman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 2000 405: 299–304.
Ohta T, Gillespie JH. 1996. Development of Neutral and Nearly Neutral Theories. Theoretical Population Biology 49: 128–142.
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, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44: D733–D745.
Oren A, Garrity GM. 2021. Valid publication of the names of forty-two phyla of prokaryotes. International Journal of Systematic and Evolutionary Microbiology 71: 005056.
Owen R. 1848. On the archetype and homologies of the vertebrate skeleton. Murray: England, London.
Park C, Zhang J. 2012. High Expression Hampers Horizontal Gene Transfer. Genome Biology and Evolution 4: 523–532.
Parks DH, Rigato F, Vera-Wolf P, Krause L, Hugenholtz P, Tyson GW, Wood DLA. 2021. Evaluation of the Microba Community Profiler for Taxonomic Profiling of Metagenomic Datasets From the Human Gut Microbiome. Frontiers in Microbiology 12: 731.
Patterson C. 1988. Homology in classical and molecular biology. Molecular biology and evolution 5: 603–625.
Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, Whitman WB, Hallin S. 2010. The ecological coherence of high bacterial taxonomic ranks. Nature Reviews Microbiology 8: 523–529.

Ponting CP, Russell RR. 2003. The Natural History of Protein Domains. Annual Review of Biophysics and Biomolecular Structure 31: 45–71.
Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. Current Opinion in Microbiology 14: 615–623.
Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Research 21: 599–609.
de Queiroz K. 2005. Ernst Mayr and the modern concept of species. Proceedings of the National Academy of Sciences of the United States of America 102: 6600–6607.
de Queiroz K. 2007. Species concepts and species delimitation. Systematic Biology 56: 879–886.
Reydon TAC, Kunz W. 2019. Species as natural entities, instrumental units and ranked taxa: new perspectives on the grouping and ranking problems. Biological Journal of the Linnean Society 126: 623–636.
Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. 2021. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Nature Reviews Microbiology 19: 347–359.
Rosselló-Mora R, Amann R. 2001. The species concept for prokaryotes. FEMS Microbiology Reviews 25: 39–67.
Rost B. 1999. Twilight zone of protein sequence alignments. Protein Engineering, Design and Selection 12: 85–94.
Rost B. 2002. Enzyme function less conserved than anticipated. J Mol Biol 318: 595–608.
Schmidt-Rohr K. 2015. Why Combustions Are Always Exothermic, Yielding about 418 kJ per Mole of O2. Journal of Chemical Education 92: 2094–2099.
Simpson GG. 1951. The species concept. Evolution 5: 285–298.
Skippington E, Ragan MA. 2012. Phylogeny rather than ecology or lifestyle biases the construction of Escherichia coli–Shigella genetic exchange communities. Open Biology 2: 120112.
Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. Journal of Molecular Biology 147: 195–197.
Snel B, Bork P, Huynen M. 2000. Genome evolution: gene fusion versus gene fission. Trends in Genetics 16: 9–11.
Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet 16: 472–482.
Sousa FL, Nelson-Sathi S, Martin WF. 2016. One step beyond a ribosome: The ancient anaerobic core. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1857: 1027–1038.
Stevens PF. 1984. Homology and Phylogeny: Morphology and Systematics. Systematic Botany 9: 395-409.
Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: the bacterial pan-genome. Current Opinion in Microbiology 11: 472–477.
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 Biology and Evolution 13: evab055.
Tria FDK, Martin WF. 2021. Gene Duplications Are At Least 50 Times Less Frequent than Gene Transfers in Prokaryotic Genomes. Genome Biology and Evolution 13: evab224.
Unden G, Bongaerts J. 1997. Alternative respiratory pathways of Escherichia coli: energetics and transcriptional regulation in response to electron acceptors. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1320: 217–234.
Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. Current Opinion in Microbiology 23: 148–154.
Ward DM. 1998. A natural species concept for prokaryotes. Current Opinion in Microbiology 1: 271–277.

Wetlaufer DB, Rustow S. 1973. Acquisition of three-dimensional structure of proteins. Annual review of biochemistry 42: 135–158.
Woese CR. 1965. Order in the genetic code. Proceedings of the National Academy of Sciences of the United States of America 54: 71–75.
Wright ES, Baum DA. 2018. Exclusivity offers a sound yet practical species criterion for bacteria despite abundant gene flow. BMC Genomics 19: 1–12.
Yang R, Folk R, Zhang N, Gong X. 2019. Homoploid hybridization of plants in the Hengduan mountains region. Ecology and Evolution 9: 8399–8410.
Yu HY, Meade A, Liu SJ. 2019. Phylogeny of Clostridium spp. Based on Conservative Genes and Comparisons with Other Trees. Microbiology 88: 469–478.