Studien zur Rolle von Sauerstoff und Gentransfer in der prokaryotischen Genomevolution

Inaugural-Dissertation

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

vorgelegt von Katharina Trost
geboren in Dülmen

Düsseldorf, November 2025

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

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

Berichterstatter

I. Prof. Dr. William F. Martin
II. Prof. Dr. Sven B. Gould

Tag der mündlichen Prüfung: 05. März 2026

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass diese Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heiner-Universität Düsseldorf“ erstellt worden ist. Die Arbeit wurde bisher keiner Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht. Ich habe bisher keinen erfolglosen Promotionsversuch unternommen.

_____

Katharina Trost, Düsseldorf 2025

für

Annette & Andreas

Im Laufe dieser Arbeit wurden mit Zustimmung des Betreuers folgende Beiträge veröffentlicht:

Publikationen in Fachzeitschriften thematisiert in dieser Thesis

I. **Katharina Trost**, Michael R. Knopp, Jessica L. E. Wimmer, Fernando D. K. Tria, William F. Martin (2024). A universal and constant rate of gene content change traces pangenome flux to LUCA. *FEMS Microbiology Letters* 371, fnae068.

II. Natalia Mrnjavac, Falk S. P. Nagies, Jessica L. E. Wimmer, Nils Kapust, Michael R. Knopp, **Katharina Trost**, Luca Modjewski, Nico Bremer, Marek Mentel, Mauro Degli Esposti, Itzhak Mizrahi, John F. Allen, William F. Martin (2024). The radical impact of oxygen on prokaryotic evolution – enzyme inhibition first, uninhibited essential biosynthesis second, aerobic respiration third. *FEBS Letters* 598, 1692–1714.

III. **Katharina Trost**, Robert B. Gennis, John F. Allen, Dan B. Mills, William F. Martin (2026). Oxygen reductase origin followed the great oxidation event and terminated the Lomagundi excursion. *Biochimica et Biophysica Acta (BBA) – Bioenergetics* 1867, 149575.

Weitere Publikationen in Fachzeitschriften

IV. Lu Fan, Dingfeng Wu, Vadim Goremykin, **Katharina Trost**, Michael Knopp, Chuanlun Zhang, William F. Martin, Ruixin Zhu (2022). Reply to: Phylogenetic affiliation of mitochondria with Alpha-II and Rickettsiales is an artefact. *Nature Ecology & Evolution* 6, 1832–1835.

Inhaltsverzeichnis

1 Zusammenfassung

Im Gegensatz zu eukaryotischen Genomen, deren Genomvariation hauptsächlich aus Duplikationen und differenziellem Genverlust resultiert, wird die Evolution prokaryotischer Genome vor allem durch den Verlust von Genen sowie deren Gewinn über lateralen (horizontalen) Gentransfer geprägt. Diese Mechanismen erzeugen einen ständigen Genfluss zwischen prokaryotischen Genomen, der zur Ausbildung der Pangenomstruktur führt, welche aus einem konservierten Kerngenom besteht, das von einem variableren akzessorischen Genom umgeben ist. Im Laufe der Zeit führt dieser Genfluss zu Veränderungen im Genomrepertoire auf allen taxonomischen Ebenen.

Im Rahmen dieser Arbeit werden prokaryotische Genflussraten anhand von Sequenzdivergenz konservierter Gene und Genomdivergenz kultivierter als auch metagenomisch assemblierter Genome berechnet und verglichen. Es wird untersucht, ob eine konstante Genflussrate in prokaryotischen Genomen vorliegt und inwieweit diese Genflussrate während der gesamten Genomevolution bestehen blieb. Dabei wird deutlich, dass die langfristige, durchschnittliche Rate des Genflusses über höhere prokaryotische Taxa hinweg konstant ist, wohingegen die Größe des akzessorischen Genoms, der Anteil des Genoms, der Unterschiede im Gengehalt für Genompaare aufweist, variiert. Die Ergebnisse weisen darauf hin, dass die Pangenomstruktur seit der Divergenz von Bakterien und Archaeen ein allgemeines Merkmal prokaryotischer Genome ist und somit auf LUCA, den letzten universellen gemeinsamen Vorfahren, zurückzuführen ist.

Der kontinuierliche Genaustausch zwischen prokaryotischen Linien prägte nicht nur die Pangenomstruktur, sondern begünstigte auch die Verbreitung zentraler metabolischer Innovationen. Eines der wichtigsten Ereignisse, die eine starke Verbreitung metabolischer Innovationen zur Folge hatte, ist das Auftreten von molekularem Sauerstoff (engl. *Great Oxidation Event*, GOE) vor etwa 2,4 Milliarden Jahren, bei dem der atmosphärische Sauerstoffgehalt von 0 % auf 1 % der aktuellen atmosphärischen Konzentration (engl. *present atmospheric level*, PAL) anstieg.

Im zweiten Teil der Arbeit werden anhand von 365 O_2-abhängigen enzymatischen Reaktionen die wichtigsten physiologischen Anpassungen, die durch O_2-abhängige Enzyme bewirkt wurden, untersucht. Traditionell wird Sauerstoff mit Atmung und damit einhergehender Energiegewinnung assoziiert. Die Ergebnisse zeigen jedoch, dass die Resistenz gegen Sauerstofftoxizität, teils durch den Ersatz O_2-empfindlicher Enzyme durch neuartiger

O_2-abhängiger Enzyme, bereits vorhanden gewesen sein musste, bevor O_2 als Endakzeptor in die Atmungskette integriert werden konnte. Zellen mussten zunächst in der Lage sein in sauerstoffreichen Umgebungen zu überleben und erst dann konnten sie O_2 tatsächlich zur Steigerung der Energieeffizienz nutzen. Auch der Einfluss von LGT auf O_2-abhängige Enzyme im Vergleich zu O_2-unabhängigen Enzymen wurde untersucht. Die O_2-abhängigen Enzyme zeigen eine deutlich höhere Einwirkung von LGT als O_2-unabhängige Gene, was darauf hindeutet, dass sie den Organismen, die sie beibehielten einen physiologischen Vorteil verschafften.

Auch Sauerstoffreduktasen (*bd*-typ, HCO, AOX, PTOX), respiratorische Enzyme, die O_2 im terminalen Schritt der O_2-abhängigen Atmungskette zu Wasser reduzieren, gehören zu den O_2-abhängigen Enzymen welche stark von LGT betroffen sind. Im letzten Teil der Arbeit wird das Alter von Genen, die für Sauerstoffreduktasen codieren untersucht sowie deren Verbreitung über prokaryotische Abstammungslinien hinweg anhand eines zeitkalibrierten, phylogenetischen Baumes dargestellt. Die daraus resultierenden Daten deuten darauf hin, dass Cytochrom-*bd*-Oxidasen (*bd*-typ), Häm-Kupfer-Oxidasen (HCO) und alternative Oxidasen (AOX, PTOX) im Zuge des GOE vor etwa 2,4 Milliarden Jahren entstanden sind und infolgedessen erheblichen lateralen Gentransfer unterzogen wurden. Die Ergebnisse beleuchten die Physiologie im Zusammenhang mit dem GOE und decken ein biologisches Modell auf, das die bisher ungeklärte $\delta^{13}C$-Isotopenanomalie der Lomagundi-Jatuli-Exkursion (LJE) vor etwa 2,3 Milliarden Jahren als Produkt eines einzigen cyanobakteriellen Enzyms direkt erklären kann.

2 Abstract

Variation in gene content across eukaryotic genomes mainly results from gene duplications and differential gene loss. In contrast, the evolution of prokaryotic genomes is primarily driven by gene loss and the acquisition of new genes through lateral (horizontal) gene transfer. These mechanisms generate a continuous flux of prokaryotic genes, leading to the formation of a pangenome structure, composed of a conserved core genome surrounded by a more variable accessory genome. Over time, this gene flux progressively shapes the genomic gene repertoire at all taxonomic levels.

In this work, prokaryotic gene flux rates are calculated and compared based on sequence divergence of a conserved, universally distributed gene set and genome divergence of cultivated as well as metagenomically assembled genomes. Linear regression models were used to investigate whether a universal and constant rate of gene flux exists in prokaryotic genomes and to what extent this gene flux remained constant throughout prokaryotic genome evolution. The analysis revealed a constant long-term average rate of gene flux across higher prokaryotic taxa. However, the size of the accessory genome and the proportion of the genome differing in gene content between genome pairs varies between taxa. The results suggest that the pangenome structure has been a common feature of prokaryotic genomes since the divergence of the bacterial and archaeal lineages and can therefore be traced back to LUCA, the last universal common ancestor.

The continuous exchange of genes between prokaryotic lineages not only shapes the pangenome structure, but also promotes the spread of key metabolic innovations. One of the most important events that led to the widespread distribution of metabolic innovations was the Great Oxidation Event (GOE) approximately 2.4 billion years ago, during which atmospheric oxygen levels rose from 0% to 1% of the present atmospheric level (PAL).

In the second part of the thesis, the most important physiological adaptations caused by O_2-dependent enzymes are examined on the basis of 365 O_2-dependent enzymatic reactions. Traditionally, oxygen is associated with respiration and the associated energy production. However, the results show that resistance to oxygen toxicity, partly through the replacement of O_2-sensitive enzymes with novel O_2-dependent enzymes, must have already been present before O_2 could be integrated into the respiratory chain as the terminal acceptor. Cells first had to be able to survive in oxygen-rich environments before they could actually use O_2 to increase energy efficiency. The influence of LGT on O_2-dependent enzymes compared to O_2-

independent enzymes was also investigated. O_2-dependent enzymes show a significantly higher impact of LGT than O_2-independent genes, suggesting that they provided a physiological advantage to the organisms that retained them.

Oxygen reductases (*bd*-type, HCO, AOX, PTOX), respiratory enzymes that reduce O_2 to water in the terminal step of the O_2-dependent respiratory chain, are also among the O_2-dependent enzymes that are strongly affected by LGT. The last part of this thesis investigates the age of genes encoding oxygen reductases and their distribution in the course of the GOE using an independently generated time-calibrated phylogenetic tree. The resulting data suggest that cytochrome-*bd*-oxidases (*bd*-type), heme-copper oxidases (HCO), and alternative oxidases (AOX, PTOX) arose during the GOE about 2.4 billion years ago and have consequently undergone extensive lateral gene transfer. The findings shed light on microbial physiological adaptations surrounding the GOE and reveal a biological model that can directly account for the previously unexplained $\delta^{13}C$ isotope anomaly of the Lomagundi-Jatuli Excursion (LJE) 2.3 billion years ago as the product of a single cyanobacterial enzyme.

3 Einleitung

3.1 Variation prokaryotischer Genome und lateraler Gentransfer

Seit 1965, als die ersten Argumente für molekulare Phylogenien in den Vordergrund traten, sind Gensequenzen immer wichtiger für Studien der Prokaryotenevolution geworden (Zuckerkandl & Pauling 1965, Doolittle 1999). Es werden immer schnellere Methoden zur Sequenzierung von prokaryotischen Genomen entwickelt, wie zum Beispiel die Hochdurchsatz-Sequenzierungsmethoden (engl. *Next-Generation Sequencing*, NGS) Roche-454 Life Sciences (Margulies *et al.* 2005), Solexa-Illumina (Hu *et al.* 2015), oder ABI-SOLiD (McKernan *et al.* 2009). Aus der stetig wachsenden Anzahl an sequenzierten Genomen und großflächigen Sequenzvergleichen wurde zunehmend deutlich, dass die Varianz zwischen prokaryotischen Genomen derselben sowie unterschiedlicher Spezies viel höher war als zuvor angenommen (Pallen & Wren 2007). Zum einen weisen prokaryotische Genome große Unterschiede in ihrer Genomgröße auf, die von ca. 150 Kilobasenpaare (kb) in intrazellulären Endosymbionten (McCutcheon & Moran 2012) bis zu ungefähr 14,8 Megabasenpaare (Mb) in Myxobakterien (Han *et al.* 2013) reichen. Des Weiteren variieren sie stark in ihrem Genrepertoire (Perna *et al.* 2001), auch zwischen nah verwandten Stämmen. Ein Beispiel hierfür sind die *Escherichia coli* Stämme K-12 und O157:H7, welche sich bei vergleichbarer Genomgröße im Gengehalt um ca. 25 % unterscheiden (Hayashi *et al.* 2001).

Diese beobachteten Unterschiede zwischen prokaryotischen Genomen lassen sich vor allem auf drei Mechanismen zurückführen: Generwerb, Genverlust und Genveränderungen (Mira *et al.* 2001, Puigbò *et al.* 2014). Dabei tragen Einzel-Nukleotid-Polymorphismen (engl. *single-nucleotide polymorphisms*, SNPs), Duplikationen und Rekombinationen nur geringfügig zur Variation prokaryotischer Genome bei (Pallen & Wren 2007, Treangen & Rocha 2011, Tria & Martin 2021). Prozesse, die in Prokaryoten den größten Teil genomischer Variation ausmachen, sind Genverlust sowie der Austausch von Genen über Artgrenzen hinweg, auch als lateraler Gentransfer bezeichnet (LGT; Pallen & Wren 2007, Treangen & Rocha 2011). Im Gegensatz dazu haben Duplikationen beispielsweise einen größeren Einfluss auf die Variation eukaryotischer Genome, da eukaryotische Genome nicht durch lateralen Gentransfer nennenswert beeinflusst werden (Albalat & Cañestro 2016, Stull *et al.* 2021).

LGT umfasst den Transfer von Genen innerhalb von Arten sowie über Artgrenzen hinweg (Doolittle 1999, Koonin *et al.* 2001, Arnold *et al.* 2022). Dabei unterscheidet man zwischen drei klassischen Mechanismen: Transformation, Konjugation und Transduktion (Abe *et al.*

2020, Arnold *et al.* 2022). Bei der Transformation nehmen prokaryotische Zellen Erbgut direkt aus ihrer freien Umgebung auf, was dazu führen kann, dass Desoxyribonukleinsäure (DNA) zwischen stark divergenten Organismen ausgetauscht wird. Es unterstützt jedoch auch die Verbreitung von Genen innerhalb von Arten, da diese oft unter ähnlichen Umweltbedingungen leben (Dubnau 1999, Ochman *et al.* 2000). Wie bei der Transformation, ist es auch bei der Transduktion nicht erforderlich, dass Donor- und Akzeptororganismus gleichzeitig am selben Ort auftreten. Als Transduktion wird der Transfer von DNA über Bakteriophagen hinweg beschrieben. Wenn ein Bakteriophage innerhalb des Donororganismus DNA-Fragmente in das Phagenkapsid aufnimmt, können diese in das Erbgut eines anderen Wirtes des Bakteriophagen eingebaut werden. Dieser Prozess ist jedoch davon abhängig, dass die Wirtsorganismen spezifische Rezeptoren für den Bakteriophagen besitzen (Jiang & Paul 1998, Ochman *et al.* 2000). Anders als bei der Transformation und Transduktion müssen bei der Konjugation Donor- und Akzeptorzelle in Kontakt miteinander treten. Der unidirektionale Transfer von Plasmiden und Transposons wird über den Pilus bei Gram-negativen Bakterien oder oberflächenlokalisierten Proteinadhäsinen bei Gram-positiven Bakterien sowie einer Konjugationsbrücke ermöglicht (Ochman *et al.* 2000, Chen *et al.* 2005). Zusätzlich zu den klassischen drei Mechanismen von LGT wird prokaryotische DNA auch lateral über weitere nicht-kanonische Wege transferiert. Dazu gehören Gentransfer-Agenten, Nanoröhren (engl. *Nanotubes*) und Membran-Vesikel (Abe *et al.* 2020, Arnold *et al.* 2022). Gentransfer-Agenten sind Phagen-ähnliche DNA-Transporter, die von einer Donorzelle produziert werden und in die Umwelt freigelassen werden (Lang und Beatty 2007, Popa & Dagan 2011). Dieser Prozess ist der Transduktion sehr ähnlich. Jedoch sind Gentransfer-Agenten bisher nicht dafür bekannt, Gene über Artgrenzen hinweg zu transferieren. Auch die Länge der DNA-Sequenzen, die über Gentransfer-Agenten weitergegeben werden können ist mit ca. 4,4 – 14kb geringer als bei Bakteriophagen, die bis zu 100kb DNA aufnehmen können (Ochman *et al.* 2000; Lang und Beatty 2007). Nanoröhren sind Kanäle, die zwischen bakteriellen Zellen der gleichen sowie unterschiedlicher Spezies gebildet werden. Darüber können die Zellen kleine zytoplasmatische Moleküle, Proteine und nicht-konjugative Plasmide bidirektional austauschen (Dubey und Ben-Yehuda 2011). Ein weiterer nicht-kanonischer Mechanismus von LGT zwischen Prokaryoten sind Membran-Vesikel, von einer Lipid-Doppelschicht umschlossene Partikel, die DNA aus der Umwelt tragen und somit das Potenzial für LGT besitzen. Sie sind in Gewässern weit verbreitet und können Plasmide innerhalb einer Art sowie über Artgrenzen hinweg weitergeben (Abe *et al.* 2020).

Generell wird angenommen, dass bis zu 96 % aller Gene in einem prokaryotischen Genom mindestens einmal durch LGT transferiert wurden (Kunin & Ouzounis 2003, Dagan *et al.* 2008). Da die Genomgrößen prokaryotischer Organismen im Vergleich zu eukaryotischen Genomen nur begrenzt variieren, muss Generwerb, der aus LGT resultiert, durch balancierende Prozesse ausgeglichen werden. In prokaryotischen Genomen wird dies durch den Verlust von Genen gewährleistet, deren Funktion aufgrund früherer Mutationen in der Gensequenz inaktiviert wurde (Mira *et al.* 2001, Kunin & Ouzounis 2003). Zusätzlich können durch genetische Drift auch ganze Gruppierungen von Genen verloren gehen (Mira *et al.* 2001). Häufig zeigen kleinere prokaryotische Genome einen erhöhten Anteil an Genverlust auf, der bis zu zwei- bis dreimal höher ist als der Beitrag lateralen Gentransfers, wohingegen größere Genome tendenziell stärker durch LGT beeinflusst werden (Kunin & Ouzounis 2003, McCutcheon & Moran 2012, Puigbò *et al.* 2014).

Durch das Einwirken von LGT und Genverlust über geologische Zeiträume entsteht eine hohe Diversität in prokaryotischen Genomen, welche wichtig für die Anpassung an verschiedenste Lebensräume ist. Die Genome spiegeln ein Mosaik wider, dessen Elemente Gene darstellen, die ursprünglich von unterschiedlichen Spezies aus unterschiedlichen Abstammungslinien stammen und somit hochdynamisch sind (Lawrence & Ochman 1998, Martin 1999, Arnold *et al.* 2022). Die Diversität erschwert jedoch auch eine Einteilung prokaryotischer Organismen in Spezies (Doolittle 1999). Mehrere Wissenschaftler bevorzugen deswegen phylogenetische Netzwerke, um dynamische Beziehungen zwischen prokaryotischen Stämmen darzustellen. Im Gegensatz dazu setzen klassische verzweigte Bäume voraus, dass alle lebenden Arten von einer kleineren Anzahl an ursprünglichen Vorfahren, bis zurück zu dem ersten universellen Vorfahren, abstammen (Darwin 1860, Doolittle 1999, Martin 1999, Kunin *et al.* 2005, Huson & Bryant 2006).

3.2 Artbildung und das prokaryotische Pangenom

In der Zeit vor dem Einsatz von molekularen Sequenzen in der Phylogenie, basierte die Systematik aller Einzeller auf physiologischen und morphologischen Merkmalen (Whittaker 1969). Die sequenzbasierte Klassifikation prokaryotischer Genome in Spezies hat sich im Laufe der Zeit kontinuierlich weiterentwickelt. Zunächst wurde in den 1960ern die Methode der DNA-DNA Hybridisierung (DDH) entwickelt, bei der zwei Genome zu der gleichen Spezies gezählt werden, wenn mehr als 70 % DNA-DNA Übereinstimmung vorhanden ist und

ähnliche phänotypische Charakteristika nachweisbar sind (Wayne *et al.* 1987). Die DDH wurde später in den 80er Jahren standardisiert, jedoch ist die Methode, aufgrund der hohen Menge an DNA die benötigt wird, zeitaufwendig und arbeitsintensiv (Goris *et al.* 2007). Carl Woese und George Fox führten daraufhin 1977 die Nutzung von 16S ribosomalen Ribonukleinsäure-Genen (rRNA) in der Phylogenetik ein (Woese & Fox 1977). 16S rRNAs galten lange als universelle, vertikal vererbte Gene, die einen molekularen Uhrencharakter sowie ein starkes phylogenetisches Signal aufweisen (Woese 1987, Yarza *et al.* 2008, Boughner & Singh 2016). Jedoch haben Berichte über LGT- und Rekombinations-Ereignisse in 16S rRNA Zweifel an ihrer universellen Verwendung in der Spezies-Phylogenie genährt (Tourova *et al.* 2001, Kitahara & Miyazaki 2013, Jain *et al.* 2018). Auch die Nutzung von nur einem Gen als Repräsentant eines ganzen Genoms wurde hinterfragt (Konstantinidis & Tiedje 2004). Infolgedessen entstanden mit der Zeit Methoden, welche auf Sequenzvergleichen mehrerer Gene beruhen. Dazu zählt zum Beispiel die Multilocus-Sequenztypisierung, welche Unterschiede in der Nukleotidsequenz hauswirtschaftlicher (engl. *housekeeping*) Gene nutzt, um Mikroorganismen zu gruppieren (Maiden *et al.* 1998). Eine weitere Alternative für sequenzbasierte Gruppierungen von Genomen in Spezies ist die durchschnittliche Nukleotid-Identität (engl. *average nucleotide identity*, ANI) konservierter Gene. Im Vergleich zu vorherigen Methoden wie der 16S rRNA-Sequenzdivergenz oder der DDH zeigt sich, dass die ANI, besonders auf der Speziesebene, Ähnlichkeiten und Unterschiede zwischen Genomen detaillierter aufdecken kann (Konstantinidis & Tiedje 2004, Goris *et al.* 2007). Die anhand von ANI definierten Gruppierungen von Genomen (ca. 95 % ANI) scheinen aber nicht immer zur selben Spezies zu gehören. Bisher genannte Methoden liefern somit zwar erste Ergebnisse für die Bestimmung prokaryotischer Spezies, sind jedoch in keinem Fall fehlerfrei. Dadurch blieb die Suche nach einem Ansatz der Speziesdefinierung, der flexibler ist und die ökologischen Besonderheiten der Spezies einbindet, zunächst erfolglos (Konstantinidis & Tiedje 2004).

Aufgrund der stetig wachsenden Anzahl sequenzierter Genome wurde eine immer größere intraspezifische Diversität erfasst (Pallen & Wren 2007, Konstantinidis & Tiedje 2004). Es stellte sich die Frage, wie viele Sequenzen innerhalb einer prokaryotischen Spezies benötigt werden, um sie überhaupt akkurat darstellen zu können (Tettelin *et al.* 2005, Tettelin *et al.* 2008). Einige Wissenschaftler folgerten, dass eine Einteilung in Cluster von Genomen eine realitätsnähere Gruppierung darstellt. Dazu gehörte auch die Gruppe um Hervé Tettelin, die 2005 das Pangenom als eine Darstellung von Genomunterschieden und -gemeinsamkeiten einer Spezies einführte (Tettelin *et al.* 2005). Ein Pangenom lässt sich in das Kerngenom (engl. *core genome*), das Gene umfasst, die in allen untersuchten Genomen vorkommen, und in das

akzessorische Genom (engl. *accessory genome*), in dem alle restlichen Gene zusammengefasst sind, einteilen (Tettelin *et al.* 2005, Tettelin *et al.* 2008). Später wurde auch eine Einteilung in vier Gruppen vorgeschlagen, bei der das Pangenom in das Kern-, weiche Kern- (engl. *soft core*), akzessorische und Wolkengenom (engl. *cloud*) eingeteilt wurde. Das weiche Kerngenom umfasst Gene, die in mindestens 95 % aller Genome vorkommen, jedoch nicht universell sind. Das akzessorische Genom hingegen beinhaltet Gene, die in mehr als 1 % der Genome vorkommen, aber in weniger als 95 % und im Wolkengenom befinden sich alle Proteinfamilien, die genomspezifisch sind (Sonnenberg *et al.* 2020, Matthews *et al.* 2024). Generell wird das Kerngenom oft als die Essenz der Spezies bezeichnet, weil sich hauptsächlich Gene, die hauswirtschaftliche oder regulative Funktionen für den Organismus haben, darin finden lassen (Tettelin *et al.* 2005, Medini *et al.* 2005). Diese Gene sind im Allgemeinen weniger durch LGT beeinflusst, wodurch die Größe des Kerngenoms gleich bleibt, unabhängig davon wie viele Genome dem Pangenom hinzugefügt werden. Im akzessorischen Genom hingegen codiert ein Großteil der Gene für mobile Elemente, Antibiotikaresistenzen oder speziesspezifische Funktionen (Tettelin *et al.* 2005, Tettelin *et al.* 2008, Vernikos *et al.* 2015). Sie bilden ein Reservoir an Funktion und unterliegen häufig LGT, um neue Eigenschaften zur Anpassung an neue Nischen zu erlangen (Tettelin *et al.* 2005, Tettelin *et al.* 2008, Vernikos *et al.* 2015, Segerman 2012). Das akzessorische Genom spiegelt somit die Vielfalt der Spezies wider und kann mit steigender Anzahl an Genomen schneller oder langsamer anwachsen (Medini *et al.* 2005). Dieses Wachstum kann anhand der Anzahl genomspezifischer Gene, die mit jedem weiteren Genom hinzugefügt werden, gemessen werden und ermöglicht eine Klassifikation des Pangenoms als offen oder geschlossen. Ein offenes Pangenom zeichnet sich dadurch aus, dass die Anzahl an genomspezifischen Genen wächst, je mehr Genome dem Pangenom hinzugefügt werden. Spezies, die ein offenes Pangenom aufweisen, sind oft an wechselnden Umweltbedingungen angepasst und offener für Gentransfer (Medini *et al.* 2005). Um das Genrepertoire einer Spezies mit einem offenem Pangenom darstellen zu können, würden somit eine sehr hohe Anzahl an Sequenzen benötigt werden (Tettelin *et al.* 2008, Vernikos *et al.* 2015). Hogg *et al.* (2007) erweiterte die Theorie des offenen Pangenoms durch ein Modell, das anhand der Genverteilung über verschiedene Genome hinweg die finale Anzahl der Gene im Pangenom schätzt, die benötigt wird, um eine Spezies akkurat darstellen zu können.

Wenn die Anzahl an neuen Proteinfamilien jedoch gegen null tendiert, je mehr Genome hinzugefügt werden, wird das Pangenom als geschlossen bezeichnet. Das lässt darauf schließen, dass das gesamte Genrepertoire der Spezies charakterisiert wurde (Tettelin *et al.* 2005, Medini *et al.* 2005, Vernikos *et al.* 2015). Spezies, die in isolierten Nischen leben und

konserviert sind, weisen eher ein geschlossen Pangenom auf, da sie durch ihre Lebensweise einen eingeschränkten Zugriff auf den globalen, mikrobiellen Genpool haben (Medini *et al.* 2005).

Im Vergleich zu früheren Methoden zur prokaryotischen Speziesdefinierung, erlaubt das Pangenom-Konzept eine flexiblere Einteilung, da es sowohl quantifizierbare, genomische Unterschiede als auch Gemeinsamkeiten einer Spezies aufzeigt und genetische sowie ökologische Einflüsse in die Klassifizierung von Arten mit einbezieht.

3.3 Metagenome

Trotz der starken Zunahme an neu sequenzierten Genomen, die durch Methoden wie NGS in Datenbanken hinterlegt werden konnten, wird davon ausgegangen, dass nur ein minimaler Anteil aller prokaryotischer Organismen bis heute sequenziert ist (Rappé & Giovannoni 2003, Segerman 2012, Lok 2015). Der Hauptgrund hierfür ist, dass viele Organismen nicht oder nur in Verbindung mit aufwändigen und teuren Laborbedingungen kultivierbar sind, da ihre natürlichen Lebensräume umweltbezogene, physikalische, biochemische und genetische Komplexitäten aufweisen (Garza & Dutilh 2015). Diese Organismen werden oft auch unter dem Begriff „mikrobielle dunkle Materie“ (engl. *microbial dark matter*) zusammengefasst (Marcy *et al.* 2007, Rinke *et al.* 2013, Lok 2015).

Aufgrund dessen werden metagenomisch assemblierte Genome (engl. *metagenomic-assembled genomes*, MAGs) mit Hilfe von bioinformatischen Algorithmen erstellt. Zunächst wird dazu jegliche DNA aus Umweltproben sequenziert, woraus eine Probe von einzelnen DNA-Stücken unterschiedlichster Organismen resultiert, das sogenannte Metagenom (Garza & Dutilh 2015, Setubal 2021). Die einzelnen DNA-Stücke werden dann anhand von ähnlichen Eigenschaften (z.B. Tetranukleotid-Frequenzen, komplementäre Markergene, taxonomischen Alignments und Codonverwendung), so gruppiert, dass sie demselben Organismus angehören, um die ursprüngliche Genomsequenz zu rekonstruieren (Garza & Dutilh 2015, Yang *et al.* 2021).

Die aktuell verfügbaren Algorithmen zur Rekonstruktion von MAGs sind jedoch noch nicht ausgereift, wodurch sich erhebliche Unterschiede in den Qualitäten der einzelnen MAGs ergeben (Mardis *et al.* 2002, Chain *et al.* 2009). Um qualitativ unverlässliche MAGs von Studien ausschließen zu können, wurden bioinformatische Programme entwickelt, welche die Qualität von MAGs anhand ihrer Vollständigkeit und Kontamination schätzen. Vollständigkeit

beschreibt die Menge von taxonspezifischen oder -unspezifischen Markerproteinen, welche in einem MAG erwartet werden. Kontamination bezieht sich auf den Anteil der Sequenzen, die nicht zum Zielorganismus gehören (Manni *et al.* 2021, Parks *et al.* 2015). Zu diesen Programmen zählen BUSCO (Benchmarking Universal single-Copy Orthologue tool; Simao *et al.*, 2015), CheckM (Parks *et al.* 2015) und CheckM2 (Chlovski *et al.* 2023). Oft wird die Qualität eines MAGs anhand einer Kombination aus beiden Qualitätsparametern durch die folgende Gleichung (Parks *et al.* 2017) bestimmt:

$$\text{Qualität} = \text{Vollständigkeit} - 5 * \text{Kontamination}$$

Ein Threshold von 50 % Qualität wird des Öfteren als geeignet angesehen, um ein MAG als hoch qualifiziert und vertraulich einzustufen (Parks *et al.* 2018, Pasolli *et al.* 2019, Nayfach *et al.* 2021, Almeida *et al.* 2021). Jedoch ist bei der Interpretation von MAGs Vorsicht geboten, da niedrige Qualitätsschwellenwerte das Risiko fehlerhafter Ergebnisse in wissenschaftlichen Analysen erhöhen können.

3.4 Sauerstoffevolution

Die Veränderung von Pangenomen über geologische Zeitspannen beruht auf mehreren evolutionären als auch ökologischen Mechanismen. Dazu gehören zum einen LGT, ökologische Nischenvielfalt, aber auch selektiver Druck (McInerney *et al.* 2017, Touchon *et al.* 2020). Umweltereignisse, welche einen starken selektiven Druck ausüben, können zu einer erhöhten Menge an Genfixierung in Organismen führen und begünstigen damit die Verbreitung zentraler metabolischer Innovationen über prokaryotische Abstammungslinien hinweg. Dazu zählt auch das Große Sauerstoffereignis vor ca. 2,4 Milliarden Jahren, bei dem molekularer Sauerstoff (O_2) das erste Mal in der Erdatmosphäre auftrat (Holland 2002, Gumsley *et al.* 2017). Cyanobakterien, die maßgeblich zur Produktion von O_2 während des GOE beitrugen, nutzen zwei auf Chlorophyll basierende Photosysteme (PSI & PSII) um Kohlenstoffdioxid (CO_2) und Stickstoff (N_2) mit Hilfe von Elektronen zu fixieren, die sie aus Wasser (H_2O) extrahiert haben. Dabei entsteht Sauerstoff im Photosystem II, der schließlich zur Erhöhung der atmosphärischen O_2-Konzentration führte (Shen 2015, Kato *et al.* 2021, Fischer *et al.* 2016, Demoulin *et al.* 2024). Einige Wissenschaftler argumentieren jedoch, dass aufgrund von Spurenmetallanreicherungen, die auf oxidative Verwitterung von terrestrischen Sulfiden

hinweisen, bereits sogenannte Hauche (engl. *whiffs*), kurzzeitige Spuren von Sauerstoffkonzentration, circa 50 bis 600 Millionen Jahre vor dem Großen Sauerstoff Ereignis in der Atmosphäre auftraten (Anbar *et al.* 2007, Czaja *et al.* 2012, Crowe *et al.* 2013). Es wurden auch zeitlich begrenzte Schwefelisotopensignale oder Quecksilbervorkommen genutzt, um frühe Veränderungen des atmosphärischen O_2-Gehalts zu bestimmen (Kaufman *et al.* 2007, Meixnerová *et al.* 2021). Diese Hypothesen werden jedoch kontrovers diskutiert und neuere Studien nehmen an, dass die kurzzeitigen Signale für O_2-Anreicherungen vor dem GOE erklärbar sind durch die Oxidation von 2,45 Milliarden alten Sedimentproben, die bereits unter anaeroben Bedingungen abgelagert wurden (Slotznick *et al.* 2022). Eine weitere Hypothese für Sauerstoffkonzentrationen vor dem GOE besagt, dass O_2 durch abgeriebenen Quarz in Küstenzonen synthetisiert wurde (He *et al.* 2021, He *et al.* 2023, Stone *et al.* 2022). Jedoch entsteht dabei eine hohe Menge an Wasserstoffperoxid (H_2O_2), welches durch seine hohe Reaktivität nicht zu einer signifikanten Steigerung der O_2-Konzentration in der Atmosphäre beitragen könnte (Koppenol & Sies 2024, Mrnjavac *et al.* 2024a, Mrnjavac *et al.* 2024b). Auch Mangan-Knollen, die auf dem Meeresboden zu finden sind, sollen Sauerstoff synthetisiert haben (Sweetman *et al.* 2024). Es ist aber fragwürdig, ob Mangan-Knollen O_2 vor dem GOE produziert haben könnten, da sie sich nur mit Hilfe von O_2 formen und anreichern. Somit kam molekularer Sauerstoff mit hoher Wahrscheinlichkeit das erste Mal mit dem GOE in die Atmosphäre und stieg auf circa 1 % des aktuellen Sauerstoffgehalts an (engl. *present atmospheric level*, PAL; Fischer *et al.* 2016, Slotznick *et al.* 2022).

Diese atmosphärische Sauerstoffkonzentration blieb von vor circa 2,3 Milliarden bis vor circa 580 Millionen Jahren mehr oder weniger konstant (Lyons *et al.* 2014, Mills *et al.* 2022, Brocks *et al.* 2023). Da es in dieser Periode an geobiologisch interessanten Ereignissen mangelt, wird sie als eine der „langweiligsten" Phasen in der Erdgeschichte beschrieben und deswegen auch die langweilige Milliarde genannt (engl. *Boring Billion*; Mukherjee *et al.* 2018). Der Begriff Pasteurische Ära wird auch in diesem Kontext genutzt, da die Sauerstoffkonzentration von 1 % PAL dem Pasteur-Punkt gleicht, welcher die Sauerstoffkonzentration angibt, bei dem fakultativ aerobe Organismen von anaerober zu aerober Atmung wechseln (Martin *et al.* 2020). Weshalb die Sauerstoffkonzentration mehr oder weniger konstant bei 1 % PAL blieb wird bis heute diskutiert. Es gibt mehrere Theorien, die auf geologische oder geochemische Prozesse zurückzuführen sind (Canfield 1998, Anbar & Knoll 2002, Poulton *et al.* 2004, Alcott *et al.* 2019, Klatt *et al.* 2021). Diese Ansätze limitieren jedoch nur die Rate der O_2-Anhäufung und bestimmen nicht ihren genauen Endwert (Allen *et al.* 2019). Es gibt jedoch auch einen biologischen Ansatz basierend auf dem Enzym

Nitrogenase, welches die Reduktion von N_2 zu Ammoniak (NH_3) katalysiert (Allen *et al.* 2019). Nitrogenasen werden durch O_2 inhibiert: Bei 1 % O_2 wird die Nitrogenase-Aktivität bis zu 41 % inhibiert und bei 10 % O_2 bis zu 100 % (Stewart & Lex 1970). Sinkt die Nitrogenase-Aktivität, sinkt auch das Vorkommen von fixiertem Stickstoff, welches notwendig für das Wachstum von Stickstoff-fixierenden Cyanobakterien ist, wodurch die O_2-Produktion gehemmt wird. Sobald die O_2-Konzentration unter 2 % sinkt, nimmt die Nitrogenase-Aktivität sowie die O_2-Produktion wieder zu. Durch eine negative Feedbackschleife kann somit die O_2-Konzentration in der Atmosphäre durch ein einziges cyanobakterielles Enzym, die Nitrogenase, über zwei Milliarden Jahre hinweg reguliert werden (Allen *et al.* 2019, Mrnjavac *et al.* 2024a, Mrnjavac *et al.* 2024b). Heute besitzen viele Cyanobakterien Mechanismen, um Nitrogenasen vor O_2 zu schützen. Es gibt jedoch Indizien, dass diese Schutzmechanismen erst spät, nach dem Ursprung von Landpflanzen in Cyanobakterien, entwickelt wurden (Mrnjavac *et al.* 2024b). Sie wären somit nicht verantwortlich für den Sauerstoffanstieg am Ende des Proterozoikums, sondern haben sich eher als Schutzmechanismus für bereits angestiegene, hohe Sauerstoffkonzentrationen entwickelt (Allen *et al.* 2019, Mrnjavac *et al.* 2024b).

Vor circa 500 Millionen Jahren, mit dem Ende der langweiligen Milliarde, stieg die Sauerstoffkonzentration in der Atmosphäre auf aktuelle 21 % an, was 100 % PAL entspricht. Der Mechanismus, durch den die erhöhte Sauerstoffkonzentration entstand, ist ein verstärktes Aufkommen an Kohlenstofffixierung mit dem Ursprung der Landpflanzen (Lenton *et al.* 2016, Stolper & Keller 2018). Das dazu notwendige Enzym ist Zellulose Synthase (Mrnjavac *et al.* 2024b). Zellulose ist hauptsächlich in den Zellwänden von Stämmen und Blättern der Landpflanzen zu finden (Pedersen *et al.* 2023). Es wird synthetisiert, indem zunächst Elektronen von H_2O genutzt werden, um O_2 in der photosynthetischen Elektronentransportkette zu generieren. Diese Elektronen werden dann zur CO_2-Fixierung in den Calvin Zyklus weitergeleitet, wodurch phosphorylierte Zucker entstehen. Daraufhin können Glucose-Monomere synthetisiert werden, welche in polymerisierter Form das stickstofffreie Polymer Zellulose bilden (Mrnjavac *et al.* 2024b). Da Landpflanzen Photosynthese in Luftorganen, hauptsächlich in Blättern, ausführen, welche aufgrund von N_2-fixierenden Mikroben im Boden physisch von ihrer Stickstoffressource getrennt sind, limitieren Nitrogenasen nicht mehr die O_2-Produktion (Allen *et al.* 2019, Mrnjavac *et al.* 2024b).

Die Sauerstoffkonzentration in der Erdatmosphäre kann somit durch drei cyanobakterielle Enzyme erklärt werden: dem Sauerstoff produzierenden Komplex des

Photosystems II, Nitrogenasen, welche durch O_2 inhibiert werden, und Zellulose Synthase in Landpflanzen.

3.5 Lomagundi-Jatuli Isotopenexkursion

Am Ende des GOE zeigt sich in geochemischen Aufzeichnungen die größte, positive ^{13}Kohlenstoff-Isotopenexkursion der letzten 3,5 Milliarden Jahre, auch Lomagundi-Jatuli Exkursion (LJE) genannt (Schidlowski *et al.* 1976, Melezhik *et al.* 2005). Sie fand vor circa 2,3 bis 2,1 Milliarden Jahren statt, dauerte 100 bis 250 Millionen Jahre an und weist ^{13}C Isotopen Werte (δ^{13}C) zwischen +5 ‰ und +10 ‰ auf (Schidlowski *et al.* 1976, Karhu & Holland 1996, Martin *et al.* 2013). Die LJE ist die bemerkenswerteste Ausnahme von δ^{13}C-Werten und bis heute nicht komplett erklärbar, da sich Wissenschaftler uneinig sind, wie die hohen δ^{13}C-Werte begründet werden können. Jedoch werden zwei Ansätze präferiert: (i) die LJE beruht auf mehreren lokalen Ereignissen in Küsten- und Flachwassergebieten bei der Kohlenstoffabbau, Sedimentflüsse, Evaporation und Methanogenese eine Rolle gespielt haben könnten (Frauensteine *et al.* 2009, Prave *et al.* 2022) und (ii) die LJE war ein globales Ereignis, welches synchron mit dem Sauerstoffanstieg in der Atmosphäre, mit Start des GOE, ablief (Karhu & Holland 1996, Gumsley *et al.* 2017). Vertreter letzter Theorie nehmen an, dass starke vulkanische Eruptionen zu erhöhten oxidativen Verwitterungen von Landmassen führten und somit zu einem steigenden Fluss von Phosphor und anderen essenziellen Nährstoffen. Dadurch stieg die photosynthetische Aktivität und damit einhergehend der atmosphärische Sauerstoff während des GOE (Holland 2002, Bekker & Holland 2012, Gumsley *et al.* 2017, Hodgskiss *et al.* 2019). Hier spielt das Enzym Ribulose-1,5-bisphosphat-carboxylase/-oxygenase (RuBisCo) eine entscheidende Rolle. Es fixiert bevorzugt ^{12}Kohlenstoff (^{12}C) aus der Atmosphäre im Photosystem II der Cyanobakterien (Hayes 1993). Dies führt zu einer Anreicherung von ^{13}C in der Atmosphäre sowie von ^{12}C in organischem Material (Hodgskiss *et al.* 2023). Der Kohlenstoff-Zyklus des LJE und die Konzentration des atmosphärischen Sauerstoffs stehen somit in direkter Verbindung zueinander. Es wird oft davon ausgegangen, dass diese erhöhte Einlagerung von leichtem organischem Material (^{12}C) zu einem Anstieg der Sauerstoffkonzentration geführt haben müsste (Karhu & Holland 1996, Bekker *et al.* 2004, Lyons *et al.* 2014). Die während des LJE beobachteten hohen δ^{13}C-Werte würden jedoch unter Berücksichtigung von Standardsauerstoffmodellen bedeuten, dass der atmosphärische Sauerstoff von 0 % vor dem GOE bis zu über 21 % (v/v) angestiegen wäre, was mehr als dem

heutigen atmosphärischen Sauerstoffgehalt entspricht (Karhu & Holland 1996). Gründe dies anzuzweifeln lassen die Frage unbeantwortet wodurch die hohen $\delta^{13}C$-Werte während des LJE entstanden sind und warum sie 100 bis 250 Millionen Jahre später mit Eintreten der sogenannten langweiligen Milliarde wieder fielen (Prave *et al.* 2022).

3.6 Sauerstoffreduktasen

Heutzutage wird der meiste Sauerstoff, welcher durch Cyanobakterien und Pflanzen produziert wird, für die Atmung und Energiekonservierung genutzt, was die atmosphärischen O_2-Konzentrationen konstant hält (Li *et al.* 2021). Am Ende der Atmungskette katalysieren Sauerstoffreduktasen (terminale Oxidasen) die Reduktion von Sauerstoff (O_2) zu Wasser (H_2O; Wikstrom 1977). Zu ihren bekanntesten Haupttypen gehören *bd*-Typ Oxidasen (*bd*; Borisov *et al.* 2011, Degli Esposti *et al.* 2019, Murali *et al.* 2021), Häm-Kupfer Oxidasen (HCO; Pereira *et al.* 2001, Sousa *et al.* 2012, Murali *et al.* 2022), die mitochondriale alternative Oxidase (AOX; Atteia *et al.* 2004, Pennisi *et al.* 2016) und die plastochinol Oxidase in Plastiden (PTOX; Kuntz 2004, McDonald & Vanlerberghe 2005).

Cytochrom *bd*-Typ Oxidasen sind reine Chinol-Oxidasen, die entweder Ubichinol oder Menachinol als Substrat nutzen. Sie besitzen drei Häme, darunter b_{558}, b_{595} und *d* (Marreiros *et al.* 2016), von denen das Häm b_{595} und das Häm *d* der O_2-Reduzierung dienen. Zu ihren Funktionen gehören zum einen die Energiekonservierung in Form eines Protonantriebs, zum anderen erfüllen sie zahlreiche weitere Funktionen, etwa indem sie Organismen die Besiedlung O_2-armer Umgebungen erleichtern oder als O_2-Abbauenzym wirken, um eine Inhibition von O_2-sensiblen Enzymen, wie Nitrogenasen, zu verhindern (Borisov *et al.* 2011, Degli-Esposti *et al.* 2019, Murali *et al.* 2021).

Häm-Kupfer Oxidasen (HCO) sind die wohl am meisten studierten Sauerstoffreduktasen. Sie besitzen eine zweikernige O_2-Reduktionsstelle, welche ein High-Spin-Häm und ein Kupferion beinhaltet (Marreiros *et al.* 2016). Die HCO generieren einen Protonenantrieb, der für verschiedenste biosynthetische Aktivitäten (z.B. ATP-Synthese), mechanische Bewegung (z.B. Flagellenrotation) oder Transport von gelösten Stoffen von zentraler Bedeutung ist. Zu ihrer Familie gehören sowohl Cytochrom C Oxidasen, Chinol-Oxidasen (Pereira *et al.* 2001, Borisov *et al.* 2011) als auch die Stickstoffmonoxid (NO) terminalen Oxidasen, welche evolutionär von dem Vorfahren der Sauerstoffreduktasen abstammen (Marreiros *et al.* 2016, Borisov *et al.* 2011, Pennisi *et al.* 2016; Murali *et al.* 2021,

Murali *et al.* 2022, Murali *et al.* 2024). Sowohl *bd*-Typ als auch Häm-Kuper Oxidasen sind stark von LGT beeinflusst und weit über prokaryotische Taxa hinweg verbreitet. *Bd*-typ Oxidasen finden sich vermehrt unter fakultativ anaeroben und mikroaeroben Organismen, wohingegen Häm-Kupfer Oxidasen sowohl in anaeroben als auch in aeroben Organismen vorkommen (Pereira *et al.* 2001, Borisov *et al.* 2011, Sousa *et al.* 2012, Borisov *et al.* 2015, Pennisi *et al.* 2016, Soo *et al.* 2019, Degli Esposti *et al.* 2019, Murali *et al.* 2021, Murali *et al.* 2022).

Die alternative Oxidase ist Cyanid-resistent und benutzt Ubichinon als Elektronendonor. Sie besitzt ein aktives Zentrum für die O_2-Reduktion auf der Basis von Eisencarboxylaten und produziert im Gegensatz zu HCO und *bd*-Typ Oxidasen keinen Protonenantrieb. AOX sind wichtig für die Hitzegenerierung in bestimmten Oberflächen, sie spielen eine Rolle in der Regulation des Energiestoffwechsels sowie im Schutz vor oxidativen Stress und bei der Aufrechterhaltung der Homöostase und Redox-Balance (Atteia *et al.* 2004, Pennisi *et al.* 2016). Außerdem sollen sie die Erzeugung reaktiver Sauerstoff Spezies (engl. *reactive oxygen species*, ROS) reduzieren und somit die übermäßige Verringerung des Ubichinon-Vorrats verhindern (Maxwell *et al.* 1999). Vertreten sind AOX in Mitochondrien von Pflanzen, Pilzen und Protisten, aber auch in wenigen Bakterien und Tieren (Pereira *et al.* 2001, Atteia *et al.* 2004, McDonald & Vanlerberghe 2005, Borisov *et al.* 2011, Pennisi *et al.* 2016).

Die mit den AOX verwandte plastoquinol terminale Oxidase (PTOX) weist ebenfalls ein aktives Zentrum auf der Basis von Eisencarboxylaten auf, welches eine ähnliche sekundäre Struktur wie das aktive Zentrum von AOX hat (Berthold & Stenmark 2003). PTOX ist auch Cyanid-resistent und kommt am Ende der photosynthetischen Elektronentransportkette vor (Kuntz 2004). Dementsprechend tritt PTOX nur in photosynthetischen Organismen, wie zum Beispiel Pflanzen, Algen, Diatomen und Cyanobakterien, auf (Atteia *et al.* 2004, McDonald & Vanlerberghe 2004, McDonald & Vanlerberghe 2005, McDonald *et al.* 2011). Ähnlich wie AOX hat PTOX die Funktion, die übermäßige Verringerung des Plastochinon-Vorrats zu verhindern und die Redox-Balance aufrechtzuerhalten. Außerdem wirkt PTOX als Komponente in der Entsättigung von Carotinoiden (Kuntz 2004, McDonald & Vanlerberghe 2005).

Bis heute ist nicht geklärt, wann genau Sauerstoffreduktasen entstanden sind. Basierend auf ihrer Sauerstoffaffinität wurde jedoch eine gewisse Reihenfolge vorgeschlagen. *Bd*-Typ terminal Oxidasen sollen aufgrund ihrer hohen Sauerstoffaffinität der erste bekannte Typ der Sauerstoffreduktasen gewesen sein. Sie können auch sauerstoffarme Umgebungen besiedeln,

wohingegen alle anderen drei Typen (HCO, AOX und PTOX) eine sehr niedrige Sauerstoffaffinität aufweisen und dementsprechend Umweltbedingungen mit hohen Sauerstoffkonzentrationen benötigen (Sharma & Wikström 2014, Degli Esposti *et al.* 2019). Darüber hinaus ist es möglich, dass die Funktion von Sauerstoffreduktasen zu Beginn nicht bioenergetischer Form war, sondern der Beseitigung von überschüssigem, toxisch wirkendem Sauerstoff diente (Degli Esposti *et al.* 2019). Außerdem wird angenommen, dass HCO von Bakterien über LGT auf Archaeen übertragen wurden und somit nicht ihren Ursprung im letzten gemeinsamen Vorfahren (engl. *last universal common ancestor*, LUCA) haben (Pereira *et al.* 2001, Degli Esposti *et al.* 2019). Die alternative Oxidase (AOX) sowie die plastochinol terminal Oxidase (PTOX) wurden wahrscheinlich vom Vorfahren der Mitochondrien und Chloroplasten vertikal vererbt (Atteia *et al.* 2004) und könnten somit ihren Ursprung in Alphaproteobakterien und Cyanobakterien haben (McDonald & Vanlerberghe 2005). Für alle vier Typen gilt jedoch, dass sie mit hoher Wahrscheinlichkeit entstanden sind, nachdem Sauerstoff bereits in der Atmosphäre war, da sie ohne das Substrat O_2 keine selektierbare Funktion gehabt haben können.

4 Zielsetzung

Prokaryoten können ihre Gene nicht nur vertikal an ihre Nachkommen weitervererben, sondern auch lateral über Artgrenzen hinweg austauschen und transferieren. Über die Zeit hinweg betrachtet, entsteht durch LGT und Genverlust ein Prozess des Genflusses in prokaryotischen Genomen. Dieser Genfluss beeinträchtigt jedoch die Kategorisierung von Genomen in Spezies (Doolittle 1999), da durch die Mechanismen des Genflusses ein hoher Grad an Diversität in prokaryotischen Genomen zu finden ist. Heute ist bekannt, dass prokaryotische Spezies in Pangenomen organisiert sind, welche ein Kerngenom mit essenziellen Genen, die in allen Genomen einer Spezies vorkommen, und ein akzessorisches Genom, welches metabolische Gene oder Gene für die Anpassung an bestimmte Umweltbedingungen beinhaltet, besitzen (Tettelin *et al.* 2005).

Vor diesem Hintergrund ist es ein Ziel dieser Arbeit, den Einfluss von Gentransfer, insbesondere LGT, auf die Genomevolution von Prokaryoten zu untersuchen. Mit Hilfe von Sequenzdivergenz vertikal vererbter und universell verbreiteter Gene sowie An- oder Abwesenheit von Proteinfamilien in prokaryotischen Genomen werden Genflussraten für verschiedene prokaryotische Taxa ermittelt. Es wird analysiert, ob die berechneten Genflussraten spezifisch für einzelne Taxa und taxonomische Ebenen oder universell vertreten sind. Außerdem wird diskutiert, inwieweit sich durch Genflussraten Rückschlüsse auf vorhandene Spezieskonzepte, wie das Pangenom, ziehen lassen.

Im weiteren Verlauf der Arbeit wird näher auf den Einfluss von Sauerstoff auf die Genom- sowie die biochemische Evolution von Prokaryoten eingegangen. Durch die Generierung von Sauerstoff (O_2) in Cyanobakterien stieg der Gehalt an O_2 vor ca. 2,4 Milliarden Jahren in der Atmosphäre an (Holland 2002, Kato *et al.* 2021). Es stellt sich die Frage, welche evolutionären Auswirkungen Sauerstoff in Prokaryoten neben Atmung und Energiegewinnung hat und inwieweit LGT sich in der Verbreitung von O_2-abhängigen Enzymen bemerkbar macht. Dazu wurden Cluster und phylogenetische Bäume von O_2-abhängigen und O_2-unabhängigen Enzyme miteinander verglichen und analysiert.

Traditionell wird Atmung und Energiegewinnung als die größte evolutionäre Auswirkung von Sauerstoff gesehen. Terminale Oxidasen katalysieren die Reduktion von Sauerstoff zu Wasser (H_2O) am Ende der Atmungskette (Wikstrom 1977). Der genaue Zeitpunkt der Entstehung von terminalen Oxidasen ist jedoch unklar. Im letzten Teil der Arbeit wird der Ursprung terminaler Oxidasen untersucht, indem die Verteilung von *bd*-, HCO und

alternativen terminalen Oxidasen (AOX, PTOX) über prokaryotische Taxa auf einen zeitlich kalibrierten phylogenetischen Baum abgebildet wird (Mahendrarajah *et al.* 2023). Anhand dieser Analyse soll ein Modell erstellt werden, welches die Physiologie rund um das GOE sowie Gründe für die Entstehung und die Auflösung des LJE, der größten, positiven ^{13}Kohlenstoff-Isotopenexkursion in den letzten 3,5 Milliarden Jahren, veranschaulicht (Schidlowski *et al.* 1976).

5 Publikationen

I A universal and constant rate of gene content change traces pangenome flux to LUCA

Katharina Trost, Michael R. Knopp, Jessica L. E. Wimmer, Fernando D. K. Tria, William F. Martin (2024).

Institut für Molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Deutschland.

Dieser Artikel wurde am 20.08.2024 in *FEMS Microbiology Letters* Ausgabe 371 veröffentlicht.

Beitrag von Katharina Trost (Erstautor und Korrespondenz):
Ich habe anhand der bereits erstellten Proteinfamilien die Messwerte für Genomdivergenz und Sequenzdivergenz in den universellsten und vertikalsten Proteinfamilien berechnet. Anhand dessen habe ich, bis auf die funktionale Annotation der Proteinfamilien, alle Analysen durchgeführt. Die Abbildungen wurden von mir erstellt, ich habe das initiale Manuskript geschrieben und war an der Überarbeitung beteiligt.

FEMS Microbiology Letters, 2024, 371, fnae068
DOI: 10.1093/femsle/fnae068
Advance access publication date: 20 August 2024
Research Article – Taxonomy, Systematics & Evolutionary Microbiology

A universal and constant rate of gene content change traces pangenome flux to LUCA

Katharina Trost[1,*], Michael R. Knopp[1], Jessica L.E. Wimmer[1], Fernando D.K. Tria[1,2], William F. Martin[1]

[1] Faculty of Mathematics and Natural Sciences, Institute of Molecular Evolution, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany
[2] Los Alamos National Laboratory, Los Alamos, NM, United States
*Corresponding author. Faculty of Mathematics and Natural Sciences, Institute of Molecular Evolution, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. E-mail: katharina.trost@hhu.de
Editor: [Aharon Oren]

Abstract

Prokaryotic genomes constantly undergo gene flux via lateral gene transfer, generating a pangenome structure consisting of a conserved core genome surrounded by a more variable accessory genome shell. Over time, flux generates change in genome content. Here, we measure and compare the rate of genome flux for 5655 prokaryotic genomes as a function of amino acid sequence divergence in 36 universally distributed proteins of the informational core (IC). We find a clock of gene content change. The long-term average rate of gene content flux is remarkably constant across all higher prokaryotic taxa sampled, whereby the size of the accessory genome—the proportion of the genome harboring gene content difference for genome pairs—varies across taxa. The proportion of species-level accessory genes per genome, varies from 0% (Chlamydia) to 30%–33% (Alphaproteobacteria, Gammaproteobacteria, and Clostridia). A clock-like rate of gene content change across all prokaryotic taxa sampled suggest that pangenome structure is a general feature of prokaryotic genomes and that it has been in existence since the divergence of bacteria and archaea.

Keywords: gene flux; pangenomes; prokaryotes; accessory genome; core genome; metagenomes

Introduction

The evolution of genome diversification in eukaryotes is mostly driven by gene duplication and differential loss (Albalat and Cañestro 2016, Stull et al. 2021). In contrast, prokaryotic genome evolution is driven mostly by gene loss and gene acquisition via lateral (or horizontal) transfer (LGT), while gene duplication is rare (Mira et al. 2001, Treangen and Rocha 2011, Tria and Martin 2021). Genetic recombination and gene transfer in prokaryotes involves unidirectional transfer of genes from donors to recipients via transformation, transduction, conjugation, gene transfer agents, or membrane vesicles (Arnold et al. 2021). Over time, these mechanisms generate a process of DNA flux through constant gene loss and gain in prokaryotic chromosomes. How much gene flux they generate and whether such flux has been in operation throughout evolutionary history are questions of interest. In early work using codon bias as a proxy for laterally acquired genes, Lawrence and Ochman (1998) estimated that about 18% of the genes in *Escherichia coli* MG1655 genome correspond to recent lateral acquisitions. Today, it is recognized that the gene content of prokaryotic species is typically organized as pangenomes (Tettelin et al. 2005), with a conserved core genome consisting of genes present in all genomes of a given taxonomic sample, such as strains, and an accessory genome consisting of genes that are differentially present in all sample members that is, in evolutionary terms, in a state of continuous flux (Medini et al. 2005, Tettelin et al. 2005, 2008, Vernikos et al. 2015, Brockhurst et al. 2019).

The rate of gene flux between genomes of prokaryotes has been extensively studied at the species and genus level and several studies uncovered a clear relationship between phylogenetic distance and the frequency of gene differences that may arise by gain or loss (gain/loss). For example, Hao and Golding (2006) examined the association between the rate of gene-flux and point mutations in the core genome of seven *Bacillus cereus* strains; they estimated a rate of 4.4 gene gain or loss per nucleotide substitution per site in the strains' core genome. Applications of the same method yielded similar inferences of 1.17 and 1.18 gene gain/loss events per point mutation per site for 12 *Streptococcus* genomes (Marri et al. 2006) and five *Corynebacterium* genomes (Marri et al. 2007), respectively. Higher rates of gene gain/loss were inferred for 27 *Pseudomonas syringae* strains, where a comparison of gene gain/loss among closely related strains showed that up to 5000 gain/loss events may have occurred before 1% amino acid sequence divergence in the core genomes was reached (Nowell et al. 2014). Other studies reported a positive association between sequence divergence of core genes and divergence in gene content (Wolf et al. 2016). For example, the more distant the genomes of two *E. coli* strains are, the fewer genes they share, although the relationship was weak (Touchon et al. 2009, Rocha 2018, Haudiquet et al. 2022). Similarly, a study of 22 *Myxococcus xanthus* strains showed that the number of gene differences is increased with phylogenetic distance as inferred from amino acid differences in the core genome (Wielgoss et al. 2016). Since the estimation of diverged gene content may be biased by differences in genome size, the "genome fluidity" metric was proposed as an unbiased estimate (Kislyuk et al. 2011), which is positively associated with the core genome sequence divergence at synonymous sites (Andreani et al. 2017).

Received 22 March 2024; **revised** 15 May 2024; **accepted** 19 August 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of FEMS. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

While these studies arrived at similar conclusions regarding the rate of gene flux and phylogenetic relatedness, a direct comparison of the estimated rates across studies and taxa is challenging for several reasons. First, the size and the composition of the core genome varies strongly depending on the pangenome taxonomic composition (Vernikos et al. 2015), such that estimates of core genome divergence vary across studies. Second, the criteria for quantifying the number of shared genes between genomes (or differences in shared genes) can be very similar or even identical across studies, but the gene sets used to estimate sequence divergence are not. As a consequence, most estimates for rates of gene flux relative to sequence divergence are not comparable across species, samples, taxonomic levels, or studies.

Here, we ask whether prokaryotic genomes harbor evidence for a general correlation between the rate of gene flux into and out of genomes as a function of sequence divergence. For this purpose, we compared sequence divergence in a universal set of 36 proteins that are present in almost all genomes across higher taxa and are sufficiently conserved to be useful for comparisons at the deepest taxonomic levels (Hansmann and Martin 2000, Charlebois and Doolittle 2004, Dagan and Martin 2006). We then tested for an association between sequence divergence in these core genes and gene content divergence in the complete genomes of 5655 taxonomically diverse isolates as well as in 2872 metagenomic assemblies (MAGs).

Methods

Prokaryotic dataset

The prokaryotic clustering set was used from Brueckner and Martin (2020) including 5655 prokaryotic genomes, 19 050 992 protein sequences from the Reference Sequence database (RefSeq), September 2016 from the National Center for Biotechnology Information (NCBI; O'Leary et al. 2016). The clustering was created using the Markov Cluster Algorithm (MCL; van Dongen 2008) as previously described (Brueckner and Martin 2020, Nagies et al. 2020). In total, 450 283 protein families were detected and for protein families with at least four protein sequences multiple alignments were made with Mafft L-INS-I version 7.130 (Katoh 2002).

Due to large sample numbers, Proteobacteria and Firmicutes were divided into classes. Archaea were divided into orders to allow multiple groups to be assessed. The resulting 59 prokaryotic taxa comprise 41 bacterial and 18 archaeal groups that are called higher taxa in the following (Table S1).

The dataset including MAGs was obtained from Garg et al. (2021) including 103 assemblies from NCBI BioProject PRJN270657 and 2546 assemblies from BioProject PRJNA288027 downloaded in 2018. Additionally, 223 assemblies form the Microbial dark matter project were added (Rinke et al. 2013). The dataset was clustered using the following pipeline: to search for local alignments, an all versus all blastp was conducted using diamond version 2.0.11 (Buchfink et al. 2015). Reciprocal best blast hits (Wolf and Koonin 2012) with an e-value ≤1E-10 were aligned globally with the Needleman–Wunsch Algorithm (Emboss Needle version 6.6.0.0; Rice et al. 2000). All global alignments with an Identity ≥25% were used for clustering into protein families, using MCL (van Dongen 2008) version 14–137 with pruning parameters -P 180000, -S 19800, and -R 25200 (P = Pruning, S = Selection, and R = Recovery). In total, 285 787 protein families were detected. The completeness and contamination of all MAGs was measured by using CheckM v1.2.1 (Parks et al. 2015). Two out of the 2872 MAGs were not included in the clustering.

Protein family annotation via KEGG

For protein family annotation, all clustered sequences from the RefSeq dataset 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 1E-10 were used for annotation. Based on these hits, a KO and a name were assigned to the clusters based on majority rule. Protein families, which contained equal to or more than 75% of unknown sequences, were not annotated to preserve the strict nature of the annotation method.

Verticality values

Verticality is a measure for how often members of a given protein family tend to recover monophyly of prokaryotic phyla in their respective gene trees (Nagies et al. 2020). Verticality values used in this analysis were obtained from the study by Nagies et al. (2020), based on the same dataset used in this study.

Determination of IC gene set

For the 260 972 prokaryotic protein families of the RefSeq dataset with a determined verticality value, a weight was calculated for the number of genomes, the number of higher taxa as well as the verticality. To obtain only one value per protein family, the average weight was calculated. To avoid taking genes that are particularly universal but not sufficiently vertical, or vice versa, the 50 protein families with the best average weights were compared to the 100 most vertical, the 100 most universal protein families based on number of genomes and to the 100 protein families that are most widely distributed across all higher taxa. Protein families present in all three lists and among the lowest 50 weights were assigned as informational core (IC) genes (Table S2).

The corresponding metagenomic IC clusters were obtained by using the reciprocal best cluster approach described in Ku et al. (2015): if 50% of all sequences of a prokaryotic cluster have their best hit in another cluster and if in this cluster also 50% of all sequences have their best hit in the prokaryotic cluster, it can be defined as a reciprocal best cluster. In the dataset, 36 IC genes were found. For every IC gene a multiple sequence alignment was calculated using Mafft L-INS-I version 7.505 (Katoh 2002).

Calculation of sequence divergence in the IC gene set (ICD)

For all 15 986 685 prokaryotic genome comparisons in the RefSeq dataset and for 4 100 275 genome pairs in the metagenomic dataset, the average sequence divergence in the IC gene set was calculated. In every genome comparison, for each IC gene (x), the proportion of different sites (ICD_x) in the multiple sequence alignment (a, $length_a$ = *number of sites in the multiple sequence alignment*) was calculated with the following formula:

$$ICD_x = \frac{diff\_sites_a}{length_a}.$$

The average proportion of different sites of all IC genes (n) is then determined as IC gene divergence ICD per genome comparison.

$$ICD = \frac{\sum_{x=1}^{n} ICD_x}{n}.$$

If a genome was not present in all 36 IC genes, the average divergence (ICD) was calculated for all IC genes in which the genome is represented by at least one protein.

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

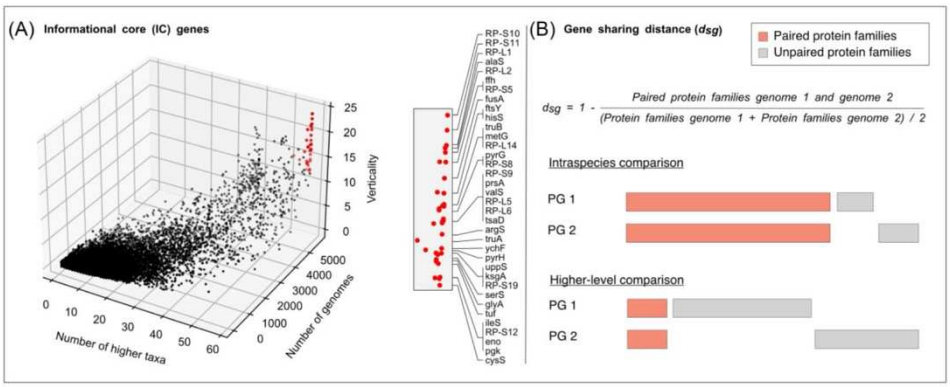


Figure 1. Comparison of number of higher taxa, number of genomes, and verticality for the IC genes (A) and calculation of gene sharing distance d_{sg} (B). (A) The number of higher taxa (x-axis) is plotted in relation to the number of genomes (z-axis) and verticality (y-axis) for the 36 defined IC genes (red) and for all other protein families from Nagies et al. (2020) (black). Higher taxa include all phyla present in the dataset except of Proteobacteria and Firmicutes, which were divided into classes due to large sample sizes and Archaea that were divided into orders to allow multiple groups to be assessed. Gene names for the 36 IC genes are shown on the right part of (A) and are sorted by verticality (B). The gene sharing distance (d_{sg}) was calculated by assigning protein families from the compared genomes into two groups. Protein families present in both genomes are defined as paired protein families, the others as unpaired protein families. The gene sharing distance (d_{sg}), which represents differences in gene content, was then calculated by subtracting the proportion of paired protein families from one. We expect that genomes belonging to the same species (intraspecies comparisons) have a higher proportion of paired protein families than genomes belonging to different higher-levels and thus show a lower gene sharing distance (PG = protein families of genome).

Calculation of gene sharing distance (d_{sg})

For all 15 986 685 prokaryotic genome comparisons in the RefSeq dataset and for 4 100 756 genome pairs in the metagenomic dataset, the difference in gene content was calculated by comparing the protein families corresponding to the compared genomes (Fig. 1B). Two groups were defined, the paired protein families and the unpaired protein families. Paired protein families are protein families in which both genomes were represented by at least one protein. If only one genome is present in the protein family, the protein family corresponds to the group of unpaired protein families. The gene sharing distance d_{sg} is defined as one minus the number of paired protein families divided by the average number of all protein families in the compared genomes.

Filtering for genome size and strains

To avoid overplotting, the original prokaryotic RefSeq dataset was filtered for genomes with genome sizes within average higher taxa genome size plus/minus one standard deviation. Additionally, to reduce phylogenetic bias, only one random strain per species was retained in the dataset. The resulting dataset comprises 1630 strains (Table S1).

Calculation of average nucleotide identity values

The average nucleotide identities (ANI) were calculated using FastANI v1.34 (Jain et al. 2018).

Calculation of rRNA sequence divergence

For every species represented by at least 10 strains in the prokaryotic RefSeq dataset, rRNA sequences were downloaded from the RefSeq database of NCBI. A multiple alignment for all sequences was made with Mafft L-INS-I version 7.505 (Katoh 2002). Equal sites were then counted and the proportion of equal sites from all sites was subtracted from one to obtain the rRNA sequence divergence per genome pair. Values for d_{sg} calculated from the prokaryotic dataset were assigned by using a random strain per species.

Statistical tests

To analyse the relationship between the ICD and d_{sg} across prokaryotic taxa, Spearman rank correlations and linear regression lines were calculated. The regression lines were evaluated doing an analysis of residuals, including residual plots, mean square residuals, QQ plots of residuals, and the Kolmogorov–Smirnov test to test whether residuals are normally distributed. The Spearman rank correlation and the linear regression were also applied for the relationship between species-level accessory genome proportion and y-axis intercept as well as for the relationship between ICD and verticality of paired and unpaired protein families. To compare the distributions of paired and unpaired protein families, a paired t-test was applied. The cloud gene analysis was evaluated using a one-way ANOVA. All statistical tests were performed using python.

Calculation of average species level accessory genome proportion per higher taxon

The average species level accessory genome proportion per higher taxon are calculated based on intraspecies genome comparisons. For every species, the average gene sharing distance (d_{sg}) of intraspecies genome comparisons was calculated. The mean of all species level accessory genome proportions corresponding to a specific higher taxa represents the average species level accessory proportion per higher taxon. The analysis was made for all higher taxa present in the filtered dataset, that are represented by at least 10 strains, 2 species, and with a P-value lower or equal to .05 from the Spearman correlations between ICD and d_{sg}.

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

Calculation of absolute gene flux rates

Absolute gene flux rates were calculated by use of the relative gene flux rates, which show the percentage of gene differences in the genome per one % ICD. The absolute number of substitutions in IC genes then correspond to one % of the average number of sites present in the IC gene alignments. Gene differences at one % ICD were then calculated by multiplying the relative gene flux rate with the average number of genes present in the higher taxa. Then the number of gene differences at one substitution in the IC genes could be calculated by dividing the number of gene differences at one % ICD by the number of substitutions at one % ICD. The analysis was made for all higher taxa present in the filtered dataset, that are represented by at least 10 strains and with a P-value lower or equal to .05 from the Spearman correlations between ICD and d_{sg}.

Eukaryotic organelle dataset

From the RefSeq Database (NCBI) 207 plastid and 68 mitochondrial proteins were downloaded in October 2022 for the species *Porphyra umbilicalis* and *Reclinomonas americana* as well as 189 089 nuclear protein sequences for the higher eukaryotic group Discoba. Additionally, for 150 eukaryotic genomes, 3 420 731 protein sequences were downloaded from the RefSeq Database, January 2018 (NCBI; O'Leary et al. 2016). The nuclear genome of *P. umbilicalis* was downloaded from the universal protein knowledgebase (UniProt Consortium 2021) in October 2022, including 13 559 protein sequences. To obtain mitochondrial homologous for prokaryotic IC genes, alphaproteobacterial IC gene sequences were searched in mitochondrial protein sequences of model organism *R. americana* by using diamond 2.0.1 (Buchfink et al. 2015). After filtering for identities higher or equal to 25%, a maximum e-value of 1E-10 and a search for best hits, 11 mitochondrial homologous for IC genes were defined based on majority rule. The remaining genes were searched in the Discoba proteome (24 IC genes) and the eukaryotic genomes (25 IC genes). The same search was made for cyanobacterial IC genes in plastid proteins of *P. umbilicalis* (14 IC genes) and in nuclear *P. umbilicalis* proteins (19 IC genes) as well as in eukaryotic genomes (22 IC genes). This resulted in four groups that represent the IC genes in mitochondria or plastids combined from organelle and nuclear protein sequences.

Informational core gene divergence between organelle genes and corresponding alphaproteobacterial or cyanobacterial genes

For the four datasets of mitochondrial or plastid universal genes, multiple alignments were made for every IC gene, using Mafft L-INS-I v7.505 (Katoh 2002) including the organelle (mitochondrial or plastid) protein sequence and the alphaproteobacterial or cyanobacterial protein sequences corresponding to the IC gene family. Average IC gene divergence (ICD) were then calculated between the organelle sequences and the corresponding prokaryotic sequences as described above (see the section "Calculation of sequence divergence in the IC gene set (ICD)").

Results and discussion

The IC is vertically inherited

For the estimation of sequence divergence, we selected genes that have a nearly universal distribution and a low degree of LGT , that is, a high level of verticality (Nagies et al. 2020). These proportions are plotted for all protein families including at least four genomes from two or more higher taxa, shown in Fig. 1(A). Higher taxa include all phyla present in the dataset except for Proteobacteria and Firmicutes, which were divided into classes due to large sample sizes and Archaea that were divided into orders to allow multiple groups to be assessed. The 36 genes (red) that are widely distributed across higher prokaryotic taxa and genomes and exhibit high levels of verticality are mostly involved in information processing (Rivera et al. 1998), hence we call this set the informational core (IC) (Table S2). On average, IC genes display a verticality of 17.09 and are present in over 5585 prokaryotic genomes and 58 higher taxa. Verticality is a measure for how often members of a given protein family tend to recover monophyly of prokaryotic phyla in their respective gene trees (Nagies et al. 2020). In the data set of Nagies et al. (2020), verticality values can vary between 0 and 42, however, the highest calculated verticality value was 24.0 for 30S ribosomal protein S10 (RP-S10). By contrast, most of the other non-IC genes (black) are distributed across few genomes (mean number of genomes = 66, mean number of higher taxa = 2) and are inherited much less vertically (mean = 0.15). Functional annotations of the IC gene set using KEGG (Kanehisa et al. 2017) showed that the IC, selected here on the basis of universal distribution and verticality, comprises mainly genes encoding for ribosomal proteins, in addition to amino acid tRNA synthetases, translation elongation factors, RNA modifications enzymes, and several genes involved in metabolism (enolase, 3-phophoglycerate kinase, and pyrimidine synthesis enzymes). From the multiple amino acid sequence alignments of the 36 IC genes, we used the proportion of amino acid differences in pairwise comparisons as a robust measure for evolutionary divergence between genome pairs, termed here informational core gene divergence (ICD). The functional distribution of IC genes as well as their tendency to be vertically inherited is in line with previous studies about the properties of universally conserved (core) genes (Tettelin et al. 2005).

To quantify gene content differences for each genome pair, we determined the proportion of paired protein families (Fig. 1B, light red), which is the number of protein families in our clustered sample that is present in both of the compared genomes divided by the average number of protein families present in the two genomes. The proportion of paired genes, subtracted from one, yields an estimate for gene sharing differences (d_{sg}) between prokaryotic genome pairs.

A clock-like rate of gene content change

Plots of ICD versus gene sharing distance (d_{sg}) for pairwise genome comparisons reveal a highly significant positive correlation between evolutionary divergence and gene content differences (Fig. 2, left-hand panels; Fig. S1). Genome comparisons with similar IC gene sequences show little difference in gene content, whereas genome pairs with a high proportion of substitutions in amino acid sequences in pairwise IC gene comparisons show large differences in gene content. The plots for the raw data containing all genomes per higher taxon are shown in the left-hand panels of Fig. 2.

In Alphaproteobacteria and Gammaproteobacteria, some genome comparisons have low gene content differences at high ICD, which is caused by the presence of many genomes of endosymbionts having highly reduced genomes (Moran and Bennett 2014, Wernegreen 2015) in these higher taxa (Fig. 2B and F). To mitigate the influence of reduced endosymbiont genomes, we generated data sets containing only genomes whose genome size was within the average higher taxon genome size ± 1 SD. Furthermore, to avoid the effect of oversampling and phylogenetic

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

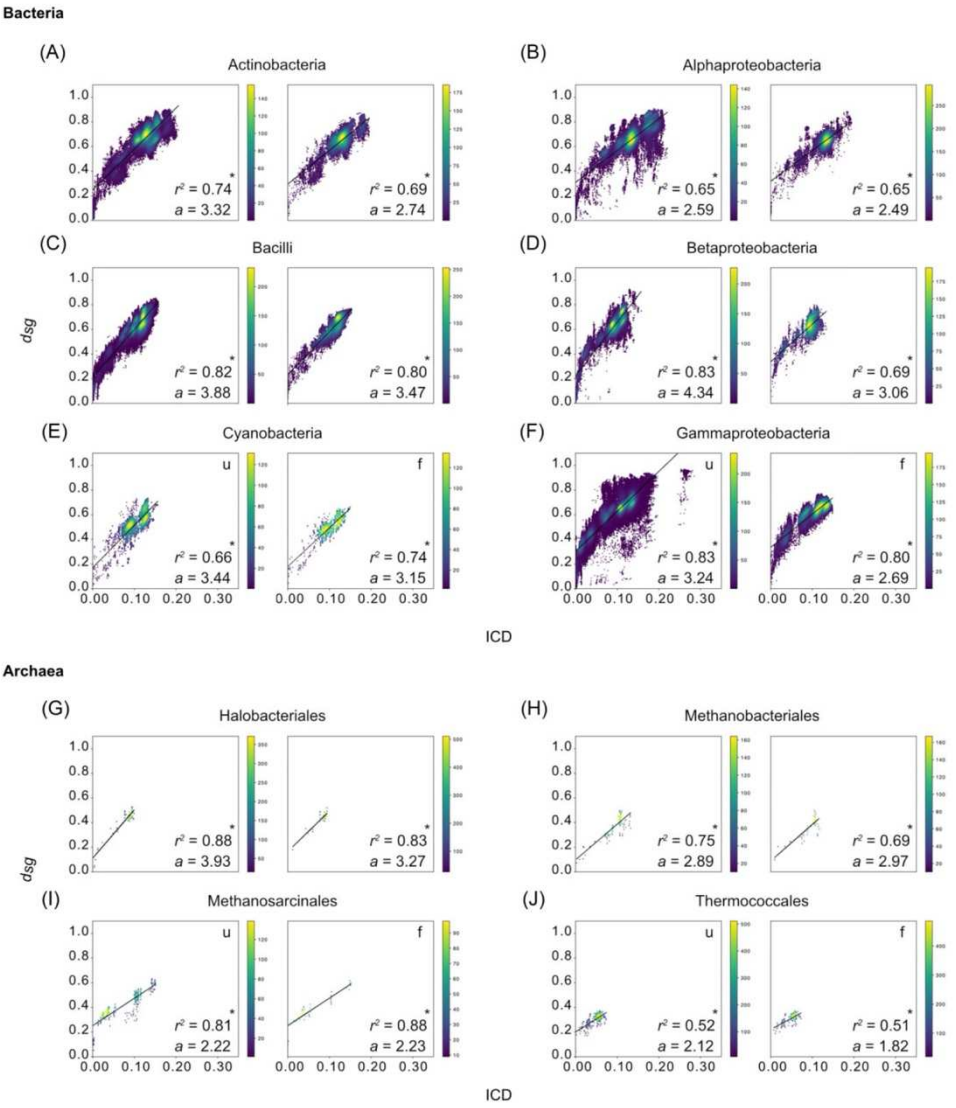


Figure 2. Correlation between ICD (x-axis) and gene sharing distance (d_{sg}; y-axis) of pairwise genome comparisons in prokaryotic higher taxa. Every point represents one genome comparison and are colored based on the density of points. Spearman rank correlations and linear regressions between ICD and d_{sg} are calculated for every higher taxon. The resulting r^2 of the linear regression as well as the slope a are shown in the lower right corner (* = Spearman rank P-value ≤ .01). The upper part of the figure shows six highly sampled bacterial higher taxa and the lower part four archaeal ones. For each prokaryotic higher taxon, two data sets were examined. Every left-hand panel, per taxon, shows the correlation in the original dataset (u). The right-hand panel contains the correlation in the filtered data set (f). Thereby, only genomes were used whose genome size was within the average higher taxon genome size ± one standard deviation. Additionally, only one random genome per species was selected. Statistics of all prokaryotic higher taxa sampled are listed in Fig. S1.

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

bias, only one genome per species was retained in the dataset. These plots for the data filtered for size and strains are shown in the right-hand panels of Fig. 2. For almost all higher taxa, the correlation remains almost unchanged relative to the unfiltered dataset (Fig. 2, right-hand panels; Fig. S1), as only outliers are removed.

We used the slopes of the linear regressions of plots of ICD versus d_{sg} as an estimate for the relative gene flux rates per higher taxon (Table S3). The calculated gene flux rates represent the proportion of gene changes per 1% amino acid differences in the IC gene set per higher taxon. When we compared taxa with sufficient sample sizes, the rate of gene flux is similar across higher taxa, varying within Bacteria from 2.04 for Chlamydiae to 3.47 for Bacilli. In archaeal higher taxa, the lowest rate is 1.82 for Thermococcales and the highest is 3.27 for Halobacteriales. On average, the bacterial gene flux rate is 2.90 and the archaeal gene flux rate 2.57, yielding an average rate of 2.83% gene content change per 1% ICD. The similar gene flux rates across higher taxa show that a regular, and averaged over long timescales, almost clock-like behavior of gene content change in genomes relative to amino acid sequence divergence in the most universal and vertically inherited genes exists in prokaryotes. The archaeal dataset used here is limited to the largest groups of Euryarchaea and does not include genomes from Asgard archaea or fast-evolving groups like DPANN and CPR. The analysis is based on referenced cultured genomes, whereby Asgard sequences from enrichment cultures comprise a very small sample. The gene flux rates within archaea are similar across groups sampled here, future studies will reveal whether genomes from other archaeal taxa show similar or anomalous rates of change for ICD versus d_{sg}.

Based on the average number of genes per higher taxon and the relative gene flux rates (Table S3) we calculated the number of different genes per substitution in the IC genes, to estimate absolute gene flux rates (Table S4). The average number of gene differences per substitution in the IC gene sequences is five for bacteria and four for archaea. However, in contrast to the relative gene flux rates (Table S3) the absolute rates vary substantially from 1.29 (Chlamydiae) to 9.03 (Betaproteobacteria). For archaea the highest rate of gene flux is found for Halobacteriales (6.27) and the lowest for Thermococcales (2.41; Table S4). The absolute gene flux rates are higher for free-living prokaryotes than for prokaryotic endosymbionts due to different genome sizes and the isolated nature of endosymbiont genome evolution (Clark et al. 1999, Shigenobu et al. 2000, van Ham et al. 2003, Moya et al. 2008, Moran and Bennett 2014, Martínez-Cano et al. 2015).

The question arises whether the similar gene flux rates are also present across different taxonomic levels. Therefore, we analysed the relationship between ICD and d_{sg} at lower taxonomic levels. The correlation remained nearly constant except of taxa on the species level (Fig. S2a and Table S5) because ICD between genomes corresponding to the same species is too close (Fig. 3D). In order to better quantify gene flux at or near the species level, we used a genome-based measure for species-level divergence, ANI (Konstantinidies and Tiedje 2005, Goris et al. 2007, Wright and Baum 2018). We calculated the ANI for all genome pairs, using FastANI (Jain et al. 2018). For direct comparison to the ICD measure, we calculated the average nucleotide differences by subtracting the ANI from 1. This measure provides a higher resolution for the intraspecies genome comparisons compared to ICD, showing a range of 0%–12% 1-ANI at an ICD range of 0%–1% (Fig. 3). The relationship between 1-ANI and gene sharing distance from species- to order-level shows that the positive correlation holds for all taxonomic levels (Fig. S2b and Table S5). As noted in previous studies at the species and genera level (Wright and Baum 2018, Hassler et al. 2022), a correlation between d_{sg} and 16S rRNA sequence divergence is not observed, as the latter tend to saturate at pairs distributed near 0.3 16S rRNA sequence divergence (Fig. S3).

By comparing the estimated relative rates of gene flux based on the slopes of the regression lines between ICD and d_{sg}, we see that the rate of gene content change is not constant across taxonomic levels. The relative gene flux rates decrease with increasing phylogenetic divergence even if the differences are small except for the species level where the rates are much higher than on all other taxonomic levels. (Fig. S4a and Table S6). This could indicate that the mechanism of gene content change within species and between species are different, as suggested by Baumdicker and Kupczok (2023). From the genus level onwards, the slopes decrease slowly with increasing phylogenetic depth, which can result from sequence saturation in distance comparisons (or potentially from, genes that are gained, lost, and regained). The highly disparate gene flux rates within species and between species are also shown by comparing the relative gene flux rates based on the slopes of the linear regression between 1-ANI and d_{sg} (Fig. S4b and Table S6). However, at taxonomic levels of genus, family and order the gene flux rates are similar.

Since the estimation of gene flux rates is based on linear regressions, which assume a normal distribution of residuals and low variance of y-values at all x-values, we performed residual analysis for the regression between ICD or 1-ANI and d_{sg} for all higher taxa and on lower taxonomic levels (Figs S5–S8). The correlations of d_{sg} with 1-ANI and ICD, though highly significant, were not strictly linear. High values of 1-ANI and ICD as well as low values of ICD show deviation from linearity (Figs S5 and S6). The deviation of high values of 1-ANI can be explained by the limitations of the method: ANI values can only be estimated up to about 20% 1-ANI because the values saturate (Fig. 4). ICD seems to be a robust measure at higher divergence. However, at values of ICD close to zero, no correlation is visible between ICD and d_{sg} because no difference in ICD values is detectable (Fig. 4B). To investigate improved fit of the regression line between ICD and d_{sg}, we performed logarithmic fits on genome comparisons of higher taxa as performed in previous studies (Touchon et al. 2009, Wielgoss et al. 2016, Andreani et al. 2017, Rocha 2018) including log–log transformations, log transformation of y-values (gene content differences, d_{sg}) and log transformation of x-values (ICD). The log–log transformations of ICD and d_{sg} values improved the r^2 values form the regression line in most higher taxa sampled and thus the fit (Fig. S7), especially in the unfiltered dataset (Fig. S8). However, the residual analysis from linear regressions on log–log transformations were often worse than (or equal to) that observed when analysing the linear regressions between ICD and d_{sg} without transformations (Figs S7 and S8). Linear regression on transformations of either ICD values or d_{sg} values resulted in worse fit than that found for linear regression between ICD and d_{sg} (Figs S7 and S8). Even though a power function (log–log transformation) gave a better fit of the regression line, we show the linear regressions of raw values without transformations, because neither are strictly linear (though obviously close to linear), and the residual analysis of raw values is better than (or not worse than) that observed with transformed data. Saturation of d_{sg} estimates at higher ICD values might stem from genes that are gained, lost, and regained at appreciable frequency in the current sample.

At low values of ICD in the raw dataset, vertical "icicle"-like structures, equidistant point groups (EPGs) can be observed (Fig. 2, left-hand panels). Genome pairs that generate EPGs have little or no ICD, but vary with respect to gene content differences. We

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

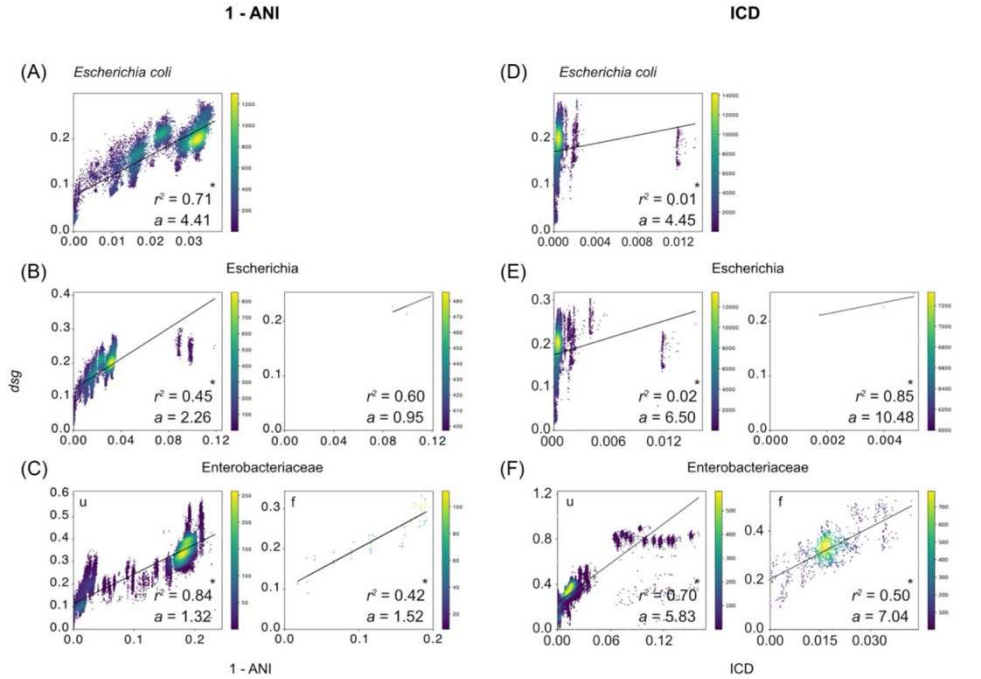


Figure 3. Relationship between 1-ANI and d_{sg} (A–C) as well as ICD and d_{sg} (D–F) for *E. coli*, *Escherichia*, and Enterobacteriaceae. Every point represents one genome comparison and are colored based on the density of points. Spearman rank correlations and linear regressions between ICD and d_{sg} are calculated for every taxon. The resulting r^2 of the linear regression as well as the slope *a* are shown in the lower right corner (* = Spearman rank *P*-value $\leq$ .01). For each prokaryotic taxon, two data sets were examined. Every left-hand panel, per taxon, shows the correlation in the original dataset (u). The right-hand panel contains the correlation in the filtered data set (f). Thereby, only genomes were used whose genome size was within the average higher taxon genome size ± 1 SD. Additionally, only one random genome per species was selected.

colored the genome pairs according to their taxonomic affiliation to highlight the dependence of EPGs upon phylogenetic structure of the data (Fig. 5). The colors indicate the lowest common taxonomic level of genome pairs. Comparisons between genomes from the same species are colored pink and comparisons between genomes from the same phylum are colored gray. EPGs are only observed when genome pairs belong to the same species or genus (Fig. 5A, upper left enlarged section). When we compared one genome of a species with all other genomes corresponding to the same species, the differences in their gene content become greater from genome to genome, forming the EPG (Fig. 5B). This indicates that every EPG reflects the growth of the accessory genome as more genomes are added to the species-level pangenome. At higher evolutionary divergence, the EPGs are lost (Fig. 5A, lower right enlarged section), because as taxonomic divergence approaches the phylum level, the shared component of accessory genomes in pairwise comparisons also approaches zero. This leads to a narrower distribution of gene sharing values for taxa with shared genes in the accessory genome in comparison to other species with increasing sequence divergence. One might argue that EPGs can also be formed by low-quality genomes. However, all genomes used in this portion of our analysis are complete genomes. The presence or absence of the EPGs represents the phylogenetic structure of the data, even though no trees were used for the calculation of the ICD. In the filtered dataset (Fig. 2, right-hand panels), the EPGs are no longer visible due to lacking intraspecies comparisons, as only one strain per species was retained in the dataset.

The *y*-axis intercept of the regression lines in Fig. 2 rarely goes through the origin. In most higher taxa, *y*-axis intercept assumes a value between 0.2 and 0.4. This means that in pairwise genome comparisons with zero ICD, the differences in pairwise gene content are 20%–40%. That is, at zero ICD, roughly 60%–80% of genes in a given comparison are shared. These 60%–80% paired genes are similar to the 70% value of DNA–DNA sequence hybridization (DDH), used in the 1960s and 1970s to delineate prokaryotic species (Wayne et al. 1987). The value of DDH correlates positively with the proportion of conserved DNA sequences (>90% sequence identity) in genomes pairs (Goris et al. 2007). In the present study, the *y*-axis intercept reflects the gene content divergence in genome pairs with identical ICD. In DDH, nonconserved sequences (<90% sequence identity) correspond to the proportion of nonshared genes in pangenome analysis, which belong to the accessory genome (Tettelin et al. 2005). In previous species level pangenome analyses, the proportion of accessory genes per genome typically varied between 0.2 and 0.5, calculated as 1 minus the proportion of species-level core genes in prokaryotic genomes (Tettelin et al. 2005, Rasko et al. 2008,

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

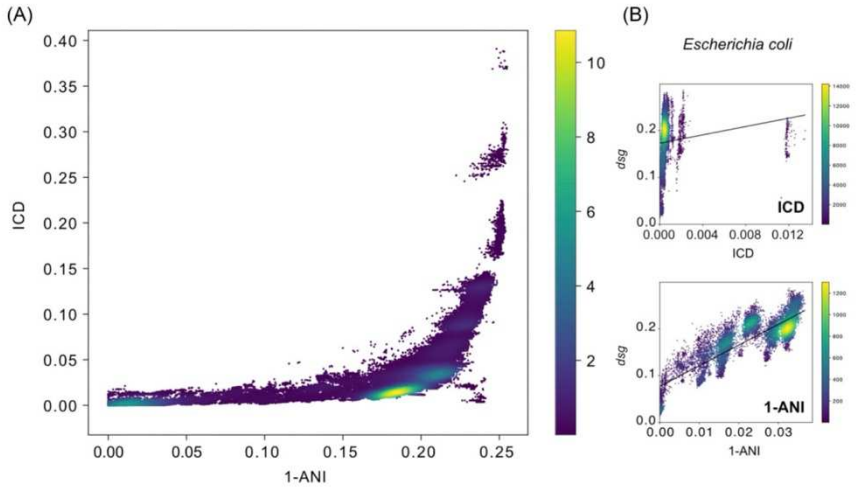


Figure 4. Relationship between 1-ANI and ICD for all genome pairs with calculated ANI from FastANI. (A) Relationship between 1-ANI (x-axis) and ICD (y-axis) for all genome pairs with calculate ANI ($n = 397\,578$). Every point represents one genome comparison and is colored based on the density of points. (B) Relationship between ICD or 1-ANI (x-axis) and d_{sg} (y-axis) for prokaryotic taxon *E. coli*. Every point represents one genome comparison and is colored based on the density of points.

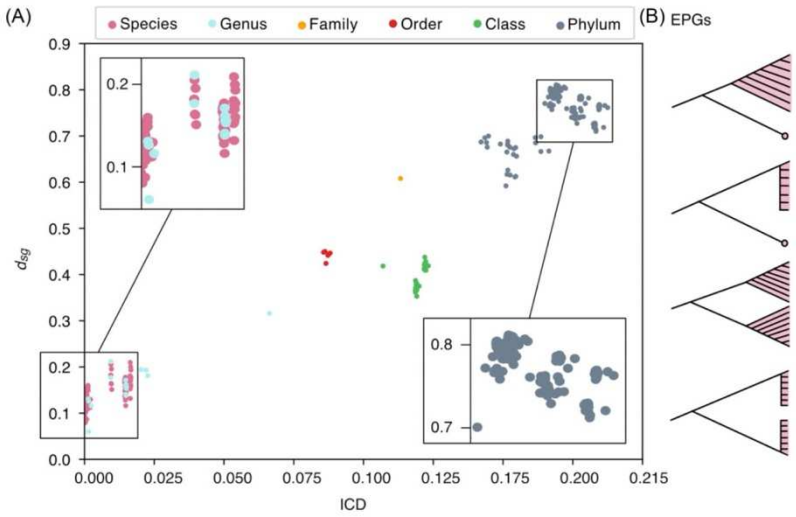


Figure 5. Correlation of ICD and d_{sg} of Chloroflexi. (A) Enlarged view of the correlation between ICD (x-axis) and d_{sg} (y-axis) of the bacterial taxon Chloroflexi. Points represent genome comparisons and are colored by the lowest equal taxon of the genomes. Equidistant point groups (EPGs) are shown in intraspecies comparisons in the upper left enlarged section. In lower taxonomic groups, like intraspecies and intragenus comparisons, EPGs are formed by genome comparisons that have a nearly identical average frequency of substitutions in IC genes but differ in their gene content. In intraphylum comparisons, the EPGs are lost (lower right enlarged section). Panel (B) shows possible tree structures that could represent EPGs of intraspecies comparisons.

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

Schoen et al. 2008, Scaria et al. 2010, van Schaik et al. 2010, Budroni et al. 2011, Park et al. 2019). This value is close to the average proportion of gene content differences between genomes having zero ICD across higher taxa found here.

Thus, the y-axis intercept provides a rough estimate for the proportion of accessory genes in genomes of the same species within the higher taxon. To test this, we calculated average species-specific accessory genome proportions per higher taxon based on the proportion of unpaired genes in pairwise genome comparisons corresponding to the same species. To avoid bias from small samples, only higher taxa represented by at least 10 strains, 2 species and a P-value ≤ .05 from the correlations calculated in Fig. 2 and Fig. S1 were used. The average species accessory genome proportions as well as the y-axis intercept were taken from the filtered dataset to avoid phylogenetic bias. The average species-level accessory genome proportions and the y-axis intercept values inferred from the regression lines are quite similar (Table S7) and correlate positively at P ≤ .01 (Fig. S9 and Table S7). That is, the y-axis intercept inferred from the entire history of the higher taxon delivers an estimate for the average accessory genome proportion in current genomes. This is interesting *per se*, but it is also evidence for the existence of pangenomes throughout the evolutionary history of prokaryotes.

To analyse the influence of cloud genes, genes that are present in only a few genomes or in only one genome, we excluded protein families containing proteins from less than 5%, 10%, 15%, or 20% of all genomes from the gene sharing distance calculation. We then compared the statistical results of the correlation and regression between the ICD and each filtered gene sharing distance dataset and the dataset including all protein families. For the Spearman rank correlation coefficient (r_s), the slope a, and r^2 of the regression line, no differences are shown (Fig. S10 and Table S8). The y-axis intercept decreases as more genes are excluded. This makes sense because the y-axis intercept represents the proportion of accessory genes per genome, including genes that are only present in a few genomes or only in one. Therefore, we cannot see any effect of cloud genes on the analysis.

The accessory genome is always more affected by LGT

Accessory genes are generally thought to be more often transferred between genomes and to contribute to functions important for the adaptation to a specific niche, whereas core genes are more conserved and tend to include housekeeping functions (Tettelin et al. 2005, 2008, Kung et al. 2010, Vernikos et al. 2015, Brockhurst et al. 2019). To see whether this aspect is captured by our approach, we plotted the average verticality of genes that are present in both genomes (paired) versus the average verticality of those that are missing in one genome of the pair (unpaired) for all genome pairs (Fig. 6). The verticality distributions between paired (red) and unpaired genes (gray) are significantly different for all genome comparisons corresponding to the same higher taxon (P = .0; Table S9) and are furthermore nonoverlapping sets. With evolutionary divergence, the verticality of paired genes as well as for unpaired genes increases. However, the slope for paired genes is much higher (a = 22.57) than the slope for unpaired genes (a = 2.89). The average verticality of paired genes increases because the more different two genomes are, the fewer genes they share. These genes are highly conserved and furthermore exhibit high verticality. The higher y-axis intercept for paired genes (paired genes y-axis intercept = 1.84, unpaired genes y-axis intercept = 0.34) shows that at zero ICD, accessory (unpaired) genes

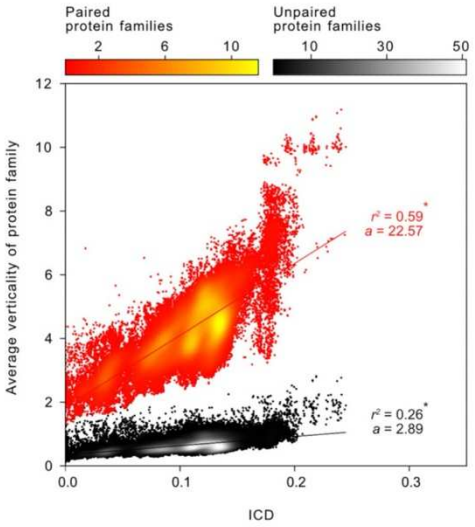


Figure 6. Comparison between average verticality (y-axis) of paired and unpaired protein families and divergence in IC genes (x-axis) in the filtered dataset. Paired protein families are defined as protein families in which both genomes of a genome comparison were represented by at least one protein. If only one genome is present in the protein family, the protein family corresponds to the group of unpaired protein families. For every genome pair two points are plotted. The average verticality of paired genes and the average verticality of unpaired genes in the compared genome pair relative to the sequence divergence in the IC genes from the compared genomes. The color of the points represents the density in the region near the point. Spearman rank correlations and linear regressions between ICD and d_{sg} are calculated for paired protein families and unpaired protein families (number of genome pairs = 124 447). The resulting r^2 of the linear regression as well as the slope a are shown next to the regression lines (* = Spearman rank P-value ≤ .01).

are always less vertical than core (paired) genes. This also holds for higher ICD, because all 124 447 compared genome pairs have a higher verticality for paired protein families than for unpaired, except for one comparison between two Mollicutes genomes (*Spiroplasma mirum* and *Spiroplasma atrichopogonis*). The analysis of the unfiltered dataset gave the same result (Fig. S11 and Table S9), whereby 88 of 2 229 958 comparisons have a higher average verticality for unpaired protein families than for paired ones, yet always belonging to the same species or genus.

The nonoverlapping verticality distribution for paired and unpaired genes in every genome comparison (Fig. 6) might suggest that the two gene sets are drawn from different samples, but because one and the same gene can be paired (core) in one comparison but unpaired in another, the two sets delineated in Fig. 6 simply indicate that the accessory genome is vastly more prone to LGT than the core genome in every genome comparison. This does not mean that core genes cannot be affected by LGT, however, it shows that accessory genes are far more readily transferred (Nesbø et al. 2001). An analysis of the frequency of functional categories (KO) from KEGG (Fig. S12; Kanehisa et al. 2017) revealed that, as expected genes belonging to the category translation were more frequent in the paired set than in the unpaired. The most common KEGG B functional annotation for unpaired genes is virus

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

information processing (Table S10). It is known that prokaryotic genomes have core and accessory components (Medini et al. 2005, Tettelin et al. 2005, 2008, Vernikos et al. 2015, Brockhurst et al. 2019). What Fig. 6 shows, is that in all comparisons, the accessory (unpaired) component is always comprised of genes that are transferred more frequently (they are less vertical). This is consistent with the observation that the tendency for a gene to undergo LGT is restricted by the presence of a preexisting copy (Nagies et al. 2020). It is also consistent with the early observation by Roger Milkman (1996) that *"The structure of genetic variation in a bacterial species is the result of recombination superimposed upon the repeated formation and spread of clones"*, whereby recombination in prokaryotes is never reciprocal and need not require donor and acceptor to belong to the same species, genus, phylum, or domain.

High-quality MAGs show similar rates of gene flux

MAGs have gained much attention in recent years, as cultured genomes are estimated to account for only a very small fraction of all prokaryotic organisms on earth (Lok 2015). Since many prokaryotes cannot be cultivated or only under difficult or expensive laboratory conditions, DNA samples are taken from the environment and are categorized into genomes—MAGs—using methods that assemble genome sized bins (Garza and Dutilh 2015). However, these methods are far from perfect (Garza and Dutilh 2015). As a consequence, programs have been developed to estimate the quality of MAGs. One of the most widely used methods for cross-checking MAGs is CheckM (Parks et al. 2015), which uses (i) marker genes that are specific to the genome's lineage, inferred from a reference genome tree, and (ii) marker genes that are usually single-copy in genomes of cultivated species, to calculate the completeness and the redundancy (contamination) of the MAG, respectively.

The two parameters estimated by CheckM are related to—but not identical to—the genomic parameters that we have investigated here, even though we do not use a reference tree or score genes as single copy. ICD contains evolutionary information (pairwise distances between genomes), but involves no tree, while d_{sg} gives information about how many genes are shared between two genomes, but not which genes, specifically, are shared.

Because a very similar pattern and correlation between ICD and d_{sg} is observed across a wide spectrum of genomes of cultured strains, it was of interest to see how MAGs appear when viewed from the standpoint of ICD versus d_{sg}. For that, we clustered a dataset of 2872 MAGs that contained a spectrum of assemblies with varying quality. The metagenomic data were clustered using the same methods as for cultured strains, values of ICD and d_{sg} were calculated accordingly. Completeness and contamination values were calculated using CheckM (Park et al. 2015) as the quality measure, defined by Parks et al. (2017), providing a filter for the quality of metagenome assemblies. This quality measure is calculated by subtracting five times the contamination from the completeness (Parks et al. 2017), a procedure that at high stringencies yields metagenome assemblies resembling genomes of cultured strains. We filtered the MAGs for different qualities, ranging from no filter (at least 0% quality) to at least 90% quality and performed correlation and linear regression analysis between ICD and d_{sg} for each dataset (Fig. 7).

At quality thresholds <50%, there is a visible tendency of pairwise comparisons from assemblies with ICD $\approx$ 0 to span the full range of $0 < d_{sg} < 0.8$. This reflects the existence of unusually disparate gene collections in low quality assemblies that are closely related by the measure of ICD divergence (ICD <0.05). Also at quality thresholds <50%, there is a tendency for assemblies to exhibit $d_{sg} \approx 1$ across all values of ICD. This reflects a class of assemblies that have gene collections more disparate than that observed for cultured strains, independent of ICD. At the same time, values of ICD often exceed 0.35 for low quality assemblies, something not observed for cultures strains, even in bacterial–archaeal comparisons (Fig. 8), suggesting that in low quality assemblies, IC genes may contain erroneous sequence information. With increasing quality filter stringency, however, MAGs reflect the properties of cultured strains with respect to gene flux rate.

In Fig. 7 it is clearly shown that r^2 from linear regression between ICD and d_{sg} increases with the quality of metagenomes data. The highest value is calculated for the dataset using MAGs with at least 90% quality (Fig. 7J). Values for 100% quality are not shown because only one MAG was estimated to have a quality of 100%. As r^2 in cultured genomes generally varies between 0.60 and 0.95 (Fig. S1) a reliable prediction of statistical values between the correlation and regression of ICD and d_{sg} for dataset including MAGs is only possible by using high-quality MAGs of more than 80%. In lower-quality datasets an overrepresentation of genome pairs that exhibit extremely high ICD values is shown (Fig. 7A–F) as well as genome pairs with low ICD but high gene content differences (d_{sg}; Fig. 7A–D). The similar properties of cultured genomes (Fig. 2 and Table S3) and high-quality MAGs (Fig. 7I–J) including r^2, the relative gene flux rates (a) and the y-axis intercept from the regression lines between ICD and d_{sg}, indicates that a constant gene flux rate across higher prokaryotic taxa, with high gene flux within the accessory genome and lower flux in the core, also applies to MAGs, if the quality of the data is high.

Pangenome structure throughout prokaryote evolution

On the basis of pairwise genome comparisons, we can observe a clock-like rate of gene turnover within higher taxa, which can be used to estimate species-level accessory and core genome sizes and which properties reflect the phylogenetic structure of the data. Since LGT events are also detectable between superkingdoms, primarily between archaea and bacteria (Rest and Mindell 2003, Gophna et al. 2004, Boto 2010), the last universal common ancestor (LUCA) (Weiss et al. 2016) might have had the ability to undergo LGT, and therefore a genome containing core genes with mainly vertical inheritance and accessory genes that are in permanent flux (Woese 2002, Nagies et al. 2020).

To test whether LUCA might have had similar rates of gene flux as found in current genomes, we expanded the analysis from Fig. 2 to all pairwise genome comparisons within prokaryotes (Fig. 8), including bacterial and archaeal genome comparisons corresponding to the same higher taxon (bacteria number of genome pairs = 123 940; archaea number of genome pairs = 507), genome comparisons corresponding to different higher taxa (bacteria number of genome pairs = 1 013 846; archaea number of genome pairs = 6753) and intersuperkingdom comparisons between archaea and bacteria (number of genome pairs = 182 589) from the dataset filtered for genome sizes and strains. The rates for bacterial and archaeal comparisons corresponding to only one higher taxon (bacterial slope = 2.51, archaeal slope = 2.07) are close to the average rates calculated in Table S3, with small differences attributable to genome comparisons corresponding to higher taxa with small sample sizes. The correlation between ICD and d_{sg} also holds for deep genome comparisons corresponding to different higher taxa and intersuperkingdom comparisons ($P = .0$)

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

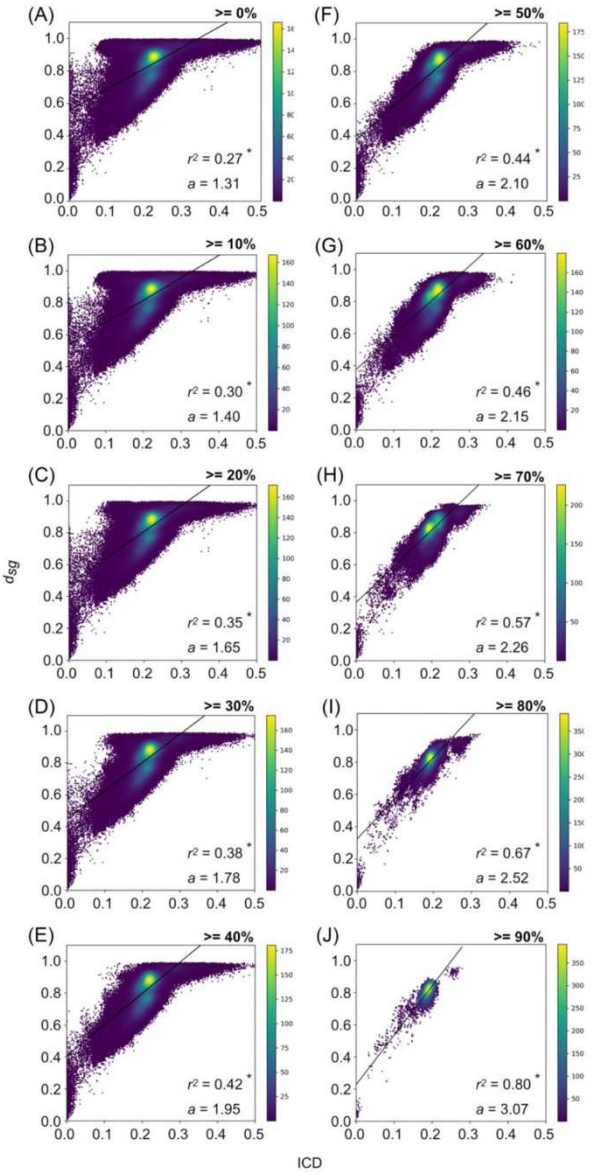


Figure 7. Correlation between ICD and d_{sg} using MAGs with a minimum quality ranging from at least 0% to at least 90%. Quality is defined as completeness minus five times the contamination and is shown at the right side over every panel (Parks et al. 2015). The slope (a) and r^2 from linear regression between ICD and d_{sg} are shown in every lower-right corner (∗ = Spearman $P \leq .01$). Points are colored based on their density.

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

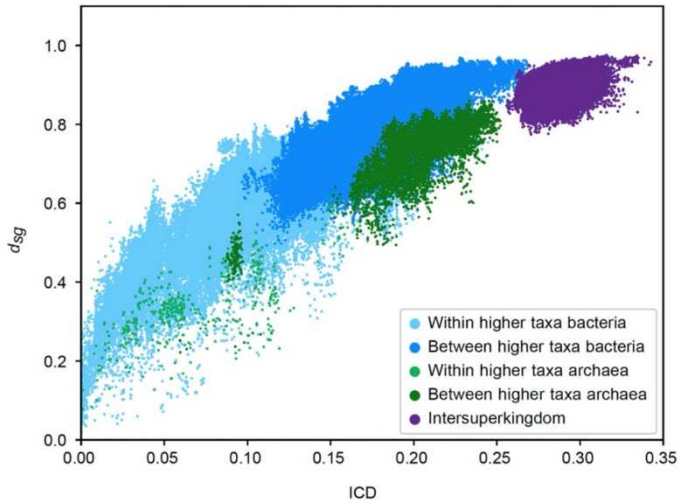


Figure 8. Correlation between ICD and gene sharing distance (d_{sg}) for prokaryotic genome pairs in higher taxonomic ranks. Correlation between ICD (x-axis) and d_{sg} (y-axis) of bacterial and archaeal genome pairs corresponding to the same higher taxon (within higher taxa), bacterial and archaeal genome pairs corresponding to different higher taxa (between higher taxa), and intersuperkingdom genome comparisons. Every point represents one pairwise genome comparison and is colored based on affiliation of the compared genomes to bacteria (blue), archaea (green), or both (purple). Lighter blue and green show pairwise comparisons between genomes corresponding to the same higher taxon. Darker blue and green shows comparisons corresponding to different higher taxa.

as well as the verticality distribution for paired (high verticality) and unpaired genes (low verticality; Fig. S13). This indicates that the basic mechanisms of gene content change in prokaryotic genomes—rapid turnover of genes in the accessory genome and slow gene flux in the core genome—has been operational since the divergence of the bacterial and archaeal lineages. However, the slope decreases, yet we have not applied any kind of correction for multiple events on either axis in order to obtain low-parameter estimates. In pairwise comparisons of bacterial genomes corresponding to different higher taxa, the estimated gene flux rate is 1.80% gene content differences at 1% ICD, between superkingdoms 0.78%.

A persistent rate of flux relative to ICD could indicate that a pangenome structure has persisted throughout prokaryotic history. Our analyses provide neither an estimate for the size of LUCA's pangenome nor an estimate of which genes, specifically, it contained. The data do, however, reveal that the rate of gene content change is roughly constant as far back as our sample probes the bacterial and archaeal lineages, indicating that the average rate of gene content change that we observe today has existed throughout prokaryotic evolution all the way back to the first free-living cells. This, in turn, traces a pangenome structure—a core genome and an accessory genome—back to the common ancestor of bacteria and archaea.

A possible application of our procedure to pangenomes concerns their impact on the origins of organelles (Esser et al. 2007, Ku et al. 2015). We embarked upon estimation of the number of genes that have persisted in any individual bacterial genome that shared common ancestry with the mitochondrial ancestor or with the cyanobacterial ancestor. However, the degree of sequence divergence between the bacterial and nuclear homologues of organelle-encoded or organelle-derived IC genes (0.4–0.5; Table S11–Supplemental Results) exceeds that observed from prokaryotic genomes (0.35; Fig. 8) such that the current method does not permit interpolation.

Conclusion

Using universally applicable proxies for genome divergence, we found very similar rates of gene flux across all higher prokaryotic taxa sampled. Gene content divergence in prokaryotic accessory genomes correlates linearly two measures of sequence divergence, ANI (nucleotides) and ICD (amino acids), which in turn indicates a clock-like rate of gene flux across higher prokaryotic taxa. The analysis shows that a universal rate of gene flux is present at all taxonomic levels, with higher gene flux within species than across higher taxonomic groups. Baumdicker and Kupczok (2023) proposed that within-population gene gain spreads genes within the population, and between-population gene gain leads to the acquisition of novel genes. This is consistent with our observation that within species, gene flux rates differ across taxa, while between species the rates remain similar.

Our use of a measure for lineage divergence that is generally applicable across comparisons of all higher prokaryotic taxa allowed us to look deeper into time than the genus or species level, where ANI performs well. The generally linear correlation between gene content differences and ICD holds as far back as any of these lineages can be traced, including the archaea–bacteria split. This indicates that the rate of genome flux and the pangenome structure of modern prokaryotic genomes is as ancient as prokaryotes themselves. This is an independent line of evidence indicating that LGT is a natural component of genome evolution in prokaryotes and has been in operation throughout the entirety of their evolution.

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

There is an ongoing debate about the main evolutionary forces shaping gene content in pangenomes (Vos et al. 2015). Here, we have observed similar and constant rates of gene flux across prokaryotic taxa. By analogy to nucleotide substitution rates, this could suggest a neutral or nearly neutral nature of gene flux in prokaryotes (Baumdicker and Kupczok 2023), compatible with Kimura's neutral theory (Kimura 1968) or the nearly neutral theory proposed by Ohta (1973). However, similar gene flux rates across species can also be obtained using selective models (Barrick et al. 2009). Neither mechanism can be expected to explain all fixation events. Both neutral and selective mechanisms likely influence the rate and frequency of genes fixed by flux between the accessory genomes of prokaryotes.

Acknowledgements

We thank Cerys Viktoria Wilke for downloading prokaryotic 16S rRNA sequences and Tal Dagan for her helpful comments and suggestions on the manuscript. Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf.

Supplementary data

Supplementary data is available at *FEMSLE Journal* online.

Conflict of interest : The authors declare no conflict of interest.

Funding

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (grant agreement number 101018894 to W.F.M.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

The data underlying this article are available in the article, in its online supplementary material and in https://uni-duesseldorf.sciebo.de/s/YiiyDz57sDala1t.

References

Albalat R, Cañestro C. Evolution by gene loss. *Nat Rev Genet* 2016;**17**:379–91.

Andreani NA, Hesse E, Vos M. Prokaryote genome fluidity is dependent on effective population size. *ISME J* 2017;**11**:1719–21.

Arnold BJ, Huang I, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol* 2022;**20**:206–18.

Barrick JE, Yu DS, Yoon SH *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 2009;**461**:1243–7.

Baumdicker F, Kupczok A. Tackling the pangenome dilemma requires the concerted analysis of multiple population genetic processes. *Genome Biol Evol* 2023;**15**. https://doi.org/10.1093/gbe/evad067.

Boto L. Horizontal gene transfer in evolution: facts and challenges. *Proc R Soc B* 2010;**277**:819–27.

Brockhurst MA, Harrison E, Hall JPJ *et al.* The ecology and evolution of pangenomes. *Curr Biol* 2019;**29**:R1094–103. https://doi.org/ 10.1016/j.cub.2019.08.012.

Brueckner J, Martin WF. Bacterial genes outnumber archaeal genes in eukaryotic genomes. *Genome Biol Evol* 2020;**12**:282–92.

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

Budroni S, Siena E, Dunning Hotopp JC *et al. Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc Natl Acad Sci USA* 2011;**108**:4494–9.

Charlebois RL, Doolittle WF. Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res* 2004;**14**:2469–77.

Clark MA, Moran NA, Baumann P. Sequence evolution in bacterial endosymbionts having extreme base compositions. *Mol Biol Evol* 1999;**16**:1586–98.

Dagan T, Martin WF. The tree of one percent. *Genome Biol* 2006;**7**:118. https://doi.org/10.1186/gb-2006-7-10-118.

Esser C, Martin W, Dagan T. The origin of mitochondria in light of a fluid prokaryotic chromosome model. *Biol Lett* 2007;**3**:180–4.

Garg SG, Kapust N, Lin W *et al.* Anomalous phylogenetic behavior of ribosomal proteins in metagenome-assembled Asgard archaea. *Genome Biol Evol* 2021;**13**. https://doi.org/10.1093/gbe/evaa238.

Garza DR, Dutilh BA From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. *Cell Mol Life Sci* 2015;**72**:4287–308.

Gophna U, Charlebois RL, Doolittle WF. Have archaeal genes contributed to bacterial virulence?. *Trends Microbiol* 2004;**12**:213–9.

Goris J, Konstantinidis KT, Klappenbach JA *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007;**57**:81–91.

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**:1655–63.

Hao W, Golding GB. The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res* 2006;**16**:636–43.

Hassler HB, Probert B, Moore C *et al.* Phylogenies of the 16S rRNA gene and its hypervariable regions lack concordance with core genome phylogenies. *Microbiome* 2022;**10**. https://doi.org/10.1186/s40168-022-01295-y.

Haudiquet M, De Sousa JM, Touchon M *et al.* Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Phil Trans R Soc B* 2022;**377**. https://doi.org/10.1098/rstb.2021.0234.

Jain C, Rodriguez-R LM, Phillippy AM *et al.* High throughput ANI analysis of 90 K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;**9**. https://doi.org/10.1038/s41467-018-07641-9.

Kanehisa M, Furumichi M, Tanabe M *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61.

Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on Fast Fourier transform. *Nucleic Acids Res* 2002;**30**:3059–66.

Kimura M. Evolutionary rate at the molecular level. *Nature* 1968;**217**:624–6.

Kislyuk AO, Haegeman B, Bergman NH *et al.* Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* 2011;**12**. https://doi.org/10.1186/1471-2164-12-32.

Konstantinidis KT, Tiedje JM. Genomic insight that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* 2005;**102**:2567–72.

Ku C, Nelson-Sathi S, Roettger M *et al.* Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* 2015;**524**:427–32.

Kung VL, Ozer EA, Hauser AR. The accessory genome of *Pseudomonas aeruginosa*. *Microbiol Mol Biol Rev* 2010;**74**:621–41.

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

Lawrence JG, Ochman H. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA* 1998;**95**:9413–7.

Lok C. Mining the microbial dark matter. *Nature* 2015;**522**:270–3.

Marri PR, Hao W, Golding GB. Gene gain and gene loss in *Streptococcus*: is it driven by habitat?. *Mol Biol Evol* 2006;**23**:2379–91.

Marri PR, Hao W, Golding GB. The role of laterally transferred genes in adaptive evolution. *BMC Evol Biol* 2007;**7**:1–14.

Martínez-Cano DJ, Reyes-Prieto M, Martínez-Romero E *et al.* Evolution of small prokaryotic genomes. *Front Microbiol* 2015;**5**. https://doi.org/10.3389 /fmicb.2014.00742.

Medini D, Donati C, Tettelin H *et al.* The microbial pan-genome. *Curr Opin Genet Dev* 2005;**15**:589–94.

Milkman R. Recombinational exchange between clonal populations. In: Neidhardt FC (ed.), *Escherichia coli and Salmonella: Cellular and Molecular Biology*. 2nd edn. Washington: American Society for Microbiology Press, 1996.

Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. *Trends Genet* 2001;**17**:589–96.

Moran NA, Bennett GM. The tiniest tiny genomes. *Annu Rev Microbiol* 2014;**68**:195–215.

Moya A, Peretó J, Gil R *et al.* Learning how to live together: genomic insights into prokaryote–animal symbioses. *Nat Rev Genet* 2008;**9**:218–29.

Nagies FSP, Brueckner J, Tria FDK *et al.* A spectrum of verticality across genes. *PLoS Genet* 2020;**16**:e1009200. https://doi.org/10.1371/journal.pgen.1009200.

Nesbø CL, Boucher Y, Doolittle WF. Defining the core of non-transferable prokaryotic genes: the euryarchaeal core. *J Mol Evol* 2001;**53**:340–50.

Nowell RW, Green S, Laue BE *et al.* The extent of genome flux and its role in the differentiation of bacterial lineages. *Genome Biol Evol* 2014;**6**:1514–29.

O'Leary NA, Wright MW, Brister JR *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**:D733–45.

Ohta T. Slightly deleterious mutant substitutions in evolution. *Nature* 1973;**246**:96–98.

Park SC, Lee K, Kim YO *et al.* Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. *Front Microbiol* 2019;**10**. https://doi.org/10.3389/fmicb.2019.00834.

Parks DH, Imelfort M, Skennerton CT *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;**25**:1043–55.

Parks DH, Rinke C, Chuvochina M *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2017;**2**:1533–42.

Rasko DA, Rosovitz MJ, Myers GS *et al.* The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* 2008;**190**:6881–93.

Rest JS, Mindell DP. Retroids in archaea: phylogeny and lateral origins. *Mol Biol Evol* 2003;**20**:1134–42.

Rice P, Longden L, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet* 2000;**16**:276–7.

Rinke C, Schwientek P, Sczyrba A *et al.*, Insight into the phylogeny and coding potential of microbial dark matter. *Nature* 2013;**499**:431–7.

Rivera MC, Jain R, Moore JE *et al.* Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA* 1998;**95**:6239–44.

Rocha EPC. Neutral theory, microbial practice: challenges in bacterial population genetics. *Mol Biol Evol* 2018;**35**:1338–47.

Scaria J, Ponnala L, Janvilisri T *et al.* Analysis of ultra low genome conservation in *Clostridium difficile*. *PLoS One* 2010;**5**:e15147. https://doi.org/10.1371/journal.pone.0015147.

Schoen C, Blom J, Claus H *et al.* Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*. *Proc Natl Acad Sci USA* 2008;**105**:3473–8.

Shigenobu S, Watanabe H, Hattori M *et al.* Genome sequence of the endocellular bacterial symbiont aphids *Buchnera* sp. APS. *Nature* 2000;**407**:81–86.

Stull GW, Qu X, Parins-Fukuchi C *et al.* Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. *Nat Plants* 2021;**7**:1015–25.

Tettelin H, Masignani V, Cieslewicz MJ *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome.". *Proc Natl Acad Sci USA* 2005;**102**:13950–5.

Tettelin H, Riley D, Cattuto C *et al.* Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 2008;**11**:472–7.

Touchon M, Hoede C, Tenaillon O *et al.* Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* 2009;**5**:e1000344. https://doi.org/10.1371/journal.pgen.1000344.

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

Tria FDK, Martin WF. Gene duplications are at least 50 times less frequent than gene transfers in prokaryotic genomes. *Genome Biol Evol* 2021;**13**. https://doi.org/10.1093/gbe/evab224.

UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Re* 2021;**49**:D480–9.

van Dongen S.Graph clustering via a discrete uncoupling process.*Siam Journal on Matrix Analysis and Applications* 2008;**30**:121–41.

van Ham RCHJ, Kamerbeek J, Palacios C *et al.* Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci USA* 2003;**100**:581–6.

van Schaik W, Top J, Riley DR *et al.* Pyrosequencing-based comparative genome analysis of the nosocomial pathogen *Enterococcus faecium* and identification of a large transferable pathogenicity island. *BMC Genomics* 2010;**11**:1–18.

Vernikos G, Medini D, Riley DR *et al.* Ten years of pan-genome analyses. *Curr Opin Microbiol* 2015;**23**:148–54.

Vos M, Hesselman MC, Beek TAT *et al.* Rates of lateral gene transfer in prokaryote: high but why?. *Trends Microbiol* 2015;**23**:598–605.

Wayne LG, Brenner DJ, Colwell RR *et al.* Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Evol Microbiol* 1987;**37**:463–4.

Weiss MC, Sousa FL, Mrnjavac N *et al.* The physiology and habitat of the last universal common ancestor. *Nat Microbiol* 2016;**1**:16116. https://doi.org/10.1038/nmicrobiol.2016.116.

Wernegreen JJ. Endosymbiont evolution: predictions from theory and surprises from genomes. *Ann NY Acad Sci* 2015;**1360**:16–35.

Wielgoss S, Didelot X, Chaudhuri RR *et al.* A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium *Myxococcus xanthus*. *ISME J* 2016;**10**:2468–77.

Woese CR. On the evolution of cells. *Proc Natl Acad Sci USA* 2002;**99**:8742–7.

Wolf YI, Koonin EV. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol Evol* 2012;**4**:1286–94.

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

Wolf YI, Makarova KS, Lobkovsky AE *et al.* Two fundamentally different classes of microbial genes. *Nat Microbiol* 2016;**2**:1–6.

Wright ES, Baum DA. Exclusivity offers a sound yet practical species criterion for bacteria despite abundant gene flow. *BMC Genomics* 2018;**19**:1–12.

Downloaded from https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnae068/7737773 by Universitaets- und Landesbibliothek Duesseldorf user on 27 September 2024

Received 22 March 2024; **revised** 15 May 2024; **accepted** 19 August 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of FEMS. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

II The radical impact of oxygen on prokaryotic evolution – enzyme inhibition first, uninhibited essential biosynthesis second, aerobic respiration third

Natalia Mrnjavac[1], Falk S. P. Nagies[1], Jessica L. E. Wimmer[1], Nils Kapust[1], Michael R. Knopp[1], **Katharina Trost[1]**, Luca Modjewski[1], Nico Bremer[1], Marek Mentel[2], Mauro Degli Esposti[3], Itzhak Mizrahi[4], John F. Allen[5] and William F. Martin[1] (2024).

1 Institut für Molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Deutschland
2 Department of Biochemistry, Faculty of Natural Sciences, Comenius University in Bratislava, Bratislava, Slovakia
3 Center for Genomic Sciences, UNAM Campus de Cuernavaca, Mexico
4 Department of Life Sciences, Ben-Gurion University of the Negev and The National Institute for Biotechnology in the Negev, Be'er-Sheva, Israel
5 Research Department of Genetics, Evolution and Environment, University College London, UK

Dieser Artikel wurde am 22. Juli 2024 in *FEBS Letters* Ausgabe 598 veröffentlicht.

Beitrag von Katharina Trost:
Ich habe an der Erstellung von Abbildung 3 mitgearbeitet und die dort abgebildeten 3-dimensionalen Punktdiagramme erstellt. Außerdem habe ich die ergänzende Abbildung S7 erstellt und war an der Überarbeitung des Manuskriptes beteiligt.

Check for updates

FEBS Letters

FEBS PRESS science publishing by scientists

RESEARCH ARTICLE

The radical impact of oxygen on prokaryotic evolution—enzyme inhibition first, uninhibited essential biosyntheses second, aerobic respiration third

Natalia Mrnjavac[1], Falk S. P. Nagies[1], Jessica L. E. Wimmer[1], Nils Kapust[1], Michael R. Knopp[1], Katharina Trost[1], Luca Modjewski[1], Nico Bremer[1], Marek Mentel[2], Mauro Degli Esposti[3], Itzhak Mizrahi[4], John F. Allen[5] and William F. Martin[1]

1 Institute of Molecular Evolution, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Germany
2 Department of Biochemistry, Faculty of Natural Sciences, Comenius University in Bratislava, Bratislava, Slovakia
3 Center for Genomic Sciences, UNAM Campus de Cuernavaca, Mexico
4 Department of Life Sciences, Ben-Gurion University of the Negev and The National Institute for Biotechnology in the Negev, Be'er-Sheva, Israel
5 Research Department of Genetics, Evolution and Environment, University College London, UK

Correspondence
N. Mrnjavac, Institute of Molecular Evolution, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Universitätsstrasse 1, 40225 Düsseldorf, Germany
Tel: +49 211 8112736
E-mail: n.mrnjavac@hhu.de

Natalia Mrnjavac, Falk S. P. Nagies, Jessica L. E. Wimmer, and Nils Kapust contributed equally to this article.

(Received 13 February 2024, revised 12 April 2024, accepted 19 April 2024)

doi:10.1002/1873-3468.14906

Edited by Peter Brzezinski

Molecular oxygen is a stable diradical. All O_2-dependent enzymes employ a radical mechanism. Generated by cyanobacteria, O_2 started accumulating on Earth 2.4 billion years ago. Its evolutionary impact is traditionally sought in respiration and energy yield. We mapped 365 O_2-dependent enzymatic reactions of prokaryotes to phylogenies for the corresponding 792 protein families. The main physiological adaptations imparted by O_2-dependent enzymes were not energy conservation, but novel organic substrate oxidations and O_2-dependent, hence O_2-tolerant, alternative pathways for O_2-inhibited reactions. Oxygen-dependent enzymes evolved in ancestrally anaerobic pathways for essential cofactor biosynthesis including NAD^+, pyridoxal, thiamine, ubiquinone, cobalamin, heme, and chlorophyll. These innovations allowed prokaryotes to synthesize essential cofactors in O_2-containing environments, a prerequisite for the later emergence of aerobic respiratory chains.

Keywords: aerobic metabolism; evolution of aerobes; evolution of respiration; great oxidation event; lateral gene transfer; oxygen inhibition

The Great Oxidation Event, GOE [1], divides Earth's history close to its midpoint. Roughly 2.4 billion years ago (Ga), photosynthetic prokaryotes with two chlorophyll-based photosystems linked in series—cyanobacteria—evolved the molecular tools needed to extract electrons from H_2O, and to use them to fix CO_2 and N_2 for growth [2,3]. Oxygen made by cyanobacteria is ground state triplet O_2, a stable diradical with two unpaired electrons of identical spin. Its structure is better written as $^{\bullet}O–O^{\bullet}$ instead of O=O to underscore its diradical nature [4] (Fig. 1). Radicals are molecules having unpaired valence electrons. They have the property of extracting single electrons from available donors so as to restore a stable octet electron configuration, converting the donor into a new radical in the process. Though most radicals are extremely reactive [5], the O_2 diradical is generally unreactive [6]. It 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

Abbreviations
GOE, great oxidation event; LGT, lateral gene transfer; OEC, oxygen-evolving complex; PAL, present atmospheric level; PLP, pyridoxal phosphate; ROS, reactive oxygen species; SLP, substrate-level phosphorylation; SOD, superoxide dismutase; *V*, verticality.

FEBS Letters (2024) © 2024 The Authors. *FEBS Letters* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

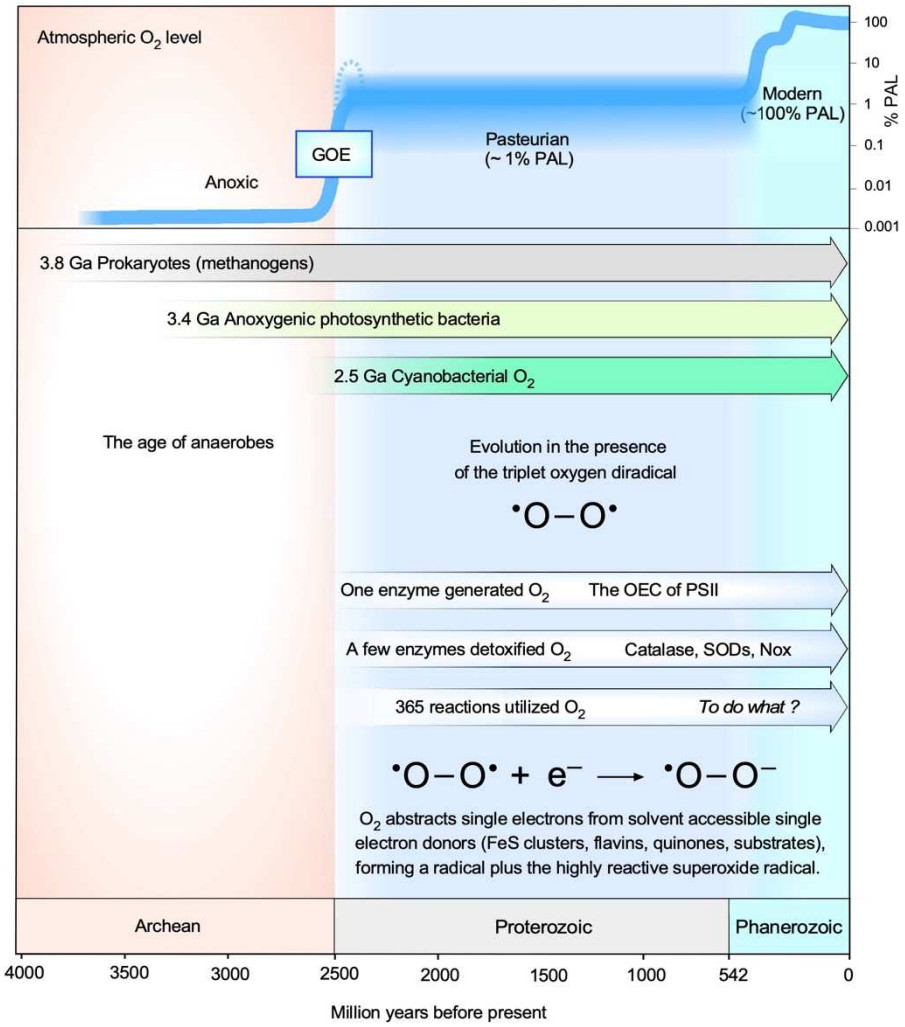


Fig. 1. A timeline of Earth history with the rise of O_2 and the appearance of relevant groups of prokaryotes. Blue boxes emphasize the GOE, the onset of O_2-dependent enzymes, and questions concerning their evolution and functions. The number 365 represents the number of O_2-dependent reactions in KEGG that we mapped to protein families (see text). SOD: superoxide dismutase; Nox: NADH oxidases (diaphorases), which are oxygen detoxifying enzymes [9]. Data from references [1,2,10–14]. Note that ultra-light carbon in 3.8 Ga rocks can be interpreted as evidence of both archaeal methanogens [15] and bacterial acetogens, which carry the same isotope signal [16], because both fix CO_2 *via* the acetyl-CoA pathway. A broken line reaching to 10% PAL around the end of the GOE indicates the Lomagundi excursion [10]. The reasons why O_2 levels remained near the Pasteur point for 1.8 billion years are still discussed. Numerous geological causes [10] and one biological cause [12] for the existence of the boring billion have been proposed. It is undebated that cyanobacteria (and their descendants, plastids) generated the current global supply of O_2 *via* one single enzyme and one single enzyme activity: the conserved Mn_4CaO_5-containing oxygen-evolving complex (OEC) of photosystem II. O_2 is written as •O–O• instead of O=O to underscore its diradical nature [4]. By 3.4–3.3 Ga, anoxygenic photosynthetic prokaryotes were generating stromatolites in aerial settings [17].

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

large resonance stabilization energy, and consequently a high activation energy barrier [4]. Despite this, O_2 has a very weak σ bond [4], which renders O_2 an extremely energy-rich molecule [7], so energy-rich that it undergoes exergonic redox reactions with every element except gold [8].

From the origin of the first microbes roughly 4 Ga to the onset of the GOE, the Earth's oceans and atmosphere were effectively devoid of O_2 [10] (Fig. 1). At the GOE, O_2 became introduced into Earth's oceans and atmosphere to approximately 1% of its present atmospheric level (PAL), and stayed more or less constant at this level for roughly 1.8 billion years. O_2 levels started to rise again about 580 million years ago, approaching modern values with the advent of land plants ca. 450 million years ago [10–13]. This protracted low oxygen phase of Earth history from 2.4 Ga to 0.58 Ga has been called the 'boring billion' [14] to emphasize the lack of geologically interesting events during that period of O_2 stasis, but it has also been called the 'Pasteurian' era of life's history [18] to emphasize the crucial observation that O_2 levels of 1% PAL correspond to the Pasteur point—the level of ambient oxygen (ca. 1% PAL or 0.2% v/v) at which facultative aerobes switch their terminal acceptors from anaerobic to aerobic respiration. The "boring" Pasteurian billion was the era during which prokaryotes learned to cope with the reactivity of O_2, to use O_2 for their own benefit, and to evolve alternatives to enzymes that were inhibited by O_2 (Fig. 1).

Prokaryotes were eye witnesses to the GOE. Since they lived before, during, and after the GOE (Fig. 1), their O_2-dependent enzymes should hold clues about the impact of O_2 on physiological evolution. To characterize the impact of O_2 in biochemical evolution, we identified 365 enzymatic reactions of prokaryotes that utilize O_2 in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [19] and that map to phylogenies of prokaryotic genes [20]. The results provide novel insights into the impact that O_2 exerted on microbial evolution, the nature of physiological traits that O_2-dependent enzymes imparted, and the dispersal of genes for O_2-dependent enzymes *via* lateral gene transfer (LGT) following the GOE.

Materials and methods

Collection of oxygen-dependent and oxygen-independent reactions

Data for 11 804 metabolic reactions were downloaded from the KEGG [19] 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 (EC) 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 reactions were subsequently filtered for reactions involving O_2 (KEGG compound C00007), yielding a set of 1949 O_2-dependent reactions, most of these being specific to eukaryotes. The reactions were mapped to prokaryotic protein families using KEGG orthology to link reactions to sequences. Protein families were created using MCL [21] as previously described [20,22], and only protein families with 4 or more sequences were used in this analysis. Only reactions that we could link to prokaryotic protein families were retained, yielding 365 O_2-dependent reactions mapped to 792 protein families occurring in prokaryotes, referred to as the O_2-dependent reaction set. Of the remaining reactions, a set of 3018 prokaryotic reactions linked to prokaryotic protein families, making up the O_2-independent reaction set (Table S1). See Data accessibility for access to protein family data. For protein family annotation, all clustered sequences were blasted against the KEGG database using Diamond 2.0.1 [23]. All best hits with at least 25% identity and a maximum *e*-value of 1×10^{-10} were used for annotation. Based on these hits, one KO (KEGG orthology identifier) annotation was assigned to each protein family based on majority rule (Table S2). Protein families which contained at least 75% of unknown sequences without any hits in the KEGG database were not annotated.

Verticality distribution of O_2-dependent and O_2-independent reactions

Verticality values (V) from Nagies *et al.* [20] were assigned to protein families (Table S1). Verticality describes the relative amount of LGT that a gene family has undergone. High verticality indicates little LGT (vertical evolution, typical for ribosomal proteins) whereas low verticality (typical for O_2-dependent enzymes) indicates abundant LGT. Only prokaryotic protein families spanning 2 or more prokaryotic phyla had an associated verticality value [20]. All the phylogenetic trees underlying this analysis are published in supplementary table 9 in Nagies *et al.* [20] and available under https://doi.org/10.25838/d5p-15. The distribution of verticality was generated for both the O_2-dependent and the O_2-independent reaction sets. The O_2-independent reaction set contained 3018 reactions and 8322 protein families with associated verticality values, while the O_2-dependent reaction set contained 365 reactions and 547 protein families with associated verticality values. Verticality values of all protein families associated with O_2-dependent and

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

O_2-independent reactions, as well as the average verticality of all protein families associated with each reaction (for O_2-dependent and O_2-independent reactions) were plotted against the number of genomes in which the protein family/reaction was detected (Fig. S1). 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 Fig. S2 and Table S3A).

Calculation of Gibbs energy and reactant count

The change in Gibbs energy ΔG was calculated for each reaction of both O_2-dependent and O_2-independent sets using eQuilibrator API [24] version 0.4.1. 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 ΔG for 288 O_2-dependent reactions and for 2139 O_2-independent reactions (Table S4).

Reactants and products of O_2-dependent reactions were counted, with reactions written in the direction of O_2 consumption. For this count, only two ROS scavenging reactions were written in the O_2-evolving direction based on their physiological function: R00009 and R02670 (annotated as catalase). See Table S5 for the most common substrates and products of O_2-dependent reactions, and Table S6 for an overview of H_2O_2-consuming and H_2O_2-evolving reactions.

Identification of cofactors

Cofactors for each reaction were identified by integrating data from the IUBMB Comments section of the BRENDA database [25], the EC 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 Table S7. The BRENDA database was queried *via* EC 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–sulfur 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. Cosubstrates, such as 2-oxoglutarate in 2-oxoglutarate non-heme iron-dependent oxygenases, were also not listed.

Occurrence of functional categories per prokaryotic phylum

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 (plotted in Fig. S3).

Statistical analysis

The calculated ΔG of the O_2-dependent and O_2-independent sets were compared with a *t*-test [26]. Comparisons of the number of O_2-dependent reactions per genome in aerobes and anaerobes and of genome size between aerobes and anaerobes (classified based on [27]) were done with Welch's *t*-test [26] (see Table S3B). In Fig. S2 the size of the largest genome of each species in the dataset was plotted against the number of O_2-dependent and O_2-independent reactions associated with the species. Regression models were chosen based on Pearson's product–moment correlation resulting in a linear model for O_2-dependent values and a logarithmic model for O_2-independent values (see Table S3A).

To compare the distributions of verticality values of O_2-dependent protein families and O_2-independent protein families, a Welch's *t*-test was performed on an unfiltered (Table S3C) and a filtered (Table S3D) dataset, in which high verticality values (> 1) were removed. This comparison was corroborated by subsampling the protein families associated with O_2-independent reactions. For this subsampling 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 that of the O_2-dependent distribution from values in the O_2-independent distribution. This subsampling procedure was repeated, but changed each time to mitigate potential effects of unequal protein family size distributions. For this, four bins of equal size for protein family sizes (number of genomes in a protein family) of the O_2-dependent reactions were defined in the range of 4 genomes (minimum protein family size to

calculate phylogenetic trees) to 3380 genomes. For each bin, the number of protein families in the O_2-dependent distribution was counted, and during sampling, bins were filled up with an equal number of protein families from the O_2-independent distribution (see Table S3E).

Results

O_2 arose in a world without respiratory oxygen reductases

O_2-dependent enzymes are most frequent in facultative anaerobes with large genomes [28,29] (Tables S3, S8, and S9, Fig. S4). Plots of the frequency of O_2-dependent and O_2-independent reactions in relation to genome size are presented in Figs S2 and S5. Although respiratory oxygen reductases are not the focus of our study, it is important to keep their evolutionary significance in perspective. In today's environment, the overwhelming majority of O_2 produced by cyanobacteria and plants is used for respiration and energy conservation in aquatic and terrestrial environments, keeping the global carbon cycle in balance and our atmospheric O_2 levels largely constant [30], forest fires and human intervention notwithstanding. But that was clearly not the case before the GOE. The goal of our study is to gain insights into the role of O_2 for prokaryotes at the time of the GOE.

For the purpose of this study, we assume that at the time of the first appearance of environmental O_2 during the GOE, functioning respiratory oxygen reductases (or other O_2-dependent enzymes) had not yet evolved. Contrary to that view, it has been proposed that O_2 reductases evolved from a more ancient family of NO reductases, such that oxygen-utilizing terminal oxidases were already present when O_2 first appeared [31]. However, that proposal is not supported by current evidence [32-38], nor have similar proposals been put forward for other O_2-utilizing enzymes. There are three large and evolutionarily unrelated (independently arisen) superfamilies of oxygen reductases in respiratory chains: (a) the heme copper oxidases (HCO) [32,33] that include cytochrome *c* oxidases, (b) the heme-containing cytochrome *bd* oxidases that oxidize membrane quinols [34-36], and (c) the alternative oxidase superfamily (AOX) of non-heme diiron proteins that oxidize quinols [37,38]. Because members of the *bd* oxidase and AOX superfamilies only have one known electron acceptor substrate, O_2 can be directly inferred as the original substrate for the founding members of those enzyme families following the GOE. Only the HCO superfamily contains members that react with another electron acceptor substrate, the nitrous oxide (NO) reductases [27,33]. Recent surveys with broad sampling have clearly shown that NO reductases evolved multiple times and independently to produce the current spectrum of HCO terminal oxidases [33]. This indicates that O_2 was also the ancestral substrate for the HCO family, as in the case of the *bd* oxidases and AOX, in line with the low bioavailability of copper (an essential cofactor of HCO enzymes) prior to the origin of environmental O_2 [32].

This indicates, in turn, that terminal oxidases, like all other O_2-dependent enzymes, arose in environments that were experiencing a gradual encroachment of O_2 into an anoxic world. As O_2 first diffused into the environment, it led to oxidation of one-electron donors and enzyme inhibition, but it also introduced a novel, energy-rich oxidant into the trajectory of biochemical evolution. Once the HCO, *bd*, and AOX families of terminal oxidases arose, they rapidly diversified into new subfamilies [33,35,38]. Moreover, prokaryotes that possessed terminal oxidases came to flourish in heterotrophic settings, amplifying both the gene copy number and protein abundance of terminal oxidases in the environment over evolutionary time. Today terminal oxidases are, in terms of substrate turnover, unchallenged as the main consumers of O_2 in contemporary environments. In modern oceans, for example, photosynthetic oxygen production and respiratory O_2 consumption occur at roughly equal rates [39]. In mammals, 90% of oxygen consumption is mitochondrial respiration [40]. Respiration requires quantitatively large amounts of terminal oxidases, which are often complex membrane proteins.

In the present study our concern is not the increase of protein abundance from the GOE to today, rather, the situation at the onset of O_2 accumulation in Earth history and how O_2 could have impacted enzyme origin in the microbes that first encountered O_2. A situation very similar to that found in terminal oxidases is encountered with another O_2-utilizing (but not strictly O_2-dependent) enzyme, one involved in CO_2 fixation, RuBisCO. Today RuBisCO is not only the most common CO_2 fixing enzyme, it is the most abundant protein in nature [41]. The Calvin cycle in which it operates is the most recently evolved [42], the least energy efficient [43] and by far the most widespread of CO_2-fixing pathways. But it did not start that way. RuBisCO's original function was not in the Calvin cycle but in fermentative breakdown of RNA [44-46]. Changes in Earth's environment, in particular O_2, impacted the evolutionary trajectory of enzymes that play the central role in Earth's modern carbon cycle: terminal oxidases that consume O_2 and RuBisCO that fixes CO_2 with electrons from photosynthetic O_2

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

production. But what was the impact of O_2 on other enzymes and pathways in the immediate wake of the GOE? We asked O_2-dependent enzymes.

O_2 is an energy-rich compound, but unused in SLP

Oxygen is a strong oxidant, with a midpoint potential of +815 mV for the O_2/H_2O pair at pH 7. The enthalpy change in combustion reactions of O_2 with organic compounds (generating CO_2) is typically on the order of $-400\ kJ{\cdot}mol^{-1}$ of O_2, regardless of the organic compound undergoing combustion, because the energy released is almost entirely the energy stored in the energy-rich O_2 molecule, not in the organic reactant [7]. We estimated the free energy change, ΔG, for 288 of the 365 reactions that use O_2 as a substrate and that map to protein families (Table S1). The ΔG values are presented in Table S4 (along with the values estimated for 2139 O_2-independent reactions; see also Fig. S6), with an average calculated change in Gibbs free energy under physiological conditions of $-234\ kJ{\cdot}mol^{-1}$ of O_2. This value is sufficiently exergonic to drive substrate-level phosphorylation (SLP). It is therefore a curious observation that strictly O_2-dependent SLP reactions are virtually unknown. A change in free energy of roughly $-70\ kJ{\cdot}mol^{-1}$ is needed for the synthesis of one ATP [47], hence there is enough energy to synthesize at least one ATP (or more) per O_2-dependent oxidation of organic substrate on average. Out of 365 prokaryotic O_2-dependent reactions, only one (0.3% of the total) generates a product capable of supporting SLP: that catalyzed by the H_2O_2-producing (phosphorylating) pyruvate oxidase (EC 1.2.3.3) [48], POX. This enzyme converts pyruvate and O_2 to CO_2, H_2O_2 and acetyl phosphate with a calculated ΔG of $-158\ kJ{\cdot}mol^{-1}$. H_2O_2-producing pyruvate oxidase has a very narrow phylogenetic distribution, occurring almost exclusively among members of the *Lactobacilli*, and the role of O_2 is to oxidize a flavin in the reaction mechanism, not pyruvate itself. If O_2 were an evolutionary vehicle to improve energy yield—a widely held premise about the role of oxygen in evolution [49,50]—why would cells not conserve the ca. $-234\ kJ{\cdot}mol^{-1}$ of free energy released in > 99% of non-respiratory O_2-reducing reactions? The absence of O_2-dependent SLP reactions suggests that at the onset of the GOE, when O_2 first became available as a substrate, the main initial role of O_2 in microbial evolution was not immediately energy conservation, otherwise cells would likely have found ways to conserve the energy released in reactions of O_2 with organic compounds. The lack of O_2-dependent SLP reactions is noteworthy. If energy was not the initial functional role for oxygen, what then?

Oxygen-dependent enzymes often act on aromatic substrates

Gene-based studies addressing the physiological role of O_2 in evolution typically focus on its use as a terminal electron acceptor in respiratory chains [51,52], but terminal oxidases represent only about 1% of known prokaryotic O_2-utilizing enzyme families [28], hence they depict only one aspect of O_2 impact on prokaryotic metabolism. The responses of anaerobic microbes to O_2 *via* O_2 detoxification enzymes such as NADH oxidases and rubredoxin:oxygen oxidoreductase are well studied [6], as is the impact of O_2 on prokaryotic gene expression *via* DNA-binding proteins [53] such as the FeS cluster-containing O_2 sensor FNR (for fumarate and nitrate respiration) [54] and the ArcAB two-component system (for anoxic redox control or aerobic respiratory control), which responds to the oxidation state of the quinone pool [55]. We found that the most common substrates of O_2-dependent reactions are redox cofactors (ferredoxins, NAD(P)H, and flavins), water and 2-oxoglutarate, while common products include water, H_2O_2, CO_2, and ammonia (Tables S5 and S6).

The main utility of O_2 in biochemical reactions is that of a strong oxidant, affording microbes that possess the corresponding genes metabolic access to chemically stable substrates in the presence of O_2. Reactions of triplet O_2 in biological systems always involve a radical mechanism and almost always involve extraction of one electron from a substrate, generating a radical to initiate the mechanism [56]. The most frequent reaction types of prokaryotic O_2-dependent enzymes in this sample are aromatic degradation and amine oxidation (Table 1). Single electron donor dioxygenases, which incorporate both atoms of O_2 into the reaction product (EC 1.13.11.-), represent the most common enzyme category ($n = 57$), followed by NAD(P)H-dependent monooxygenases ($n = 56$), which incorporate one atom from O_2 into the product (EC 1.14.13.-). The enzymes from both these groups act mainly on aromatic substrates (Table 1). Dioxygenases typically disrupt aromatic rings [57]. The next most common enzymes are 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 [58]. NAD(P)H-dependent dioxygenases (EC 1.14.12.-), all of which act on aromatic substrates [59], are the next most common category ($n = 38$), followed by

Table 1. Most common prokaryotic O_2-dependent reactions. *n*: number of reactions; Dearo: number of dearomatizing reactions (the reaction destroys carbon 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 oxygen atoms incorporated into substrate; Monooxygenase: one oxygen atom incorporated into substrate; Diverse: O_2 can be incorporated or reduced; Acceptor: O_2 is reduced, not incorporated.

EC number	*n*	e^- donor (substrate)	Fate of O_2 atoms	Dearo.	NH_3 prod.	Substr. Arom.
1.13.11.-	57	Single donor	Dioxygenase	40	0	45
1.14.13.-	56	Donor + NAD(P)H	Monooxygenase	4	0	38
1.4.3.-	48	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.-	19	Flavin and Donor	Monooxygenase	0	0	6
1.1.3.-	14	CH-OH group	Acceptor	0	0	3
1.14.11.-	12	Paired donors	Diverse	0	0	2
1.5.3.-	8	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.-[a]	3	Quinones (or cyt *c*)	Acceptor	0	0	3

[a]Given out of order in the ranking, there are EC 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 are provided in Table S10.

FeS-dependent monooxygenases [60] ($n = 28$) (EC 1.14.15.-) (Table 1).

Of the 365 O_2-dependent KEGG reactions that mapped to protein families, more than 50% act on aromatics, stable substrates that require either a strong oxidant or a strong reductant [61] to disrupt the aromatic ring (Table 1). However, aromatic degradation in prokaryotes does not require O_2 [59,61,62]. Two O_2-independent routes of aromatic degradation *via* a benzoyl-CoA intermediate are catalyzed by unrelated benzoyl-CoA reductases. One route employs flavin-based electron bifurcation to generate midpoint potentials sufficiently negative to reduce benzoyl-CoA (E'_0 = –622 mV) [61]. Importantly, the growth rates of aerobic and anaerobic aromatic-degrading microbes are very similar, both having doubling times on the order of 4–6 h [59]. This indicates that the advantage conferred by the O_2-dependent pathway of aromatic degradation is not more rapid growth.

The evolutionary rationale behind the origin of O_2-dependent aromatic degradation might reside in the O_2-sensitivity of the evolutionarily older anaerobic enzymes. The O_2 sensitivity of solvent-exposed FeS clusters is a central underlying theme of O_2 in biochemical evolution [9,63–71]. The anaerobic ATP-dependent benzoyl-CoA reductase from *Thauera aromatica* contains three O_2-sensitive FeS clusters [72], while the electron-bifurcating benzoyl-CoA reductase enzyme complex from *Geobacter metallireducens* contains over 50 FeS clusters [61]. The more ancient, O_2-sensitive enzymes cannot function in oxic environments, requiring the origin of enzymes with an alternative reaction mechanism that can operate in oxic habitats [73]. The solvent-exposed nature of FeS clusters is important for their inhibition by O_2. Yet many FeS clusters in proteins are not solvent-exposed and not O_2-sensitive, such that the mere presence of an FeS cluster in a protein is not always a proxy for O_2 sensitivity. Human mitochondrial complex I contains eight FeS clusters, and photosystem I contains three, but neither is inhibited by O_2. In order for O_2 to oxidize an FeS cluster, it has to attain physical proximity, which is possible in the case of solvent-exposed clusters.

Oxygen-dependent enzymes confer novel physiological traits

The frequency of each enzyme (protein family) among the genomes in our sample is shown in Table 2 for the 30 most common O_2-dependent enzymes (the full list is given in Table S2). The most widespread O_2-dependent enzymes are either terminal oxidases, or enzymes employed in (cofactor) biosynthesis, detoxification, and substrate mobilization. In primary metabolism, the O_2-dependent biosynthetic reactions present

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2. Most widely distributed prokaryotic O_2-dependent enzymes. KO: identifier of the KEGG orthology group used for protein family annotation; N_g: number of genomes; *V*: verticality value (see Materials and methods). Enzyme function abbreviations are in brackets. Numbers in square brackets correspond to protons pumped (data from references [33,34]). The 7,8-dihydroneopterin oxygenase activity (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 [74].

Enzyme name	EC number	KO	N_g	*V*
7,8-Dihydroneopterin oxygenase	1.13.11.81	K01633	4424	0.90
Cytochrome *bd* ubiquinol oxidase subunit II [0] (Oxphos)	7.1.1.7	K00425	4075	0.63
Cytochrome *bd* ubiquinol oxidase subunit I [0] (Oxphos)	7.1.1.7	K00426	4074	0.57
Catalase (O_2 detoxification)	1.11.1.6	K03781	3485	0.30
L-Aspartate oxidase (NAD+ synthesis; N mobilization)	1.4.3.16	K00278	3388	6.22
Pyridoxamine 5′-phosphate oxidase (PLP synthesis)	1.4.3.5	K00275	3026	0.11
Nitronate monooxygenase (N mobilization)	1.13.12.16	K00459	2974	0.34
Bacterioferritin (Iron storage)	1.16.3.1	K03594	2867	0.52
Glycolate dehydrogenase FAD-linked SU (2-OH acids)	1.1.99.14	K00104	2612	0.36
Cytochrome *o* ubiquinol oxidase subunit I [2] (Oxphos)	7.1.1.3	K02298	2561	0.14
Coproporphyrinogen III oxidase (Heme synthesis)	1.3.3.3	K00228	2473	1.18
Cytochrome *c* oxidase subunit I [4] (Oxphos)	7.1.1.9	K02274	2448	0.86
Cytochrome *c* oxidase subunit II [4] (Oxphos)	7.1.1.9	K02275	2435	0.39
2-polyprenylphenol 6-hydroxylase (UQ synthesis)	1.14.13.240	K18800	2431	1.49
Cytochrome *o* ubiquinol oxidase subunit II [2] (Oxphos)	7.1.1.3	K02297	2406	0.57
4,5-DOPA dioxygenase (Amino acids & aromatics ox.)	1.13.11.-	K15777	2266	1.15
Catalase-peroxidase (substrate oxidation and O_2 detox)	1.11.1.21	K03782	2197	0.14
Superoxide dismutase, Cu-Zn family (O_2 detoxification)	1.15.1.1	K04565	2180	1.07
2-octaprenyl-6-methoxyphenol hydroxylase (UQ synth.)	1.14.13.-	K03185	2157	0.72
Malate dehydrogenase, quinone (TCA cycle)	1.1.5.4	K00116	2095	0.31
Glycine oxidase (Thiamine biosynthesis)	1.4.3.19	K03153	1886	0.97
(S)-2-hydroxy-acid oxidase (2-OH acids, glycolate DH)	1.1.3.15	K11473	1849	0.90
Alkanesulfonate monooxygenase (S mobilization)	1.14.14.5	K00299	1831	0.09
Gamma-glutamyl putrescine oxidase (Amino acid ox.)	1.4.3.-	K09471	1809	0.09
tRNA 5-MAM-2-thiouridine bifunctional protein (tRNA)	1.5.-.-	K15461	1789	0.18
4-Hydroxyphenylpyruvate dioxygenase (Amino acid ox.)	1.13.11.27	K00457	1699	0.38
Alkanesulfonate monooxygenase (S mobilization)	1.14.14.5	K04091	1660	0.02
3-Phenylpropanoate dioxygenase (Amino acid ox.)	1.14.12.19	K00529	1540	0.28
4-Hydroxyphenylacetate 3-monooxygenase (AA ox.)	1.14.14.9	K00484	1539	0.09
Protoporphyrinogen oxidase (Heme synthesis)	1.3.3.4	K00231	1488	0.62
Cytochrome *c* oxidase subunit III [4] (Oxphos)	7.1.1.9	K02276	1433	0.49

post-GOE alternatives to O_2-independent reactions that existed in cells before the GOE. The oxidation of amino acids, sulfonates, or 2-hydroxy acids also present alternatives to preexisting anaerobic pathways.

The same principle holds for the HCO, *bd* and AOX superfamilies of terminal oxidases [27,51,75]. The family of cytochrome *bd* oxidases became integrated into previously-existing anaerobic electron transfer chains that used terminal acceptors such as sulfite and metals, producing a branching from membrane quinols to oxygen that is very common in extant facultatively anaerobic prokaryotes [34,51]. The appearance of HCOs introduced a branching in the electron transfer chains that pivot on *c*-type cytochromes and often use metals as source of electrons [76]. Large-scale phylogenetic studies of the HCO superfamily indicate that the most ancient HCOs are the A-type HCOs [33], which, like *bd* oxidases and AOX, normally do not reduce NO, as outlined above. The cytochrome *bd* oxidases are found among many archaeal groups and might be an ancient lineage of oxygen reductases, but their evolutionary history is, like that of HCOs, complicated by lateral gene transfers [32–38,75,76]. Though the relative ages of terminal oxidase families are not clarified because the protein families arose independently, and because the history of all three families has been affected by LGT [32–34,38], sulfide provides hints. HCO terminal oxidases require copper, which has very low bioavailability in sulfidic environments [75], and are strongly inhibited by sulfide [77,78], whereas *bd* oxidases are sulfide tolerant [79,80] and do not require copper. This would speak in favor of a greater antiquity for the *bd* family. The AOX family is probably the youngest of the three

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

oxygen reductase superfamilies, based on its very low affinity for O_2 [51].

In the present sample of prokaryotes, cytochrome *bd* ubiquinol oxidases appear to be the most common O_2-reducing terminal oxidases (Table 2). These transmembrane enzymes usually oxidize quinols but do not translocate protons across the membrane [34] although they do generate proton motive force *via* scalar protons (the localization of proton-consuming and proton-generating reactions on opposite sides of the membrane).

Another novel class of enzymes that arose in response to O_2 are detoxification enzymes that scavenge ROS (reactive oxygen species) that result from reactions of O_2: catalases [81], catalase-peroxidases [82], superoxide dismutases [83], and other detoxification enzymes [9], some of which are not represented in KEGG. The main cofactors of the ROS scavenging enzymes are metals [9,81–83]. Their main function is detoxification of reactive oxygen species such as the superoxide radical, $O_2^{\cdot-}$, that forms from one-electron transfers during O_2-dependent metabolism, or hydrogen peroxide, H_2O_2. Many novel enzymatic functions that did not exist prior to the existence of O_2 are found in the biosynthesis and degradation of secondary metabolites [28,84,85].

Genes for O_2-dependent enzymes undergo LGT more frequently than others

Evidence for O_2 accumulation in Earth history is geochemical (Fig. 1). Within that temporal framework, genes for O_2-dependent reactions can be inherited vertically, in which case their advantage is realized only within lineages, or they can be inherited *via* lateral gene transfer, in which case their advantage can be transmitted to any lineage. Several gene-based studies of O_2 in evolution address the timing of O_2 appearance [81,83,84] by using molecular clocks to date the age of O_2-dependent enzymes. Dating the appearance of O_2 with phylogenies of O_2-metabolizing enzymes requires, however, that the genes have not been subject to lateral gene transfer during evolution. A recent molecular-clock-based study of oxygen in evolution embraced the vertical view and ascribed the origin of oxygen-dependent enzymes to a single hypothetical lineage called the last universal oxygen ancestor [81]. In the real world of microbial genomes, it is known that all genes in prokaryotic genomes can be transferred and that most, if not all, have been subject to LGT at some point in evolution [20,85–87]. O_2-dependent enzymes are no exception.

If an O_2-dependent enzyme confers the ability to survive in oxic habitats, and if oxic habitats became more widespread across Earth history (Fig. 1), what is useful for one microbe can be useful for another. This predicts that O_2-dependent enzymes should readily spread across lineages *via* LGT during evolution. To quantify the role of LGT in genes for O_2-dependent enzymes, we utilized values of verticality, V, which provide a measure for how often a gene has been transferred between lineages in evolution [20]. Verticality measures only LGT events across major taxonomic boundaries (phyla or domains), evolutionarily recent fine-scale transfers are ignored in the calculation of verticality such that it provides a robust measure of LGT frequency [20]. High values of verticality indicate low levels of LGT for members of a gene family, for example ribosomal proteins, while low values of verticality reflect a high frequency of LGT for a given prokaryotic gene during evolution.

For the 547 protein families of O_2-dependent enzymes in prokaryotes for which we could assign a value of V, the mean verticality or V_{avg} is 0.273 ± 0.626 (avg ± SD). For comparison, the 3018 O_2-independent reactions in the data map to 11 754 protein families, of which 8322 had an associated verticality value (Table S1). The O_2-independent enzymes have a mean verticality of $V_{avg} = 0.781 \pm 1.926$. Although the means overlap, the difference in the two distributions is highly significant at $P = 7.43 \cdot 10^{-47}$ (t-value = −14.92, DF = 1406; Fig. 2). The difference remains significant even when higher verticality values (> 1) are filtered out ($V \leq 1$, t-value = −6.22, DF = 603, $P = 9.02 \cdot 10^{-10}$; see Table S3C,D).

Figure 2 shows that genes for O_2-dependent enzymes underwent more LGT in their evolution (expressed in lower verticality values) than genes for O_2-independent enzymes. Even the more widely distributed protein families and reactions have comparable low verticality (Fig. S1). The excess of more vertical protein families among O_2-independent enzymes (Fig. 2, upper right) is the signal of vertically evolving enzymes, many of them involved in information processing, such as aminoacyl-tRNA synthetases (for the complete list see Table S1). Genes for ribosomal proteins, though among the most vertically evolving genes in prokaryotes [20], are not plotted in Fig. 2 because their products do not catalyze chemical reactions, hence they are not represented in KEGG.

Figure 3 underscores the small contribution of vertical evolution in prokaryotic genes that code for enzymes involved in O_2-dependent reactions (blue points) relative to genes that are currently in use for molecular systematics studies, including ribosomal and other information processing proteins (black points),

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

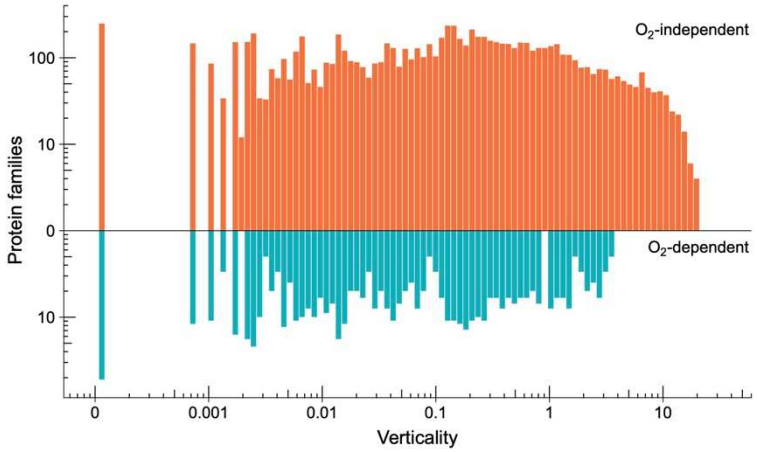


Fig. 2. Distribution of verticality values for protein families catalyzing oxygen-independent and oxygen-dependent reactions. The distribution of 8322 protein families with verticality values associated with O_2-independent reactions are shown in the upper histogram (orange) and the 547 protein families with verticality values associated with O_2-dependent reactions in the lower histogram (blue) in logarithmic scale, 100 bins. O_2-dependent gene families show a lower verticality, which means they have been subject to more LGT. For details of the statistical procedures see Materials and methods and Table S3.

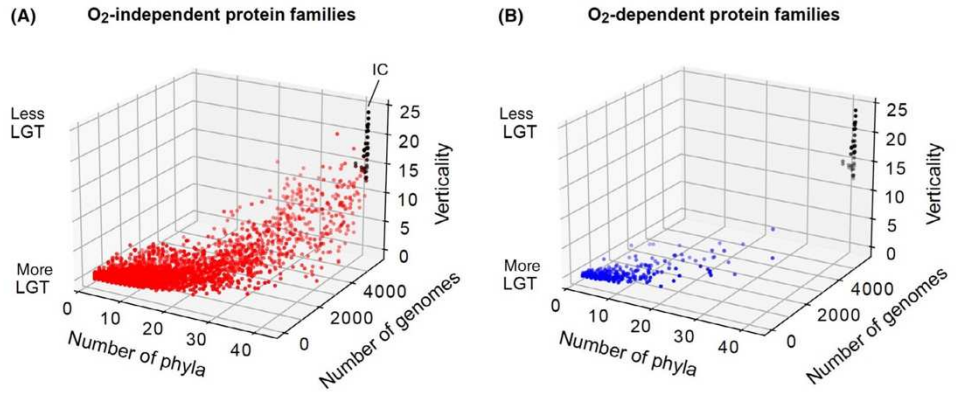
and genes that code for O_2-independent reactions (red points). The genes for O_2-dependent enzymes have almost no tendency at all to undergo vertical evolution, they are freely passed around lineages, where they can become fixed or not.

Within the 10 functional categories in KEGG with the highest frequency of O_2-dependent enzymes, the most frequently transferred O_2-dependent genes relative to O_2-independent genes within the same functional category encode products involved in the metabolism of 'other' amino acids (D-amino acids, glutathione, taurine, selenocompounds, etc.), followed by the functional category 'Protein families: metabolism' (which includes various catabolic reactions), genes coding for enzymes involved in xenobiotics degradation (includes cytochrome P_{450} enzymes), amino acid oxidation and breakdown of secondary metabolites (Fig. 3, Figs S3 and S7). In lipid metabolism, fatty acid and carotenoid oxidation, O_2-dependent enzymes are common, as these act on non-fermentable substrates. A summary of transfers across the nodes of a phylogenetic tree is presented in Fig. S8 (see also Table S11).

Evolutionary rationale behind O_2-dependent synthesis of essential cofactors

The most common genes for O_2-dependent enzymes we identified (Table 2) encode steps in the synthesis of essential cofactors including NAD^+, pyridoxal phosphate (PLP), heme, ubiquinone (UQ), and thiamine (Thi). There are also O_2-dependent and O_2-independent biosynthesis pathways for cobalamin and chlorophyll [73,88,89]. Why should organisms evolve O_2-dependent pathways for the biosynthesis of essential cofactors in the presence of preexisting O_2-independent pathways? One suggestion is that O_2-dependent pathways evolved in the presence of the older anaerobic pathways because they are favored for thermodynamic and kinetic reasons [90]. But the observation that microbial growth rates are similar for O_2-dependent and O_2-independent aromatic degradation [59] offers no hints that the thermodynamics or kinetics of the O_2-dependent pathways confer advantage. Why are there two kinds of pathways for seven essential cofactors?

In the case of parallel O_2-dependent and O_2-independent pathways for chlorophyll synthesis, Chew and Bryant [73] reached a conclusion that is probably applicable in most, if not all, cases of O_2-dependent and O_2-independent alternative pathways: "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 oxygen as a substrate, then evolved." In other words, the inactivation



(C)

KEGG functional category	O_2-dependent Protein families	O_2-dependent V_{avg}	O_2-independent Protein families	O_2-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. 3. Genes for enzymes catalyzing O_2-dependent reactions tend to be laterally transferred. Comparison of number of phyla in which a protein family is present, number of genomes where it occurs and verticality values for prokaryotic protein families. (A) The number of phyla (*x*-axis) is plotted in relation to the number of genomes (*z*-axis) and verticality values (*y*-axis) for 36 highly vertical and universal informational core genes (black) and for the 8322 O_2-independent protein families with associated verticality values (red). (B) The number of phyla (*x*-axis) is plotted in relation to the number of genomes (*z*-axis) and verticality values (*y*-axis) for 36 highly vertical and universal informational core genes (black) and the 547 protein families corresponding to oxygen-dependent reactions with associated verticality values (blue). The following functions correspond to the 36 informational core (IC) genes and are sorted by verticality: rpsJ, rpsK, rplA, alaS, rplB, ffh, rpsE, fusA, ftsY, hisS, truB, metG, rplN, pyrG, rpsH, rpsI, prsA, valS, rplE, rplF, tsaD, argS, truA, ychF, pyrH, uppS, ksgA, rpsS, serS, glyA, tuf, ileS, rpsL, eno, pgk, cyxS. (C) Average verticality values for O_2-dependent (blue points) and O_2-independent (red points) protein families across the 10 functional categories with the highest frequency of O_2-utilizing enzymes. The last column shows the ratio of average verticality between O_2-independent and O_2-dependent protein families per functional category. Values for all functional categories can be found in Table S12, and alternative plots showing the average verticality per functional category, including error bars, can be found in Fig. S7.

by O_2 of an anaerobe's preexisting enzyme generated the selection pressure for the evolution of an alternative enzyme that tolerated the presence of O_2, sometimes even by using O_2 in the radical-dependent mechanism. We suggest that the same evolutionary reason probably applies generally to the origin of O_2-dependent pathways for the synthesis of essential cofactors (or other essential functions) in the presence of preexisting O_2-independent pathways.

The O_2-independent pathways for essential cofactors must be older than the O_2-dependent pathways because prokaryotes that existed prior to the origin of oxygenic photosynthesis undoubtedly required and synthesized NAD^+ [91], PLP [92], ubiquinone, UQ [93,94], thiamine [95,96], heme [88], chlorophyll [73] and cobalamin [88,89]. The preexisting O_2-independent routes typically involve enzymes with O_2-sensitive low-potential metal centers, such as [4Fe4S] clusters,

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

O_2-sensitive pathway intermediates, or both. The reactions tend to rely on radical mechanisms, which are known to be common to many strict anaerobes [5]. Although the mechanism of O_2 activation is not known for many oxygen-dependent enzymes, the chemistry of O_2 in biological reactions prescribes that one-electron activation is necessary, which is why O_2 can interfere as an inhibitor of other biological radical reactions by extracting the radical. O_2-dependent biosynthetic pathways for essential cofactors offered microbes the tools needed to colonize oxic habitats, from which they would otherwise have been excluded. Enzymes with mechanisms of O_2 tolerance that do not involve O_2 as a substrate fall outside the scope of this paper, but the evolutionary rationale behind O_2 tolerance would also apply. The principle of oxygen's impact in evolution—inhibition first, then the ability to grow, then respiration—is illustrated in Fig. 4 (and highlighted in the title).

Flavins and iron are the most common cofactors across O_2-dependent reactions

As a consequence of its kinetically stable triplet diradical state, dioxygen has high activation energy barriers [4]. Because organic substrates usually exist in their singlet ground state, direct reactions of triplet oxygen with organic compounds are spin-forbidden [97] such that 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 by 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 [98], or an organic cofactor such as a flavin or a pterin [99]. 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 most common cofactors for the 365 reactions for which data could readily be identified are given in Table 3 (complete list in Table S7).

Physiological reactions of O_2 are radical reactions that require single electron donors to initiate the reaction with the O_2 diradical. This is reflected in the frequency of single electron donors (marker with an asterisk in Table 3) among the cofactors for O_2-dependent reactions sampled here. In some reactions, the substrate can provide the radical. In some reactions, the mechanism and the single electron source are unknown. Some of the cofactors are not directly involved in O_2 activation. The frequencies of the cofactors reflect their occurrence across individual reaction types, not their frequency within a protein or their occurrence in the environment. For example, a reaction catalyzed by a heme copper oxidase containing several hemes and several copper ions in all marine cyanobacteria (or in mammalian mitochondria, not sampled here) produces one count of heme and copper each. Similarly, the most abundant enzyme on Earth, RuBisCO, accounts for the lone occurrence of magnesium in the table.

The most common cofactors in O_2-dependent reactions are 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 flavin monooxygenases such as UbiH and UbiI involved in ubiquinone synthesis, PhzS from the phenazine pathway, or cyclohexanone monooxygenases ChnB. 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 [99,100].

Iron is the most common transition metal cofactor in oxygenases [101]. 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 center and a Rieske iron–sulfur cluster. The latter accepts electrons from the two-electron donor NADH through a reductase component (often flavin-dependent), sometimes *via* an additional electron carrier. The electrons are transferred to the mononuclear iron center in order to activate dioxygen. The proposed mechanism includes rearrangements encompassing several oxidation states of the mononuclear iron center and a variety of radical and non-radical intermediates [102].

Flavins and iron are present individually, and sometimes together, in over 40% of the enzymes in our sample. NAD(P)H is very common as an electron donor (present in 23% of the reactions), providing electrons to oxygenases through flavins or the reductase component of a multi-component enzyme, as in the Rieske non-heme iron oxygenases. Iron–sulfur clusters are present as prosthetic groups in roughly 16% of the enzymes in our sample and about 17% bind the soluble one-electron carrier ferredoxin. Cytochromes as electron donors are rare for these

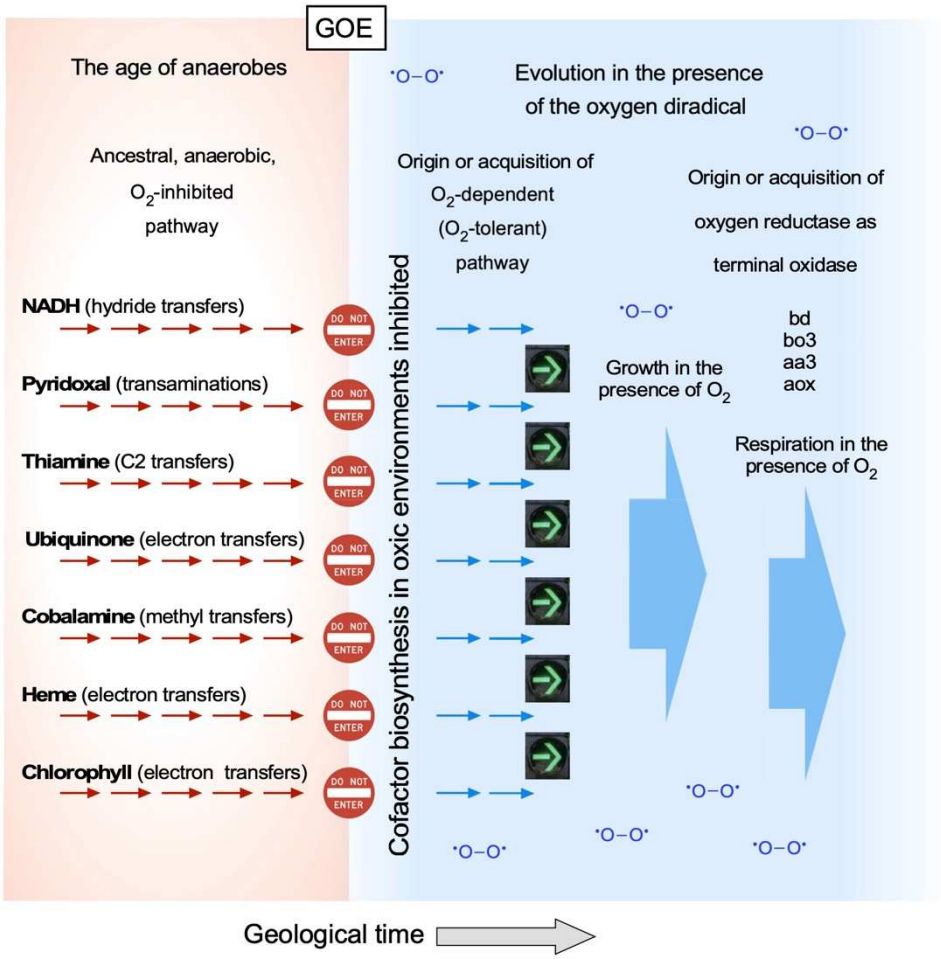


Fig. 4. Inhibition, growth, respiration. Several essential cofactors have O_2-inhibited and O_2-dependent biosynthetic pathways. The anaerobic pathways are older, obviously, because the cofactors shown are essential to various groups of prokaryotic anaerobes that existed prior to the origin of oxygenic photosynthesis. For references to the pathways see text. O_2-inhibited pathways preclude colonization of oxic niches ("do not enter" signs). Auxotrophy for the cofactor in an oxic niche is only an option if another group in the oxic niche has already evolved an O_2-tolerant biosynthetic route, in which case gene acquisition *via* LGT becomes an alternative to auxotrophy. The geological time of origin of oxygen respiration is not an issue here, only the relative timing of cofactor biosynthesis, which enables growth in an oxic environment, followed by origin or acquisition of O_2 respiratory terminal oxidases.

prokaryotic enzymes. The only O_2-dependent *S*-adenosyl methionine (SAM) binding enzyme identified in our sample was MnmC, a bifunctional enzyme found primarily in γ-Proteobacteria that catalyzes a specific modification of the tRNA wobble base by a mechanism including methylation preceded by oxidative cleavage [103].

Mononuclear iron centers and FeS clusters are often inhibited by oxygen [9]. This inhibition is likely the reason for the emergence of alternative pathways following

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3. The most common cofactors across 365 O_2-dependent reactions of prokaryotes. One-electron donors are marked with an (*).

Cofactors	Frequency
Flavins*	156
Iron*	152
NAD(P)+	84
Ferredoxin*	61
Iron–sulfur 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 (unspecified)	1
Thiamine-pyrophosphate	1

the GOE. Why, then, are many alternative oxygen-tolerant enzymes replete with Fe and FeS clusters? The reactivity of oxygen with low-potential metal centers that results in cofactor inactivation in ancient O_2-independent enzymes, whose active sites are not adapted to dealing with oxygen, was not only circumvented, but rather used to good advantage for oxygen activation in the new, alternative pathways: once activated, O_2 can initiate a large number of radical reactions.

The KEGG list of O_2-dependent reactions in prokaryotes includes 21 having no associated cofactors, or where the cofactors were unknown. Several cofactor-independent oxygenases have recently been described [104–108]. Their mechanisms involve a substrate anion, generally generated by base catalysis, that donates a single electron to dioxygen, yielding a stabilized radical pair. The most familiar example of an oxygenase that requires no cofactors to activate dioxygen is the Calvin cycle enzyme RuBisCO, whose oxygenase mechanism involves a substrate-derived radical [107].

No evidence for pre-GOE environmental O_2 in genomes

Evidence for the appearance of O_2 in the atmosphere by the GOE 2.4 billion years ago is universally accepted and uncontroversial [1,2,10,11]. Some studies even date the GOE to a slightly more recent time, 2.3 billion years ago [108]. Several recent reports argue that atmospheric O_2 is older than the GOE. The proposals have it that geochemical interactions between quartz and water generated H_2O_2 early in Earth history and this quartz-derived H_2O_2 could have given rise to O_2 independent of the cyanobacterial OEC [109–111]. From a biological standpoint, an origin of O_2 from H_2O_2 seems extremely unlikely for several reasons. With a standard potential $E_o = +1.7$ V, H_2O_2 is a stronger oxidant than Cl_2 ($E_o = +1.36$ V) or O_2 ($E_o = +0.81$ V), hence a powerful disinfectant and general-purpose biocide, excluding a physical proximity between (so far unobserved) geological H_2O_2 production and life.

At the heart of H_2O_2-dependent abiotic O_2 origin models [109–111] is the chemical reaction catalyzed by catalase,

$$2\,H_2O_2 \rightarrow O_2 + 2\,H_2O$$

which is exothermic by $\Delta H = -101\,kJ{\cdot}mol^{-1}$ of H_2O_2 at 100 °C [111] and exergonic with $\Delta G^{o\prime} = -189.9\,kJ{\cdot}mol^{-1}$ (1 M H_2O_2, 25 °C, pH 7, 1 atm O_2) and $\Delta G_R = -154.2\,kJ{\cdot}mol^{-1}{\cdot}mol^{-1}$ (for 1 mM H_2O_2, 100 °C, pH 7 and 0.1 atm O_2), making it an irreversible reaction under physiological conditions. The same reaction can be catalyzed by bifunctional catalase-peroxidases (Table S6), enzymes that degrade H_2O_2 either as a catalase (reaction above) or alternatively as a peroxidase [112] according to the reaction.

$$H_2O_2 + 2\,H - Substrate \rightarrow 2\,H_2O + 2\,Substrate$$

using the same active site, with ΔG depending on the substrate. Catalase is the most common H_2O_2 detoxifying enzyme in genomes (Table S6). Proponents of a sand/quartz origin of O_2 might argue that the abundance of catalases in genomes (Table S6) provides evidence in favor of an H_2O_2 origin of pre-GOE O_2, with catalase representing a widespread and ancient relict of a time when O_2 was made from H_2O_2. But catalase activity is essentially ubiquitous among prokaryotes, even among strict anaerobes [113], which generally lack O_2-dependent enzymes and pathways that could utilize O_2 (Figs S4 and S5). The function of catalase in strict anaerobes is to detoxify reactive oxygen species generated by one electron carriers during redox metabolism or by other organisms in the environment [77,114], not to provide anaerobes with O_2 from environmentally-supplied H_2O_2.

Furthermore, the proposed quartz-H_2O_2 mechanism of O_2 origin would, in principle, operate before the

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

origin of cells [111], in which case the genetic code, aminoacyl-tRNA synthetases, and ribosome biogenesis would have evolved in the presence of O_2 and H_2O_2. We found no O_2- or H_2O_2-dependent enzymes involved in core information processing functions. Thus, data from genomes and enzymes are distinctly at odds with suggestions for a pre-GOE (or prebiotic) origin of O_2 from H_2O_2 or quartz–water interactions.

Moreover, a recent incisive study by Koppenol and Sies [115] effectively rules out the possibility that abiotic H_2O_2 could have generated O_2 in amounts that would in any way impact the biosphere. They used simple and uncontroversial kinetics to calculate the expected half-life of H_2O_2 in the presence of 20–200 μM iron(II), a reasonable value for Archaean oceans. The key to their reasoning is the classical Fenton reaction, the rapid reaction of H_2O_2 with Fe^{2+} to yield Fe^{3+} and hydroxyl radicals, which rapidly react with almost any organic substrate. They obtain a conservative estimate for the half-life of H_2O_2 in the presence of Fe^{2+} as 0.7 s. Because Fe^{2+} was ubiquitous in all water-bearing environments on the early Earth, including the oceans, the miniscule 0.7 s half-life of H_2O_2 leaves no alternative to the conclusion that "before oxygenic photosynthesis, organisms were not exposed to H_2O_2, $HO^{\cdot}$, O_2^{-}, or CO_3^{-}" [115].

Another line of evidence suggesting pre-GOE O_2 comes from the possible presence of small amounts ("whiffs") of O_2 appearing in rocks 2.5 billion years of age, or ~50 MY prior to the GOE, based on the redox state of redox-sensitive minerals in sediments [116]. The evidence for pre-GOE whiffs of O_2 has recently been debated, however, with a new report indicating that the oxidation state of the molybdenum minerals originally supporting the existence of whiffs resulted from oxidation that took place after the GOE, long after the sediments had been deposited [117]. Those interpretations are challenged [118], and the challenge is in turn rebutted [119], suggesting anoxia (< 1 ppm O_2) in the sediment in question 2.5 billion years ago. From a biological perspective, the difference between oxygen appearance 2.4 billion years ago (the standard GOE model), or 50 million years earlier (the whiffs model), or 50 million years later [108] corresponds to a 2% difference in oxygen age, which has no impact on the verticality or functions (Fig. 3) of O_2-dependent enzymes surveyed here. It merely alters the time at which O_2 started to impact enzyme evolution by a factor of 0.02. In this study, we conservatively take the age of O_2 to correspond to the age of the GOE and we do not use molecular clocks.

Discussion

The main evolutionary impact of the origin of O_2 is traditionally viewed from the standpoint of eukaryote origin [49,84] or from the standpoint of mitochondrial respiration and animal evolution [50,120]. Yet eukaryotes appear roughly 1 billion years after the advent of O_2 [13,121], and animals appear in the fossil record near the base of the Cambrian, almost 2 billion years later than the appearance of O_2 [122,123]. Because both eukaryotes and animals emerged in a world that already contained biologically relevant amounts of O_2, their immediate ancestors were already equipped, enzymatically, to deal both with anoxia and with O_2 as a toxin and/or a substrate [124]. The fossilization of early invertebrates required the existence of O_2 for the synthesis of collagen by proline hydroxylases to generate hard body parts [125]. The origin of large animals also required the presence of O_2 [126]. Life on land was accompanied by adaptations to a permanently oxygenated atmosphere [127].

From the moment it was first produced by cyanobacteria, O_2 became an inhibitor of prokaryotic proteins with solvent-exposed FeS clusters, including ferredoxins, by acting as a one-electron acceptor in redox reactions [9,63]. The appearance of environmental O_2 following the GOE confronted prokaryotic life with a few challenges and numerous chemical opportunities. The main challenge was that a very small number of enzymes in anaerobes—sometimes only one enzyme per species, but often in physiologically key positions in metabolism—are inhibited by exposure to O_2 or activated forms thereof [6,9,63–73]. Such enzymes, for example pyruvate:ferredoxin oxidoreductase, a key enzyme in carbon and energy metabolism of anaerobes [65], or nitrogenase, the biosphere's N_2-fixing enzyme [128], typically have either a radical mechanism [5] or harbor low-potential metal centers, very often FeS centers, that can spontaneously react with the O_2 diradical [71], 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 [6].

The evolutionary significance of O_2 is traditionally viewed in terms of energetic efficiency, O_2 having enabled improved ATP yield from heterotrophic substrate breakdown. The underlying reasoning is often 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" [50]. *Escherichia coli* gains 15 ATP per glucose during O_2

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

respiration under optimal conditions and 4 ATP per glucose from anaerobic fermentation [129], the difference in maximal ATP yield is roughly a factor of 3.8, not 18.

Terrestrial, multicellular eukaryotes are, of course, highly specialized to an O_2-containing atmosphere and possess an O_2-dependent metabolism, yet this specialization took place relatively late in evolution [13,124,127]. Our present investigation deals exclusively with prokaryotes, as they were the only organisms to directly witness the advent of O_2 in evolution. In prokaryotes, O_2 presence does not always mean more energy, and more energy is not always in line with the physiological needs of the cell. The membrane potential, $\Delta\Psi$, used by *E. coli* for substrate import and ATP synthesis under anaerobic growth, −130 mV, is roughly the same as that under aerobic growth, −140 mV [130]. Yeast exhibits a similar response to *E. coli*'s overflow metabolism, called the Crabtree effect, the preference for fermentation in the presence of glucose and oxygen, which results in an increased rate of ATP production at low efficiency [18,131], not in an increased efficiency of ATP production.

In organisms that respire O_2, the diversity of terminal respiratory oxidases exhibits a variety of bioenergetic capacity. Among the enzymes of the HCO superfamily [51,75,76], the *aa3* type cytochrome *c* oxidases and the bo_3 type ubiquinol oxidases, both belonging to family A, pump 4 protons per O_2 reduced, while the cbb_3 type cytochrome *c* oxidases (family C) usually pump 2 protons per O_2 [35,75,132]. Among the growing number of oxidases belonging to the B family of HCO [33], there are subfamilies that pump 4 protons per O_2, various subfamilies that pump 2 protons per O_2 as the prototypic B-family oxidase of *Thermus* [132] and five subfamilies that do not pump protons, also because they do not react with oxygen for the lack of an active-site Tyr, which is involved in the formation of the His-Tyr cofactor in oxygen reductases. Conversely, the *bd* type ubiquinol oxidases and AOX pump 0 protons per O_2 [35,51], although the *bd* oxidases generate protonmotive force [34,35]. The subunits of non-pumping *bd* type oxidases are among the most widespread O_2-dependent enzymes in the present genome sample, apparently more widespread than those of the terminal oxidases that pump 2 or 4 protons per O_2 reduced (Table 2) (see references [33,34] for recent analyses of the HCO and *bd* oxidase superfamilies).

As with the Crabtree effect (2 ATP per glucose *via* cytosolic glycolysis in the presence of O_2), this spectrum of energy conservation within O_2-dependent terminal oxidases from 0 to 4 protons pumped per O_2 reduced suggests that the immediate physiological role of O_2 in evolution was not improved energy yield. Despite the highly exergonic nature of reactions between O_2 and organic compounds, the absence of proton-pumping capabilities in many terminal oxidases [34] and the noteworthy absence of O_2-dependent substrate-level phosphorylation 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 Earth's increasingly oxic habitats because of enzymatic inhibition imposed by O_2.

Lateral gene transfer is the default mode of prokaryotic genome evolution [133]. It has impacted the distribution of O_2-dependent terminal oxidases across prokaryotic lineages [76,134]. Most genes in prokaryotic genomes have undergone LGT, with 97% of all genes having undergone at least one case of LGT between bacteria and archaea [135]. No gene family present in prokaryotic genomes has been completely immune to LGT between prokaryotic phyla during evolution [20]. We found that O_2-dependent enzymes underwent LGT more frequently than O_2-independent enzymes (Figs 2 and 3) and that this is true across almost all functional categories (Fig. 3 and Fig. S7, Table S12). The frequent spread of genes for O_2-dependent enzymes indicates that they conferred a physiological advantage to organisms that retained them. The nature of that advantage falls into three categories: (a) the breakdown of stable bonds in aromatic and nitrogenous compounds thereby mobilizing recalcitrant substrates in oxic environments, (b) the replacement (or supplementation) of enzymes that are inhibited by O_2, and (c) terminal oxidases (Table 2), but terminal oxidases can only be used by an organism once O_2-tolerant pathways of essential intermediate biosynthesis (or auxotrophy) have been incorporated into the genome (Fig. 4).

In primary metabolism, O_2-dependent enzymes have not generated fundamentally new biosynthetic or assimilatory traits. The central pathways of primary metabolism were in place and have remained largely unchanged since the time of the last universal common ancestor LUCA [136]. O_2-dependent enzymes presented alternative, O_2-tolerant routes to preexisting O_2-independent pathways that involved O_2-sensitive intermediates or O_2-sensitive cofactors, in particular solvent-exposed low-potential FeS clusters [64–71,73]. O_2-dependent enzymes are O_2-tolerant by nature, they allowed cells to colonize O_2-containing habitats from which they otherwise would have been excluded [73].

This factor underlies the higher rates of LGT we observe for O_2-dependent reactions. The colonization of oxic habitats not only required the presence of O_2 detoxification mechanisms [9,63], it required the presence of O_2-tolerant pathways for essential cofactor biosynthesis and substrate mobilization. O_2-tolerant enzymes that provide alternatives to the older O_2-sensitive counterparts in essential cofactor biosynthesis had to evolve and be expressible in the genome before the use of O_2 as a terminal acceptor in respiratory chains became an option for any microbe. Stated another way, O_2 is useless for a cell if the cell cannot grow in the presence of O_2 for lack of essential cofactors—NAD^+ [91], PLP [92], thiamine [95,96], quinones [93,94], chlorophyll [73], cobalamin [88,89], or heme [88]—whose anaerobic synthesis pathways are inhibited by O_2.

This selective pressure for the origin of O_2-tolerant and O_2-dependent enzymes, first suggested for chlorophyll biosynthesis [73], appears to apply very generally to O_2-dependent enzymes. Are there glaring exceptions to the rule? There is one, a big one—nitrogenase. Nitrogenase is rapidly inactivated by O_2 through damage to its solvent-exposed FeS clusters [128,137]. Yet in 4 billion years, life has never brought forth an O_2-dependent or O_2-tolerant alternative to Mo, Fe or V-dependent forms of nitrogenase. As a consequence, nitrogenase remained inhibited by atmospheric O_2 at ca. 1% PAL throughout the boring Pasteurian billion [12]. The inhibition of FeS clusters in nitrogenase by O_2 is well studied [128,137]. Nitrogenase inhibition by O_2 is a special case among O_2-dependent enzymes: it is the most important O_2-dependent enzyme that never evolved. Nitrogenase inhibition might have been the limiting factor for the increase in oxygen levels throughout the boring billion, through a negative feedback loop [12]. The essence of that negative feedback loop is that prior to the origin of cellulose synthesis by land plants, O_2 production by cyanobacteria and algae was stoichiometrically linked to synthesis of cell mass with a roughly 5 : 1 C : N ratio. Inhibition of nitrogenase brought biomass synthesis (hence O_2 production) to an immediate halt until environmental O_2 levels once again fell below 1% and nitrogenase could resume activity [12]. Even though nitrogenase inhibition by O_2 acted locally on FeS clusters in a protein, it exhibited an impact as global as cyanobacterial O_2 production itself. Our survey of O_2-dependent enzymes uncovers surprising generality of O_2 inhibition [73] and identifies it as a key mechanism that drove metabolic evolution, enabling prokaryotes to colonize oxic habitats, governing the composition of Earth's atmosphere [12] for almost 2 billion years.

Acknowledgements

We thank Karl Kleinermanns (Düsseldorf) and Yannick De Decker (Brussels) for their calculations of Gibbs free energy change for the catalase reaction. We thank Rebecca E. Gerhards for help with preparing the manuscript. We thank the Zentrum für Informations- und Medientechnologie (ZIM) of the HHU Düsseldorf for providing computational resources. This paper is dedicated to the memory of Dan Tawfik. This work was supported by the European Research Council (ERC) under the Horizon 2020 research and innovation program (grant 101018894 to WFM), the Volkswagen Foundation (grant 96742 to WFM), the Deutsche Forschungsgemeinschaft (grant MA 1426/21-1 to WFM) and the German-Israeli Project Cooperation (DIP) (grant 1426/23-1 to WFM and grant 2476/2-1 to IM). MM is supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic (VEGA) (grant 1/0457/24). Open Access funding enabled and organized by Projekt DEAL.

Author contributions

WFM, NM, FSPN, JLEW and MM conceived the study. WFM supervised the study. WFM, NM, FSPN, JLEW, NK and MRK designed the research. NM, FSPN, JLEW, NK, MRK and KT curated the data. NM, FSPN, JLEW, NK, MRK, KT, LM and NB performed the bioinformatical analysis. WFM, NM, FSPN, JLEW, MRK, KT, LM and NB visualized the data. WFM, JFA, NM, FSPN, JLEW, NK, MRK, KT, LM, NB, MM, MDE and IM analyzed and interpreted the data. WFM, NM, FSPN, JLEW and MRK wrote the original manuscript. WFM, JFA, NM, FSPN, JLEW, NK, MRK, KT, LM, NB, MM, MDE and IM revised and edited the manuscript. WFM and IM acquired funding.

Peer review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/1873-3468.14906.

Data accessibility

All data are available in the main text or the Supporting Information. The protein families used for this

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

analysis are available under https://www.molevol.hhu.de/resources.

References

1 Holland HD (2002) Volcanic gases, black smokers, and the great oxidation event. *Geochim Cosmochim Acta* **66**, 3811–3826.
2 Fischer WW, Hemp J and Johnson JE (2016) Evolution of oxygenic photosynthesis. *Annu Rev Earth Planet Sci* **44**, 647–683.
3 Demoulin CF, Lara YJ, Lambion A and Javaux EJ (2024) Oldest thylakoids in fossil cells directly evidence oxygenic photosynthesis. *Nature* **625**, 529–534.
4 Borden WT, Hoffmann R, Stuyver T and Chen B (2017) Dioxygen: what makes this triplet diradical kinetically persistent? *J Am Chem Soc* **139**, 9010–9018.
5 Buckel W and Golding BT (2006) Radical enzymes in anaerobes. *Annu Rev Microbiol* **60**, 27–49.
6 Lu Z and Imlay JA (2021) When anaerobes encounter oxygen: mechanisms of oxygen toxicity, tolerance and defense. *Nat Rev Microbiol* **19**, 774–785.
7 Schmidt-Rohr K (2015) Why combustions are always exothermic, yielding about 418 kJ per mole of O_2. *J Chem Educ* **92**, 2094–2099.
8 Brewer L (1952) The thermodynamic properties of the oxides and their vaporization processes. *Chem Rev* **52**, 1–75.
9 Khademian M and Imlay JA (2021) How microbes evolved to tolerate oxygen. *Trends Microbiol* **29**, 428–440.
10 Lyons TW, Reinhard CT and Planavsky NJ (2014) The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **50**, 307–315.
11 Lenton TM, Dahl TW, Daines SJ, Mills BJW, Ozaki K, Saltzman MR and Porada P (2016) Earliest land plants created modern levels of atmospheric oxygen. *Proc Natl Acad Sci USA* **113**, 9704–9709.
12 Allen JF, Thake B and Martin WF (2019) Nitrogenase inhibition limited oxygenation of Earth's proterozoic atmosphere. *Trends Plant Sci* **24**, 1022–1031.
13 Mills DB, Boyle RA, Daines SJ, Sperling EA, Pisani D, Donoghue PCJ and Lenton TM (2022) Eukaryogenesis and oxygen in Earth history. *Nat Ecol Evol* **6**, 520–532.
14 Mukherjee I, Large RR, Corkrey R and Danyushevsky LV (2018) The boring billion, a slingshot for complex life on Earth. *Sci Rep* **8**, 4432.
15 Mojzsis SJ, Arrhenius G, McKeegan KD, Harrison TM, Nutman AP and Friend CRL (1996) Evidence for life on Earth before 3,800 million years ago. *Nature* **384**, 55–59.
16 Blaser MB, Dreisbach LK and Conrad R (2013) Carbon isotope fractionation of 11 acetogenic strains grown on H_2 and CO_2. *Appl Environ Microbiol* **79**, 1787–1794.
17 Arndt NT and Nisbet EG (2012) Processes on the young Earth and the habitats of early life. *Annu Rev Earth Planet Sci* **40**, 521–549.
18 Martin WF, Tielens AGM and Mentel M (2020) Mitochondria and Anaerobic Energy Metabolism in Eukaryotes: Biochemistry and Evolution. De Gruyter, Berlin.
19 Kanehisa M and Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30.
20 Nagies FSP, Brueckner J, Tria FDK and Martin WF (2020) A spectrum of verticality across genes. *PLoS Genet* **16**, e1009200.
21 Enright AJ, van Dongen S and Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575–1584.
22 Brueckner J and Martin WF (2020) Bacterial genes outnumber archaeal genes in eukaryotic genomes. *Genome Biol Evol* **12**, 282–292.
23 Buchfink B, Xie C and Huson D (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60.
24 Flamholz A, Noor E, Bar-Even A and Milo R (2012) eQuilibrator—the biochemical thermodynamics calculator. *Nucleic Acids Res* **40**, D770–D775.
25 Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D and Schomburg D (2021) BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res* **49**, D498–D508.
26 Welch BL (1947) The generalization of "Student's" problem when several different population variances are involved. *Biometrika* **34**, 28–35.
27 Sousa FL, Alves RJ, Pereira-Leal JB, Teixeira M and Pereira MM (2011) A bioinformatics classifier and database for heme-copper oxygen reductases. *PLoS One* **6**, e19117.
28 Sousa FL, Nelson-Sathi S and Martin WF (2016) One step beyond a ribosome: the ancient anaerobic core. *Biochim Biophys Acta* **1857**, 1027–1038.
29 Jabøońska J and Tawfik DS (2019) The number and type of oxygen-utilizing enzymes indicates aerobic vs. anaerobic phenotype. *Free Radic Biol Med* **140**, 84–92.
30 Li C, Huang J, Ding L, Liu X, Han D and Huang J (2021) Estimation of oceanic and land carbon sinks based on the most recent oxygen budget. *Earth's Future* **9**, e2021EF002124.
31 Ducluzeau A-L, van Lis R, Duval S, Schoepp-Cothenet B, Russell MJ and Nitschke W (2009) Was nitric oxide the first deep electron sink? *Trends Biochem Sci* **34**, 9–15.
32 Pereira MM, Santana M and Teixeira M (2001) A novel scenario for the evolution of haem–copper oxygen reductases. *Biochim Biophys Acta* **1505**, 185–208.

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

33 Murali R, Hemp J and Gennis RB (2022) Evolution of quinol oxidation within the heme-copper oxidoreductase superfamily. *Biochim Biophys Acta Bioenerg* **1863**, 148907.

34 Murali R, Gennis RB and Hemp J (2021) Evolution of the cytochrome bd oxygen reductase superfamily and the function of CydAA' in Archaea. *ISME J* **15**, 3534–3548.

35 Borisov VB, Gennis RB, Hemp J and Verkhovsky MI (2011) The cytochrome bd respiratory oxygen reductases. *Biochim Biophys Acta Bioenerg* **1807**, 1398–1413.

36 Degli Esposti M, Rosas-Pérez T, Servín-Garcidueñas LE, Bolaños LM, Rosenblueth M and Martínez-Romero E (2015) Molecular evolution of cytochrome bd oxidases across proteobacterial genomes. *Genome Biol Evol* **7**, 801–820.

37 Atteia A, van Lis R, van Hellemond JJ, Tielens AGM, Martin W and Henze K (2004) Identification of prokaryotic homologues indicates an endosymbiotic origin for the alternative oxidases of mitochondria (AOX) and chloroplasts (PTOX). *Gene* **330**, 143–148.

38 Pennisi R, Salvi D, Brandi V, Angelini R, Ascenzi P and Polticelli F (2016) Molecular evolution of alternative oxidase proteins: a phylogenetic and structure modeling approach. *J Mol Evol* **82**, 207–218.

39 del Giorgio PA and Duarte CM (2002) Respiration in the open ocean. *Nature* **420**, 379–384.

40 Rolfe DF and Brown GC (1997) Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiol Rev* **77**, 731–758.

41 Erb TJ and Zarzycki J (2018) A short history of RubisCO: the rise and fall (?) of Nature's predominant CO_2 fixing enzyme. *Curr Opin Biotechnol* **49**, 100–107.

42 Fuchs G (2011) Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu Rev Microbiol* **65**, 631–658.

43 Berg IA, Kockelkorn D, Ramos-Vera WH, Say RF, Zarzycki J, Hügler M, Alber BE and Fuchs G (2010) Autotrophic carbon fixation in archaea. *Nat Rev Microbiol* **8**, 447–460.

44 Aono R, Sato T, Yano A, Yoshida S, Nishitani Y, Miki K, Imanaka T and Atomi H (2012) Enzymatic characterization of AMP phosphorylase and ribose-1,5-bisphosphate isomerase functioning in an archaeal AMP metabolic pathway. *J Bacteriol* **194**, 6847–6855.

45 Aono R, Sato T, Imanaka T and Atomi H (2015) A pentose bisphosphate pathway for nucleoside degradation in Archaea. *Nat Chem Biol* **11**, 355–360.

46 Schönheit P, Buckel W and Martin WF (2016) On the origin of heterotrophy. *Trends Microbiol* **24**, 12–25.

47 Thauer RK, Jungermann K and Decker K (1977) Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol Rev* **41**, 100–180.

48 Muller YA and Schulz GE (1993) Structure of the thiamine- and flavin-dependent enzyme pyruvate oxidase. *Science* **259**, 965–967.

49 Margulis L (1970) Origin of Eukaryotic Cells. Yale University Press, New Haven, CT.

50 Rytkönen KT (2018) Evolution: oxygen and early animals. *Elife* **7**, e34756.

51 Degli Esposti M, Mentel M, Martin W and Sousa FL (2019) Oxygen reductases in alphaproteobacterial genomes: physiological evolution from low to high oxygen environments. *Front Microbiol* **10**, 499.

52 Brochier-Armanet C, Talla E and Gribaldo S (2009) The multiple evolutionary history of dioxygen reductases: implications for the origin and evolution of aerobic respiration. *Mol Biol Evol* **26**, 285–297.

53 Allen JF (1993) Redox control of transcription – sensors, response regulators, activators and repressors. *FEBS Lett* **332**, 203–207.

54 Unden G and Bongaerts J (1997) Alternative respiratory pathways of *Escherichia coli*: energetics and transcriptional regulation in response to electron acceptors. *Biochim Biophys Acta Bioenerg* **1320**, 217–234.

55 Brown AN, Anderson MT, Bachman MA and Mobley HLT (2022) The ArcAB two-component system: function in metabolism, redox control, and infection. *Microbiol Mol Biol Rev* **86**, e00110-21.

56 Fridovich I (1989) Superoxide dismutases. An adaptation to a paramagnetic gas. *J Biol Chem* **264**, 7761–7764.

57 Vaillancourt FH, Bolin JT and Eltis LD (2006) The ins and outs of ring-cleaving dioxygenases. *Crit Rev Biochem Mol Biol* **41**, 241–267.

58 Gaweska H and Fitzpatrick PF (2011) Structures and mechanism of the monoamine oxidase family. *Biomol Concepts* **2**, 365–377.

59 Fuchs G, Boll M and Heider J (2011) Microbial degradation of aromatic compounds – from one strategy to four. *Nat Rev Microbiol* **9**, 803–816.

60 Vanoni MA (2021) Iron-sulfur flavoenzymes: the added value of making the most ancient redox cofactors and the versatile flavins work together. *Open Biol* **11**, 210010.

61 Huwiler SG, Löffler C, Anselmann SEL, Stärk HJ, von Bergen M, Flechsler J, Rachel R and Boll M (2019) One-megadalton metalloenzyme complex in *Geobacter metallireducens* involved in benzene ring reduction beyond the biological redox window. *Proc Natl Acad Sci USA* **116**, 2259–2264.

62 Fuchs G (2008) Anaerobic metabolism of aromatic compounds. *Ann N Y Acad Sci* **1125**, 82–99.

63 Imlay JA (2003) Pathways of oxidative damage. *Annu Rev Microbiol* **57**, 395–418.

64 Schlesier J, Rohde M, Gerhardt S and Einsle O (2016) A conformational switch triggers nitrogenase

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

protection from oxygen damage by Shethna protein II (FeSII). *J Am Chem Soc* **138**, 239–247.
65 Ragsdale SW (2003) Pyruvate ferredoxin oxidoreductase and its radical intermediate. *Chem Rev* **103**, 2333–2346.
66 Boyd ES, Thomas KM, Dai Y, Boyd JM and Outten FW (2014) Interplay between oxygen and Fe-S cluster biogenesis: insights from the Suf pathway. *Biochemistry* **53**, 5834–5847.
67 Khoroshilova N, Popescu C, Münck E, Beinert H and Kiley PJ (1997) Iron-sulfur cluster disassembly in the FNR protein of *Escherichia coli* by O_2: [4Fe-4S] to [2Fe-2S] conversion with loss of biological activity. *Proc Natl Acad Sci USA* **94**, 6087–6092.
68 Pan N and Imlay JA (2001) How does oxygen inhibit central metabolism in the obligate anaerobe Bacteroides thetaiotaomicron. *Mol Microbiol* **39**, 1562–1571.
69 Orme-Johnson WH and Beinert H (1969) On the formation of the superoxide anion radical during the reaction of reduced iron-sulfur proteins with oxygen. *Biochem Biophys Res Commun* **36**, 905–911.
70 Allen JF (1975) A two-step mechanism for the photosynthetic reduction of oxygen by ferredoxin. *Biochem Biophys Res Commun* **66**, 36–43.
71 Imlay JA (2006) Iron-sulphur clusters and the problem with oxygen. *Mol Microbiol* **59**, 1073–1082.
72 Tiedt O, Fuchs J, Eisenreich W and Boll M (2018) A catalytically versatile benzoyl-CoA reductase, key enzyme in the degradation of methyl- and halobenzoates in denitrifying bacteria. *J Biol Chem* **293**, 10264–10274.
73 Chew AGM and Bryant DA (2007) Chlorophyll biosynthesis in bacteria: the origins of structural and functional diversity. *Annu Rev Microbiol* **61**, 113–129.
74 Czekster CM and Blanchard JS (2012) One substrate, five products: reactions catalyzed by the dihydroneopterin aldolase from *Mycobacterium tuberculosis*. *J Am Chem Soc* **134**, 19758–19771.
75 Sousa FL, Alves RJ, Ribeiro MA, Pereira-Leal JB, Teixeira M and Pereira MM (2012) The superfamily of heme-copper oxygen reductases: types and evolutionary considerations. *Biochim Biophys Acta* **1817**, 629–637.
76 Degli Esposti M (2020) On the evolution of cytochrome oxidases consuming oxygen. *Biochim Biophys Acta Bioenerg* **1861**, 148304.
77 Nicholls P (1975) The effect of sulfide on cytochrome aa3. Isosteric and allosteric shifts of the reduced α-peak. *Biochim Biophys Acta* **396**, 24–35.
78 Nicholls P, Marshall DC, Cooper CE and Wilson MT (2013) Sulfide inhibition of and metabolism by cytochrome c oxidase. *Biochem Soc Trans* **41**, 1312–1316.
79 Forte E, Borisov VB, Falabella M, Colaço HG, Tinajero-Trejo M, Poole RK, Vicente JB, Sarti P and Giuffrè A (2016) The terminal oxidase cytochrome bd promotes sulfide-resistant bacterial respiration and growth. *Sci Rep* **6**, 23788.
80 Korshunov S, Imlay KRC and Imlay JA (2016) The cytochrome *bd* oxidase of *Escherichia coli* prevents respiratory inhibition by endogenous and exogenous hydrogen sulfide. *Mol Microbiol* **101**, 62–77.
81 Tehrani HS and Moosavi-Movahedi AA (2018) Catalase and its mysteries. *Prog Biophys Mol Biol* **140**, 5–12.
82 Khmelevtsova LE, Sazykin IS, Azhogina TN and Sazykina MA (2020) Prokaryotic peroxidases and their application in biotechnology (review). *Appl Biochem Microbiol* **56**, 373–380.
83 Bafana A, Dutt S, Kumar A, Kumar S and Ahuja PS (2011) The basic and applied aspects of superoxide dismutase. *J Mol Catal B: Enzym* **68**, 129–138.
84 Raymond J and Segrè D (2006) The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311**, 1764–1767.
85 Jabøońska J and Tawfik DS (2021) The evolution of oxygen-utilizing enzymes suggests early biosphere oxygenation. *Nat Ecol Evol* **5**, 442–448.
86 Dagan T and Martin W (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* **104**, 870–875.
87 Dagan T, Artzy-Randrup Y and Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA* **105**, 10039–10044.
88 Dailey HA, Dailey TA, Gerdes S, Jahn D, Jahn M, O'Brian MR and Warren MJ (2017) Prokaryotic heme biosynthesis: multiple pathways to a common essential product. *Microbiol Mol Biol Rev* **81**, e00048-16.
89 Bryant DA, Hunter CN and Warren MJ (2020) Biosynthesis of the modified tetrapyrroles—the pigments of life. *J Biol Chem* **295**, 6888–6925.
90 Raymond J and Blankenship RE (2005) Biosynthetic pathways, gene replacement and the antiquity of life. *Geobiology* **2**, 199–203.
91 Ollagnier-de Choudensa S, Loiseau L, Sanakis Y, Barras F and Fontecave M (2005) Quinolinate synthetase, an iron–sulfur enzyme in NAD biosynthesis. *FEBS Lett* **579**, 3737–3743.
92 Mukherjee T, Hanes J, Tews I, Ealick SE and Begley TP (2011) Pyridoxal phosphate: biosynthesis and catabolism. *Biochim Biophys Acta Proteins Proteomics* **1814**, 1585–1596.
93 Degli Esposti M (2017) A journey across genomes uncovers the origin of ubiquinone in cyanobacteria. *Genome Biol Evol* **9**, 3039–3053.
94 Pelosi L, Vo CD, Abby SS, Loiseau L, Rascalou B, Hajj Chehade M, Faivre B, Goussé M, Chenal C, Touati N *et al.* (2019) Ubiquinone biosynthesis over

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the entire O_2 range: characterization of a conserved O_2-independent pathway. *MBio* **10**, e01319-19.
95 Leonardi R, Fairhurst SA, Kriek M, Lowe DJ and Roach PL (2003) Thiamine biosynthesis in *Escherichia coli*: isolation and initial characterisation of the ThiGH complex. *FEBS Lett* **539**, 95–99.
96 Settembre EC, Dorrestein PC, Park JH, Augustine AH, Begley TP and Ealick SE (2003) Structural and mechanistic studies on ThiO, a glycine oxidase essential for thiamin biosynthesis in *Bacillus subtilis*. *Biochemistry* **42**, 2971–2981.
97 Klinman JP (2001) Life as aerobes: are there simple rules for activation of dioxygen by enzymes? *J Biol Inorg Chem* **6**, 1–13.
98 Huang X and Groves JT (2018) Oxygen activation and radical transformations in heme proteins and metalloporphyrins. *Chem Rev* **118**, 2491–2553.
99 Romero E, Gómez Castellanos JR, Gadda G, Fraaije MW and Mattevi A (2018) Same substrate, many reactions: oxygen activation in flavoenzymes. *Chem Rev* **118**, 1742–1769.
100 Wongnate T, Surawatanawong P, Visitsatthawong S, Sucharitakul J, Scrutton NS and Chaiven P (2014) Proton-coupled electron transfer and adduct configuration are important for C4a-hydroperoxyflavin formation and stabilization in a flavoenzyme. *J Am Chem Soc* **136**, 241–253.
101 Wang Y, Li J and Liu A (2017) Oxygen activation by mononuclear nonheme iron dioxygenases involved in the degradation of aromatics. *J Biol Inorg Chem* **22**, 395–405.
102 Barry SM and Challis GL (2013) Mechanism and catalytic diversity of Rieske non-heme iron-dependent oxygenases. *ACS Catal* **3**, 2362–2370.
103 Kim J and Almo SC (2013) Structural basis for hypermodification of the wobble uridine in tRNA by bifunctional enzyme MnmC. *BMC Struct Biol* **13**, 1–13.
104 Widboom PF, Fielding EN, Liu Y and Bruner SD (2007) Structural basis for cofactor-independent dioxygenation in vancomycin biosynthesis. *Nature* **447**, 342–345.
105 Frerichs-Deeken U, Ranguelova K, Kappl R, Hüttermann J and Fetzner S (2004) Dioxygenases without requirement for cofactors and their chemical model reaction: compulsory order ternary complex mechanism of 1 H-3-hydroxy-4-oxoquinaldine 2, 4-dioxygenase involving general base catalysis by histidine 251 and single-electron oxidation of the substrate dianion. *Biochemistry* **43**, 14485–14499.
106 Baas BJ, Poddar H, Geertsema EM, Rozeboom HJ, de Vries MP, Permentier HP, Thunnissen AMWH and Poelarends GJ (2015) Functional and structural characterization of an unusual cofactor-independent oxygenase. *Biochemistry* **54**, 1219–1232.
107 Tcherkez G (2016) The mechanism of Rubisco-catalysed oxygenation. *Plant Cell Environ* **39**, 983–997.
108 Luo G, Ono S, Beukes NJ, Wang DT, Xie S and Summons RE (2016) Rapid oxygenation of Earth's atmosphere 2.33 billion years ago. *Sci Adv* **2**, e1600134.
109 He H, Wu X, Xian H, Zhu J, Yang Y, Lv Y, Li Y and Konhauser KO (2021) An abiotic source of Archean hydrogen peroxide and oxygen that pre-dates oxygenic photosynthesis. *Nat Commun* **12**, 6611.
110 He H, Wu X, Zhu J, Lin M, Lv Y, Xian H, Yang Y, Lin X, Li S, Li Y *et al.* (2023) A mineral-based origin of Earth's initial hydrogen peroxide and molecular oxygen. *Proc Natl Acad Sci USA* **120**, e2221984120.
111 Stone J, Edgar JO, Gould JA and Telling J (2022) Tectonically-driven oxidant production in the hot biosphere. *Nat Commun* **13**, 4529.
112 Carpena X, Loprasert S, Mongkolsuk S, Switala J, Loewen PC and Fita I (2003) Catalase-peroxidase KatG of *Burkholderia pseudomallei* at 1.7A resolution. *J Mol Biol* **327**, 475–489.
113 Brioukhanov AL and Netrusov AI (2007) Aerotolerance of strictly anaerobic microorganisms and factors of defense against oxidative stress: a review. *Appl Biochem Microbiol* **43**, 567–582.
114 Harada M, Akiyama A, Furukawa R, Yokobori S, Tajika E and Yamagishi A (2021) Evolution of superoxide dismutases and catalases in cyanobacteria: occurrence of the antioxidant enzyme genes before the rise of atmospheric oxygen. *J Mol Evol* **89**, 527–543.
115 Koppenol WH and Sies H (2024) Was hydrogen peroxide present before the arrival of oxygenic photosynthesis? The important role of iron(II) in the Archean ocean. *Redox Biol* **69**, 103012.
116 Anbar AD, Duan Y, Lyons TW, Arnold GL, Kendall B, Creaser RA, Kaufman AJ, Gordon GW, Scott C, Garvin J *et al.* (2007) A whiff of oxygen before the great oxidation event? *Science* **317**, 1903–1906.
117 Slotznick SP, Johnson JE, Rasmussen B, Raub TD, Webb SM, Zi JW, Kirschvink JL and Fischer WW (2022) Reexamination of 2.5-Ga "whiff" of oxygen interval points to anoxic ocean before GOE. *Sci Adv* **8**, eabj7190.
118 Anbar AD, Buick R, Gordon GW, Johnson AC, Kendall B, Lyons TW, Ostrander CM, Planavsky NJ, Reinhard CT and Stüeken EE (2023) Technical comment on "reexamination of 2.5-Ga 'whiff' of oxygen interval points to anoxic ocean before GOE". *Sci Adv* **9**, eabq3736.
119 Slotznick SP, Johnson JE, Rasmussen B, Raub TD, Webb SM, Zi J-W, Kirschvink JL and Fischer WW (2023) Response to comment on "reexamination of 2.5-Ga 'whiff' of oxygen interval points to anoxic ocean before GOE". *Sci Adv* **9**, eadg1530.
120 Planavsky NJ, Reinhard CT, Wang X, Thomson D, McGoldrick P, Rainbird RH, Johnson T, Fischer WW

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and Lyons TW (2014) Low mid-proterozoic atmospheric oxygen levels and the delayed rise of animals. *Science* **346**, 635–638.
121 Zimorski V, Mentel M, Tielens AGM and Martin WF (2019) Energy metabolism in anaerobic eukaryotes and Earth's late oxygenation. *Free Radic Biol Med* **140**, 279–294.
122 Budd GE (2008) The earliest fossil record of the animals and its significance. *Philos Trans R Soc B* **363**, 1425–1434.
123 Brocks JJ, Nettersheim BJ, Adam P, Schaeffer P, Jarrett AJM, Güneli N, Liyanage T, van Maldegem LM, Hallmann C and Hope JM (2023) Lost world of complex life and the late rise of the eukaryotic crown. *Nature* **618**, 767–773.
124 Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu R-Y, van der Giezen M, Tielens AGM and Martin WF (2012) Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol Mol Biol Rev* **76**, 444–495.
125 Towe KM (1970) Oxygen collagen priority and early metazoan fossil record. *Proc Natl Acad Sci USA* **65**, 781–788.
126 Harrison JF, Kaiser A and van den Brooks JM (2010) Atmospheric oxygen level and the evolution of insect body size. *Proc R Soc B* **277**, 1937–1946.
127 Gould SB, Garg SG, Handrich M, Nelson-Sathi S, Gruenheit N, Tielens AGM and Martin WF (2019) Adaptation to life on land at high O_2 via transition from ferredoxin-to NADH-dependent redox balance. *Proc Biol Sci* **286**, 20191491.
128 Hu Y and Ribbe MW (2015) Nitrogenase and homologs. *J Biol Inorg Chem* **20**, 435–445.
129 Szenk M, Dill KA and de Graff ARM (2017) Why do fast-growing bacteria enter overflow metabolism? Testing the membrane real estate hypothesis. *Cell Syst* **5**, 95–104.
130 Tran QH and Unden G (1998) Changes in the proton potential and the cellular energetics of *Escherichia coli* during growth by aerobic and anaerobic respiration or by fermentation. *Eur J Biochem* **251**, 538–543.
131 Pfeiffer T and Morley A (2014) An evolutionary perspective on the Crabtree effect. *Front Mol Biosci* **1**, 00017.
132 Han H, Hemp J, Pace LA, Ouyang H, Ganesan K, Roh JH, Daldal F, Blanke SR and Gennis RB (2011) Adaptation of aerobic respiration to low O_2 environments. *Proc Natl Acad Sci USA* **108**, 14109–14114.
133 Arnold BJ, Huang IT and Hanage WP (2021) Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol* **20**, 206–218.
134 Osborne JP and Gennis RB (1999) Sequence analysis of cytochrome bd oxidase suggests a revised topology for subunit I. *Biochim Biophys Acta* **1410**, 32–50.
135 Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S and Martin WF (2016) The physiology and habitat of the last universal common ancestor. *Nat Microbiol* **1**, 1–8.
136 Wimmer JLE, Xavier JC, Vieira ADN, Pereira DPH, Leidner J, Sousa FL, Kleinermanns K, Preiner M and Martin WF (2021) Energy at origins: favorable thermodynamics of biosynthetic reactions in the last universal common ancestor (LUCA). *Front Microbiol* **12**, 793664.
137 Jasniewski AJ, Sickerman NS, Hu Y and Ribbe MW (2018) The Fe protein: an unsung hero of nitrogenase. *Inorganics* **6**, 25.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Correlation of protein family verticality and frequency among prokaryotic genomes.
Fig. S2. Correlation of reaction frequency versus genome size.
Fig. S3. Distribution of KEGG functional categories within prokaryotic phyla associated with protein families catalyzing oxygen-dependent reactions.
Fig. S4. Taxonomic distribution of O_2-dependent and O_2-independent reactions.
Fig. S5. Reaction frequency versus genome size for O_2-dependent reactions in prokaryotes.
Fig. S6. Distribution of Gibbs energy ΔG for oxygen-dependent and oxygen-independent reactions.
Fig. S7. Alternative plots of the average verticality across the 10 functional categories with the highest frequency of O_2-utilizing protein families.
Fig. S8. Origins of O_2-dependent reactions across a backbone phylogeny.
Table S1. List of O_2-dependent and O_2-independent reactions from KEGG that were linked to protein families.
Table S2. All protein families with respective KEGG Orthology identifier (KO), corresponding reactions, protein family size and verticality *V*.
Table S3. Statistical tests with relevant parameters.
Table S4. Gibbs energy ΔG for O_2-dependent and -independent reactions.
Table S5. Most common reactants and products across 365 O_2-dependent reactions of prokaryotes.
Table S6. Most frequent H_2O_2 – O_2 interconverting enzymes among 365 O_2-dependent reactions.
Table S7. Cofactors for 365 O_2-dependent reactions.
Table S8. O_2-dependent reactions per genome across taxonomic groups.

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table S9. Information on the 5655 prokaryotic genomes used in this study.
Table S10. Enzyme commission numbers and reaction type for 365 O_2-dependent reactions.
Table S11. Ancestral state reconstruction for 365 O_2-dependent reactions.
Table S12. Functional categories with a list of reactions they include, number of protein families per category and their respective average verticality.

18733468, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14906 by Universitaet Und Landesbibliot, Wiley Online Library on [06/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

III Oxygen reductase origin followed the great oxidation event and terminated the Lomagundi excursion

Katharina Trost[1], Robert B. Gennis[2], John F. Allen[3], Dan B. Mills [4†] and William F. Martin[1] (2025).

1 Institut für Molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Deutschland

2 Department of Chemistry, University of Illinois Urbana-Champaign, USA, Center for Genomic Sciences, UNAM Campus de Cuernavaca, Mexico

3 Research Department of Genetics, Evolution and Environment, University College London, UK

4 Department of Earth and Environmental science, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, 80333 Munich, Germany

† aktuelle Adresse: Institut für Molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Deutschland

Dieser Artikel wurde am 1. April 2026 in *Biochimica et Biophysica Acta (BBA) – Bioenergetics* Ausgabe 1867 veröffentlicht.

Beitrag von Katharina Trost (Erstautor und Korrespondenz):
Die bioinformatischen Analysen zur Datierung der terminalen Oxidasen sowie der Bestimmung von LGT in den phylogenetischen Bäumen wurden von mir durchgeführt. Die Abbildungen 1 bis 5 wurden von mir erstellt und Teile des initialen Manuskriptes wurden von mir geschrieben. An der Überarbeitung des Manuskripts war ich ebenfalls beteiligt.

BBA - Bioenergetics 1867 (2026) 149575

Contents lists available at ScienceDirect

BBA - Bioenergetics

journal homepage: www.elsevier.com/locate/bbabio

Regular paper

Oxygen reductase origin followed the great oxidation event and terminated the Lomagundi excursion

Katharina Trost [a,*], Robert B. Gennis [b], John F. Allen [c], Daniel B. Mills [d,1], William F. Martin [a]

[a] Department of Biology, Institute for Molecular Evolution, Heinrich Heine University of Duesseldorf, 40225, Duesseldorf, Germany
[b] Department of Chemistry, University of Illinois Urbana-Champaign, USA
[c] Research Department of Genetics, Evolution and Environment, University College London, UK
[d] Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, 80333, Munich, Germany

A B S T R A C T

The history of Earth's atmospheric oxygen is a cornerstone of evolutionary biology. While unequivocal evidence for an increase in atmospheric O_2 marks the Great Oxidation Event (GOE) roughly 2.4 billion years ago, evidence underlying proposals for pre-GOE O_2 accumulation is debated. Here we have investigated the distribution of genes for oxygen reductases, the enzymes that consume O_2 in respiratory chains, across independently generated molecular timescales of prokaryotic evolution. The data indicate that cytochrome *bd*-oxidases, heme-copper oxidases and alternative oxidases arose in the wake of the GOE ca. 2.4 billion years ago, after which the genes were subjected to abundant lateral gene transfer, a reflection of their utility in redox balance and membrane bioenergetics. The data lead us to propose a straightforward four-stage model for O_2 accumulation surrounding the GOE: (i) Negligible O_2 existed prior to the GOE. (ii) Cyanobacterial O_2 production started at the GOE, yet was capped at 2 % [*v*/v] atmospheric O_2, the threshold at which cyanobacterial nitrogenase is inhibited by O_2. (iii) Production of 0.02 atm of O_2 (2 % [*v*/v]) at the GOE buried roughly the entire atmospheric CO_2 inventory, causing sudden enrichment of ^{13}C in dissolved inorganic carbon (the Lomagundi ^{13}C anomaly), through RuBisCO isotope discrimination, without atmospheric O_2 exceeding 2 % [*v*/v]. (iv) High atmospheric ^{12}C at the end of the Lomagundi excursion marks the origin of oxygen reductases, their rapid spread via function in respiratory CO_2 liberation, and the onset of equilibrium between photosynthetic O_2 production and respiratory O_2 consumption at 2 % atmospheric O_2.

1. Introduction

Molecular oxygen, O_2, accumulated in the Earth's atmosphere starting ~2.4 billion years ago (Ga) during the Great Oxidation Event or GOE, as documented by several lines of evidence [1–3]. Among them, heavy stable carbon isotope ratios, $\delta^{13}C$ ($\delta^{13}C = [(^{13}C/^{12}C)_{sample}/(^{13}C/^{12}C)_{standard}] - 1$), in sedimentary rocks serves as a proxy for increased organic carbon burial, which enable the persistence of photosynthetically derived O_2 in Earth's atmosphere [4,5]. Another important indicator of Earth's atmospheric oxygenation are measurements of mass-independent sulfur fractionation, or MIFs, which put a strict upper limit of 10^{-6} present atmospheric level (PAL), or 10^{-7} atm, prior to the GOE [6]. There are, however, reports that traces of atmospheric O_2 accumulation, called "whiffs," commenced slightly earlier than the GOE [7,8]. Those reports have been challenged, however, as newer findings indicate that the whiffs are caused by later oxidation of 2.45 Ga sediment samples that were deposited in the absence of O_2 [9]. Anbar et al. [10] responded to that report and [11] responded in return. There are also reports that synthesis of O_2 from sand could have generated O_2 pre-GOE [12–14], but the proposed mechanism involves the synthesis of H_2O_2, which is too reactive to have contributed to O_2 accumulation on an atmospheric scale [15,16]. The half-life of H_2O_2 is only 0.7 s in the presence of Fe^{2+} [15], which would preclude its role as a source of environmental O_2 or as a possible precursor to H_2O in the evolution of the oxygen evolving complex (OEC) of photosystem II [17]. There are also reports that ocean floor manganese nodules can synthesize O_2 [18], but the nodules in question are formed and deposited with the help of O_2, rendering any such contribution to pre-GOE O_2 production unlikely at best.

Several molecular phylogenetic studies of oxygen-utilizing enzymes [19–23] or enzymes related to oxygen-utilizing pathways [24,25] infer an origin of oxygenic photosynthesis prior to the GOE on the basis of molecular clocks. But such studies entail the assumption of strict vertical inheritance for prokaryotic genes, that is, no lateral gene transfer (LGT) or at most one LGT from an unknown extinct donor [24], whereby it is known that all prokaryotic genes studied to date have been subjected to multiple LGTs during evolution [26], including—and in particular—O_2-dependent enzymes, which are among the most frequently transferred

* Corresponding author at: Department of Biology, Heinrich Heine University of Duesseldorf, 40225, Duesseldorf, Germany.
E-mail address: katharina.trost@hhu.de (K. Trost).
[1] current address: Institute for Molecular Evolution, Heinrich Heine University of Duesseldorf, 40225 Duesseldorf, Germany.

https://doi.org/10.1016/j.bbabio.2025.149575
Received 26 August 2025; Received in revised form 7 November 2025; Accepted 30 November 2025
Available online 1 December 2025
0005-2728/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

genes in prokaryotes [16]. Furthermore, molecular clock studies require the use of geochemical and paleontological calibration points, whereby there is no agreement as to what constitutes reliable evidence for pre-GOE O_2. For example, Davin et al. [22] calibrated their trees assuming that the Fe and U-Th-Pb isotope signatures reported by Satkoski et al. [27,28] represent a hard minimum age for photosynthetic O_2 production by 3.2 Ga, 800 MY before the GOE, whereby reports using chromium isotopes to infer pre-GOE O_2 at 3.0 Ga [29] were challenged based on evidence for later oxidative weathering [30]. Isotope-independent biomarker data supporting the existence of cyanobacteria at 2.7 Ga [31] turned out to be contamination from younger rocks [32]. Using post-GOE prokaryotic fossils as calibration points [33] dated the origin of cyanobacteria to roughly 3 Ga, but no fossil cyanobacteria of that age are known, and fossils once thought to be 3.5 Ga cyanobacteria [34] turned out to be abiotic structures of hydrothermal vents [35]. Finally, the molecular clocks of Jablońska & Tawfik [23] inferred evidence for O_2 before the GOE were not calibrated on geochemical data but using published molecular clocks. If we recall that MIFs put a strict upper limit for O_2 of 10^{-7} atm prior to the GOE [6,36], all reports of pre-GOE O_2 carry the caveat that pre-GOE O_2 production was restricted to a particular local environment, and never accumulated in the atmosphere.

It is possible that, prior to the GOE, soluble Mn served as an evolutionary precursor substrate for the primordial oxygen evolving complex prior to the use of water as electron donor, but in a process that does not produce O_2 [37,38]. There is no question that O_2 became environmentally and physiologically relevant at the GOE [6]. What if there was no rudimentary or locally restricted O_2 production before the GOE, which is possible [39]? What if the GOE is telling it like it was? In a straightforward read of the geochemical record, the appearance of biologically relevant amounts of O_2 on Earth corresponds 1:1 with the GOE. In that case, the GOE marks the maximum age of O_2 respiration by prokaryotes because without the substrate (O_2), the O_2-reducing enzymes of respiratory chains [40,41], and other O_2 dependent enzymes [16] could have no selectable O_2-dependent function. This line of reasoning—that the GOE is the calibration point for the origin of O_2-dependent enzymes—is almost entirely absent in the molecular-based literature on O_2 history, and no molecular dating studies, except of Soo et al. [42], to our knowledge have suggested an origin of O_2 dependent enzymes subsequent to the GOE, that is, molecular dating studies consistently date the origin of O_2 pre-GOE.

The GOE is not, however, a simple event. The end of the GOE is accompanied by the Lomagundi-Jatuli Excursion (LJE, also called the Lomagundi excursion), the largest event of elevated, seawater-derived $\delta^{13}C$ values over the last 3.5 billion years [43,44]. During the LJE, $\delta^{13}C$ values increased to roughly +5 to +10 ‰, indicating, at face value, massive primary production and carbon burial, which under standard geochemical models [3,4,45] corresponds to massive O_2 production (between 12 and 22 times the present atmospheric reservoir; [4]). There is no consensus about the interpretation of the LJE. It could indicate a global event or a series of coastal, shallow water events [45–48] that lasted approximately 100 to 250 Ma, from 2.3 to 2.0 billion years ago [46]. Using standard atmospheric models [3,4,45], the magnitude of $\delta^{13}C$ enrichment at the LJE would imply that O_2 rose from zero pre-GOE to levels greatly exceeding the value of 21 % (*v*/v) in today's atmosphere. There are, however, reasons to doubt that standard atmospheric models apply to the LJE, leaving the cause and impact of the $\delta^{13}C$ anomaly during the LJE, in terms of O_2 levels, an open question [48].

Following the LJE, $\delta^{13}C$ values fall to levels indicating roughly 1–10 % of present atmospheric O_2 levels (PAL) for almost 2 billion years until the appearance of land plants [49–51]. Geochemists debate the reasons for that continued phase of low oxygen [52–57], but the simplest explanation is biological, and enzymatic, in that nitrogenase is inhibited by O_2, and that inhibition limits cyanobacterial growth and O_2 production, on a global scale, until O_2 production by land plants set in ~500 MY ago [16,58–61]. During that time, oxygen reductases arose and spread, also into the eukaryotic lineage via the origin of mitochondria [60,62,63].

On the modern Earth, O_2 consumption by oxygen reductases roughly equals O_2 production [64,65]. Without biological O_2 consumption through respiratory terminal oxidases, O_2 would rise to levels that promote spontaneous combustion in forests. There are four basic types of oxygen reductases that maintain O_2 at 21 % v/v including the cytochrome *bd*-type oxygen reductases (bd), the heme-copper oxygen reductases (HCO), the alternative oxygen reductases (AOX) and the plastoquinol terminal oxidase (PTOX) (Fig. 1c) [40–42,66–70]. The bd-, HCO- types of reductases are known to be highly affected by LGT even between domains (Bacteria and Archaea) and thus are distributed over a wide range of prokaryotes [16,40,41,66,68–71]. The alternative oxygen reductases (AOX) are present in eukaryotes and in marine bacteria [68,72] while PTOX can only be found in photosynthetic organisms including higher plants, alga, diatoms and Cyanobacteria [72–74]. AOX and PTOX are membrane bound quinol reductases but have no role in energy conservation, solely serving the function of maintaining redox balance and avoidance of over reduced quinol pool in the bioenergetic membrane instead [75–77]. The *bd*-type and HCO oxygen reductases conserve energy in the form of proton gradients [40,41] and are likely no older than the GOE [42], having arisen in oxic environments [16]. The HCO family includes the nitric oxide (NO) reductases, which are evolutionarily derived from O_2 oxidase ancestors [40,41,66,68,69,78].

The timing of oxygen reductase origin is an unresolved issue, though the oxygen affinity of *bd*-type, HCO and AOX reductases suggest a sequence of order in their evolution: While *bd*-type oxidases have high oxygen affinity, typically occurring in environments with low O_2-levels, the affinity of HCO and AOX and PTOX oxygen is low, requiring O_2-rich environments for activity [79,80]. Here we investigate the timing of oxygen reductase origin and their spread across prokaryotic lineages by mapping their distributions across time-calibrated phylogenetic trees [81]. Our approach presents a radical departure from previous studies in that (i) we accept the date of the GOE as the earliest possible time of oxygen reductase origin and function, (ii) we accept the existence of LGT in oxygen reductase evolution, and (iii) we use a non-controversial molecular dating scheme for prokaryotic evolution that was generated by third parties and not for the purpose of dating oxygen reductase evolution. The findings highlight physiology surrounding the GOE and uncover a biological model that can account in a surprisingly direct manner for the $\delta^{13}C$ isotope anomaly at Lomagundi-Jatuli excursion as the product of a single cyanobacterial enzyme.

2. Methods

2.1. Prokaryotic time tree

The prokaryotic dated tree of life was obtained from Mahendrarajah et al. [81]. It comprises 863 strains including 350 bacterial, 350 archaeal and 163 eukaryotic genomes.

2.2. Balanced prokaryotic RefSeq dataset

The prokaryotic sequences were downloaded from the Reference Sequence Database (RefSeq) release 223 in May 2024 from the National Center for Biotechnology Information (NCBI; [82]) including 41,210 prokaryotic genomes. To avoid any phylogenetic bias, a balanced sample was generated using the biggest archaeal genome per species and the biggest bacterial genome per family. Additionally, 11 genomes with less than 1000 proteins were filtered out and 9 genomes from organisms that have no cytochromes and which were found by Rosenbaum and Müller [83] were added. In total, the balanced dataset comprises 953 genomes including 552 bacterial and 401 archaeal genomes.

2.3. Oxygen reductases proteins

The set of 265 *bd*-type oxygen reductase sequences were obtained

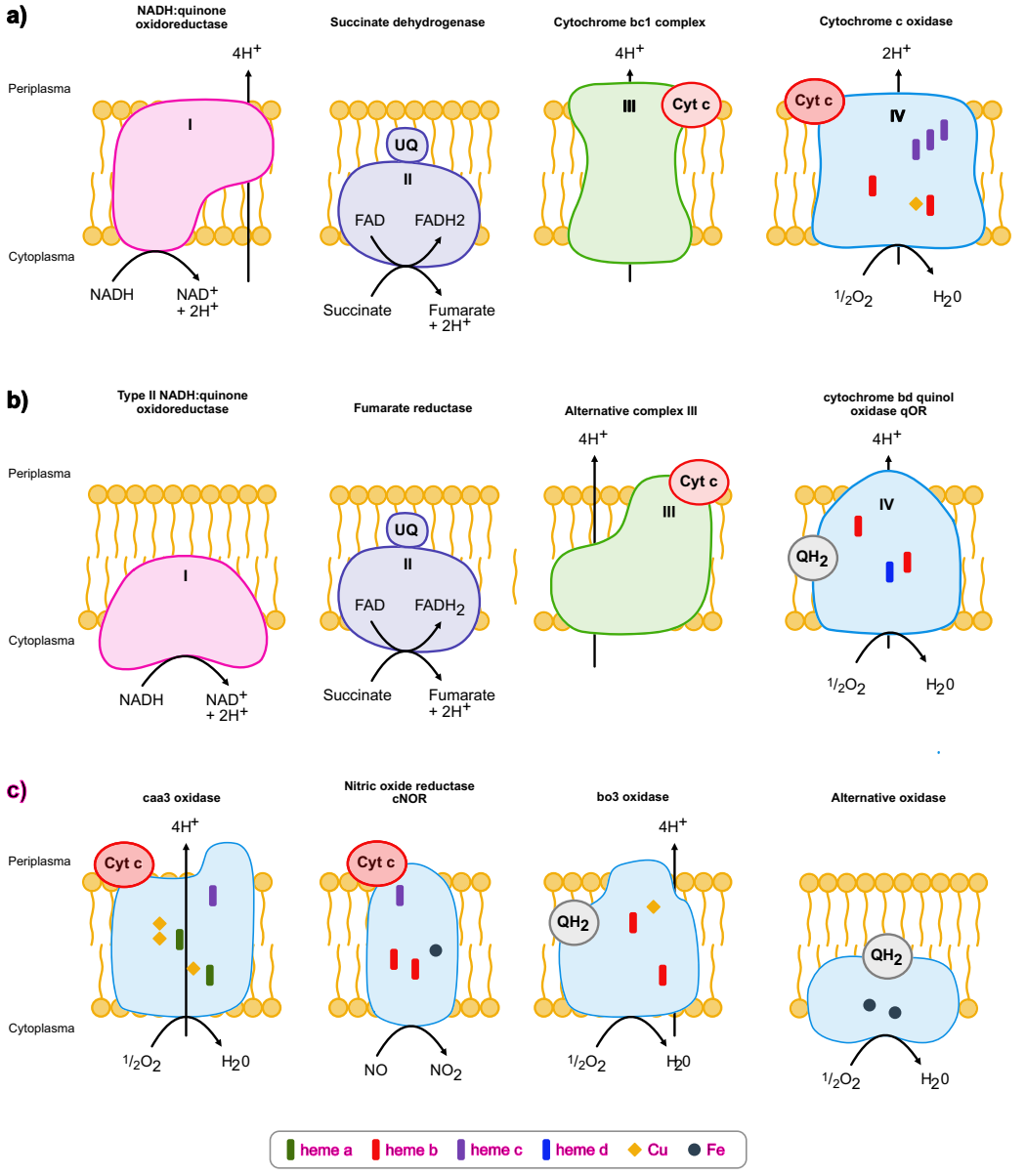


Fig. 1. Components of the respiratory chain and different types of oxygen reductases. A) components of the classical respiratory chain and b) alternative complexes of the respiratory chain. In c) different types of oxygen reductases are shown including the caa3 oxidase (HCO), the nitric oxide reductase cNOR (HCO), the bo3 oxidase (HCO) and the alternative oxidase (AOX).

from Murali et al. [40]. From Murali et al. [41] 35,352 heme-copper oxygen reductase proteins were downloaded. A set of group-specific consensus sequences for alternative oxygen reductase proteins were downloaded from Weaver & McDonald [84] including 21 sequences of eukaryotic and prokaryotic groups. The plastoquinol terminal oxidase was taken form species *Anabaena cylindrica* with the accession number AFZ5900.1, downloaded from NCBI in December 2024.

2.4. Heme biosynthesis and cytochrome b protein sequences

The heme biosynthesis proteins for the protoporphyrin pathway were downloaded from RefSeq Release 227 (NCBI, [82]). The protoporphyrinogen oxidase (PgoX) was obtained from the species *Staphylococcus aureus* and all other protoporphyrin pathway proteins were obtained from species *Klebsiella Pneumniae*. The coproporphyrin pathway proteins were all from *Staphylococcus aureus* and the proteins from the siroheme pathway proteins were from *Methanosarcina barkeri*.

The cytochrome *b* proteins corresponding to the HdrDE complex from *Methanosarcina barkeri* were downloaded from RefSeq Release 227 (NCBI, [82]). As no complete sequences for the proteins of the VhtACG complex could be downloaded from RefSeq, we used hmmer profiles from InterPro [85].

2.5. Presence and absence of oxygen reductase proteins within a dated tree of life

The 265 proteins from *bd*-type oxygen reductase, the 35,352 heme-copper oxygen reductase proteins, the 21 alternative oxygen reductase proteins and the plastoquinol terminal oxidase sequence [40,41,84] were blasted against the balanced prokaryotic RefSeq dataset using Diamond version 2.1.8 [86]. Hits with an e-value $\leq 10E^{-10}$ and local identity $\geq$25 % were retained and cross-checked by protein annotation. Taxa corresponding to strains present in the remaining hits were colored in the dated tree of life using Interactive Tree of Life (iTOL v6, [87]) and the most ancient possible gene origins were calculated based on the sum of branch length of the deepest colored nodes in the dated tree of life. For phylogenetic tree analysis python ETE3 [88] was used.

2.6. Presence and absence of heme biosynthesis and cytochrome b proteins in Methanogens and Halophiles

All heme biosynthesis proteins and proteins of the HdrDE complex including cytochrome *b* were blasted against the genomes of Methanobacteria, Methanococci, Methanopyri, Methanomicrobia, Methanoliparia, Methanonatronarchaeia, Archaeoglobi, Thermoproteota and Halobacteria using Diamond version 2.18 [86]. Hits with an e-value $\leq 10E^{-10}$ and local identity $\geq$25 % were retained and cross-checked by protein annotation. The resulting best hits per protein were used as a proxy for presence or absence within the genome.

HMMER profiles of the VhtACG complex were searched against the genomes of Methanobacteria, Methanococci, Methanopyri, Methanomicrobia, Methanoliparia, Methanonatronarchaeia, Archaeoglobi, Thermoproteota and Halobacteria using HMMER version 3.3.2 (hmmer.org). Only hits with an e-value $\leq 10E^{-10}$ were retained and cross-checked by protein annotation. The best scoring hit per genome was used to infer presence or absence within the genome.

2.7. Monophyly of possible origin groups within oxygen reductase protein trees

Best blast hits per RefSeq genome were defined from the hits generated by the Diamond *blastp* search between reductase proteins and balanced RefSeq dataset for each oxygen reductase (see Taxonomic annotation of oxygen reductase proteins). From these, multiple alignments were made using MAFFT linsi v7.505 [89] and phylogenetic trees were generated using RAxML version 8.2.12 [90] under the PROTCATWAG model. Groups of taxa corresponding to the most ancient possible gene origins were colored within the protein trees and monophyly of these groups were checked using python ETE3 and iTOL v6 [87,88]. Lateral gene transfer events per group and oxygen terminal oxidase were calculated by subtracting one from the number of clades present in the protein tree since one clade has to be the origin and all others are LGTs. To obtain a number of LGT events per terminal oxygenase the values for every group were summed up.

2.8. Statistical tests

Kernel density estimations were made for the distributions of origins of *bd*-type, HCO and AOX reductases. All statistical tests were performed using python. Kolmogorov-Smirnov test was used to compare the distribution of origin ages.

3. Results

3.1. Occurrence of oxygen reductases across prokaryotes

To date the four types of oxygen reductases we used the dated phylogenetic tree with geological time spans as branch lengths constructed by Mahendrarajah et al. [81]. Based on diamond *blastp* [86] searches between protein sequences of *bd* [40], HCO [41], AOX [84] and PTOX reductases and a balanced prokaryotic genome dataset, we colored leaves and corresponding clades of taxa with *bd*, HCO or AOX and PTOX reductases sequences in the phylogenetic time tree (Figs. 2–3, Supplemental Figure 3). Leaves and clades corresponding to eukaryotes are colored in light gray since they were not part of the analysis, as well as taxa that were not present in the balanced prokaryotic dataset and therefore cannot be hit by our blast, as these taxa mainly correspond to metagenomic assemblies (MAGs) that are not represented in our balanced prokaryotic dataset.

Cytochrome *bd* reductases are common in Actinomycetota, Bacilli, Pseudomonadota and Halobacteria and less abundant in Chlorobiota, Clostridia, Fusobacteriota, Spirochaetota Mycoplasmatota, Nitrososphaerota, Thermococci and Thermotogota (Fig. 2). This distribution is consistent with previous studies [40,66], with the exception of the occurrence of *bd* in Thermotogota, where it is however only present in one of the five possible strains (Supplemental Table 1).

HCO reductases are more common in the current data than *bd* oxidases or alternative oxidases (AOX and PTOX). They are distributed across almost all taxonomic groups except for smaller archaeal and bacterial groups including Heimdallarchaeota, Korarchaeota, Nanohaloarchaeota, Aenigmarchaeota, Mycoplasmatota and Synergistota (taxonomy of NCBI as of January 2023). Additionally, we found isolated cases of blast hits for HCO proteins in methanogens, yet only in four strains of Methanomicrobia and one of Methanonatronarchaeia (Fig. 3, Supplemental Table 1). Because (i) all HCOs contain heme and (ii) methanogens are not able to synthesize heme except of some species corresponding to Methanosarcinales, for example *Methanosarcina barkeri* [41,91], we performed Diamond *blastp* searches of heme biosynthesis proteins against methanogens and Halobacteria, to see whether the presence of HCO reductases in Methanomicrobia and Methanonatronarchaeia could be chance similarity or the result of an LGT that does not generate a functional protein (that is, a component of the accessory genome). Among methanogens, only strains of Methanosarcinales encoded a full heme biosynthesis pathway (Supplemental Fig. 1), 96 % of strains of Methanosarcinales in our dataset encoded the three key proteins for the alternative siroheme pathway (Supplemental Table 2). Additionally, we checked whether the sampled methanogens possess the VhtACG and HdrDE protein complexes, which are involved in energy conservation of species of Methanosarcinales and are known to contain cytochrome *b* [92,93]. The complete VhtACG and HdrDE protein complexes were only present in some strains of Methanosarcinales and Methanonatronarchaea (Supplemental Fig. 2). However, the VhtC protein, which includes cytochrome *b*, is also present in Halobacteria, Archaeoglobi, Thermoproteota and Methanocellales. The other cytochrome-containing protein HdrE was only detected in Methanosarcinales (all), one strain of Methanomicrobiales, and the lone Methanonatronarchaeal strain. Based on the absence of heme biosynthesis cytochrome *b* containing protein complexes VhtACG and HdrDE, the occurrence of a putative HCO in the Methanotrichales strain of Methanomicrobia is probably attributable to sequence similarity to other oxidases. Although all methanogens known are strict anaerobes,

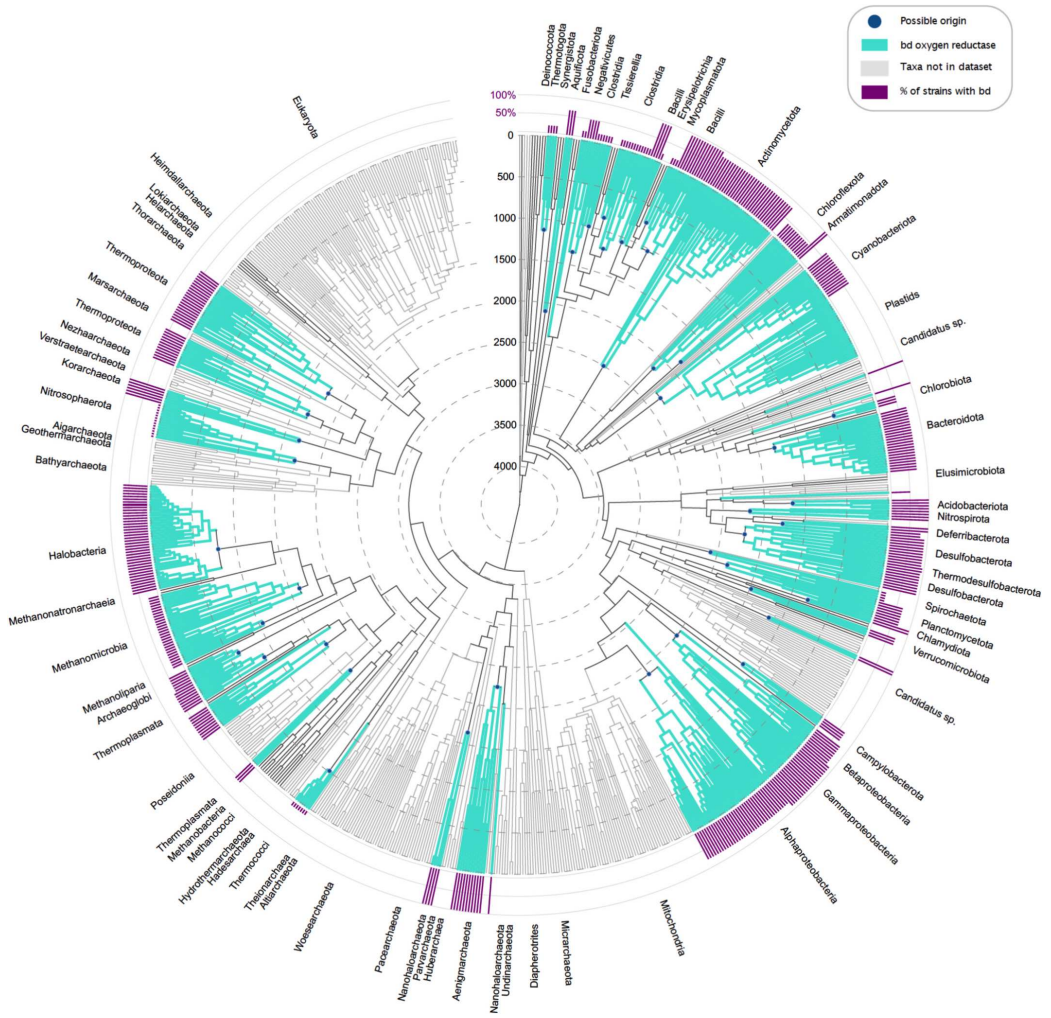


Fig. 2. Occurrence of *bd*-type oxygen reductase in a dated tree of life. Branches in the dated tree of life obtained from Mahendrarajah et al. [81] are colored according to the presence (turquoise) or absence (gray) of *bd*-type oxygen reductase. Eukaryotes were not included in the analysis and are therefore colored in higher gray tones, as are taxa that were not present in the comparative dataset. Dark blue dots at nodes represent possible origins of *bd*-type oxygen reductase. Purple bars represent the percentage of strains within the taxa that have reductases. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

HCO reductases can in principle be present in the three remaining Methanosarcinales strains, though we found no reports of their possible expression or function. Outside the methanogens, HCO reductases are otherwise well known to be present throughout the tree of life, with involvement in both aerobic and anaerobic respiration [41,78,79,94].

Alternative oxidases including AOX, an additional terminal oxidase in mitochondrial electron transport, and PTOX, the plastoquinol terminal oxidase which is the relative enzyme of the photosynthetic electron transport chain [95] are less common in prokaryotes [68,74,84,96]. Consistent with previous analyses, we found AOX reductases only in Pseudomonadota, specifically Alpha-, Beta- and Gammaproteobacteria (Supplemental Fig. 3; [68,84,96]) and PTOX sequences in Cyanobacteriota (Supplemental Fig. 3; [74]). One AOX sequence was also found in Cyanobacterium *Picosynechococcus*, but as this is likely to reflect sequence similarity between AOX and PTOX [74], we excluded this genome for further analysis with AOX.

3.2. Timing the origins and spread of oxygen reductases

To estimate the time of origin for each oxygen reductase, we used the deepest node for each colored clade and calculated the age of the possible origin by summing up the branch lengths. This conservatively

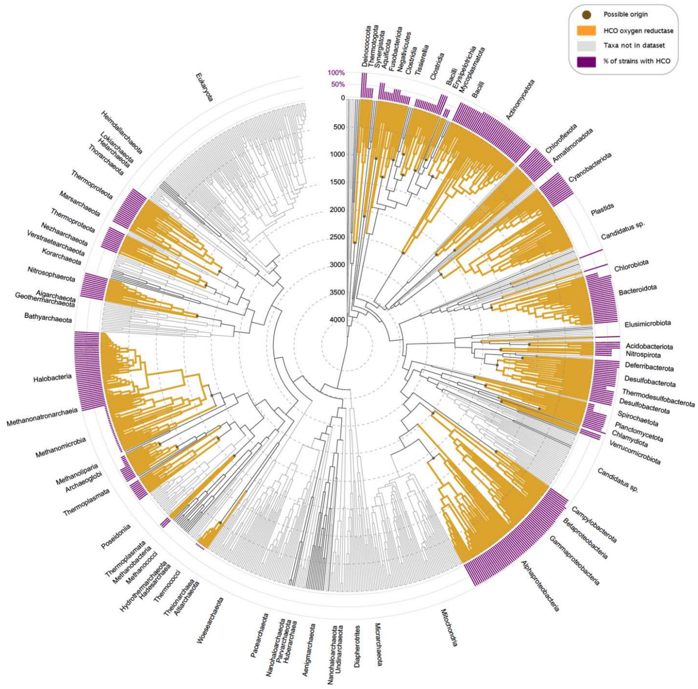


Fig. 3. Occurrence of HCO oxygen reductase in a dated tree of life. Branches in the dated tree of life obtained from Mahendrarajah et al. [81] are colored according to the presence (yellow) or absence (gray) of HCO oxygen reductase. Eukaryotes were not included in the analysis and are therefore colored in higher gray tones, as are taxa that were not present in the comparative dataset. Brown dots at nodes represent possible origins of HCO oxygen reductase. Purple bars represent the percentage of strains within the taxa that have reductases. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

delivers a maximum age for the respective reductases in each clade. For *bd* oxygen reductase we identified 41 possible origins (independent clades) and for HCO 33 possible origins. The AOX and PTOX reductases are the least frequently distributed across the prokaryotic time tree, reflecting only two possible origins for AOX and one origin for PTOX (Supplemental Table 1). The timing of the (earliest) origin of *bd* oxidases and members of the HCO family within a given prokaryotic clade can, with many caveats, be read directly off the timed tree generated by Mahendrarajah et al. [81]. We plotted the distribution of ages for each possible origin on a geological timespan (Fig. 4 and Supplemental Fig. 4). For each distribution except of PTOX (due to the sample size of one) we calculated a Kernel Density Estimation (KDE) to estimate the probability distribution of ages of origins over the entire time period.

What does the age of a *bd* clade or an HCO clade indicate? The *bd* oxidases are all related in sequence, structure and function, they descend from a single common ancestor. We observe, for example, 41 clades of prokaryotes that harbor *bd* oxidase genes. At the one extreme, these 41 clades could be the result of a single *bd* oxidase gene origin in the common ancestor of bacteria and archaea followed by differential loss. This kind of strictly vertical reasoning places all proteins present in some bacteria and some archaea in the last universal common ancestor LUCA. It would place the age of *bd* oxidases at roughly 4 billion years and entail their persistent presence, without oxygen, throughout diverse basal branches in the tree for at least 1.8 billion years, up until the GOE. This kind of "no LGT" scenario calls for geological sources of sustained O_2 production prior to the GOE—controversial sources [12–14,18]—that are however not documented in the geological record, because the first uncontested appearance of biologically useful (respirable) amounts O_2 on Earth *is* the GOE. A "no LGT" model also calls for explanation of why other studies find evidence for substantial amounts of LGT in the evolution of *bd* oxidases and all other prokaryotic genes [40,42,66,84,97].

The other extreme is that only *one* lineage among the 41 *bd* containing clades invented *bd* oxidases and that all other 40 clades are the result of subsequent lateral transfers from the original inventing clade or from secondary spread. That would entail a great deal of LGT in bd oxidase evolution, consistent with recent studies [16]. It would mean that the first origin of *bd* oxidases occurred roughly 2.5 billion years ago (the oldest *bd* origin in the tree, in Actinomycetota), and very close to the GOE (2.4 Ga), within the limits of accuracy on the Mahendrarajah et al. [81] tree. It would entail no requirements for the existence of respirable oxygen prior to the GOE. In fact, this extreme (one origin, 40 LGTs) fits the observations from gene evolution and a straight reading of the geochemical record well, with no need for corollaries.

The ages of the 41 *bd* origins are distributed between 2500 and 510 Ma ago with only one origin before the time of the GOE (Fig. 4, Supplemental Table 1). Since, for the purposes of this paper, we posit that there was no oxygen before the GOE [15,16], the possible origins before the GOE contributing to Actinomycetota (age origin Actinomycetota = 2501 Ma) is likely a result from LGT into the Actinomycetota lineage. All other possible origins are distributed at timespans after the GOE with Cyanobacteriota having the oldest origin (the age of Cyanobacteriota is 2325 Ma in the calibration of Mahendrarajah et al. [81]) with Archaeoglobi (623 Ma), Thermococci (512 Ma) and Chlorobiota (510 Ma) (see Supplemental Table 1 for a list). The KDE for *bd* shows a peak of origins

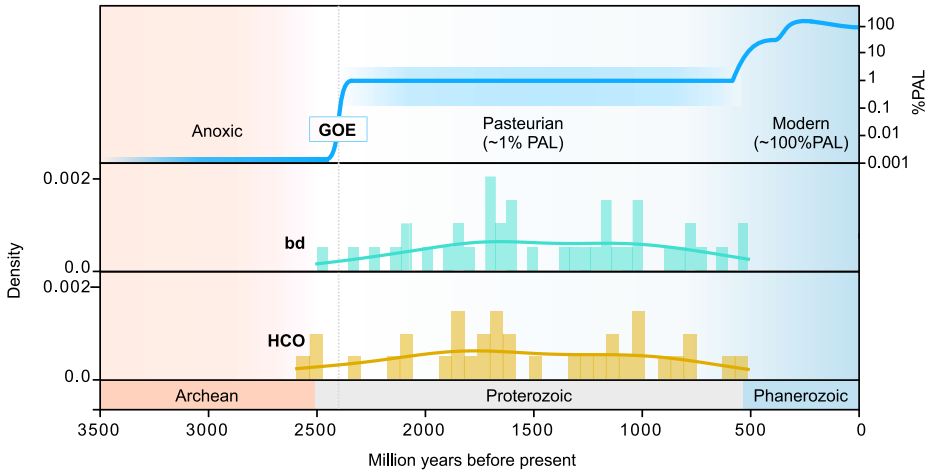


Fig. 4. Distribution of ages per group on geological timescale. Distribution of ages per possible origin (group) within the dated tree of life for the bd-type (turquoise) and HCO (yellow) oxygen reductases. The age [Ma] per group is shown on the x-axis and the corresponding kernel density function (KDE) is placed over the corresponding distribution. The distributions of AOX and PTOX can be found in Supplementary Fig. 4. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

around 1600–1700 Ma which indicates a large number of *bd* oxidase origins in different lineages (spread via LGT) during this time span (Fig. 4). In comparison, the average age of *bd* origins is 1430 Ma, slightly lower than the peak around 1600–1700 Ma (Supplemental Table 3). Due to LGT, many origins of smaller taxonomic groups could affect the average age of origins and thus easily distort it to a lower average age. Still, the peak at 1600–1700 Ma is within the range of average origin age ± one standard deviation (STD, Supplemental Table 3).

The distribution of ages of HCO reductase origins is similar to that of *bd*-type reductases (Kolmogorov-Smirnov Statistic = 0.111, $P = 0.945$). HCO origins are distributed between 512 and 2593 Ma with two possible origins before the GOE (Fig. 4, Supplemental Table 1). These two origins correspond to the taxa Deinococcota, Thermotogota (age = 2593 Ma) and Actinomycetota (age = 2501 Ma). After that, the next origin is located in Beta-, Gamma- and Zetaproteobacteria (age = 2477; Supplemental Table 1) which is consistent with a previous study, suggesting that HCO may originate in basal lineages of Pseudomonadota [98]. Taxa including late possible origins for HCO reductase are Chlamydiota, Archaeoglobi and Thermococci (age origin Chlamydiota = 790 Ma, age origin Archaeoglobi = 623 Ma, age origin Thermococci = 512 Ma; Supplemental Table 1). The KDE has a peak of origin frequency at 1700–1800 Ma, as for *bd*-type reductases, and a second peak around 1000 Ma (Fig. 4). The average age of all HCO origins is at 1523 Ma, again slightly lower as the peak within the KDE. Noticeable for both distributions and KDEs of *bd*-type and HCO reductases is that the origins only occur within the timespan of the Pasteurian billion (also called the boring billion [50,99,100], between 1800 and 800 Ma. Thus, the data indicate that oxygen reductases arose and were spread across prokaryotes (i) after the GOE and (ii) during the time period of low oxygen in Earth history (the Pasteurian billion). Similar results were found for AOX and PTOX (Supplemental Figs. 3–4, Supplemental Table 1).

3.3. Oxygen reductases are strongly affected by LGT

Because *bd*-type and HCO oxygen reductases are known to be subject to frequent transfer by LGT, we tested whether our sample produces similar results as previous studies [40,41,66,68,69,71]. For each reductase we generated a protein tree based on the best blast hits from the balanced RefSeq dataset. The leaves of the protein trees are colored according to their affiliation to groups, representing possible origins in the time tree and were checked whether they are monophyletic or not (Fig. 5, Supplemental Table 4). Reductases were defined as highly affected by LGT if the groups were mainly represented by several clades in the protein tree. In *bd*-type and HCO reductase protein trees, the groups per possible origin are widely spread and usually not monophyletic (Fig. 5a-b). Only three groups are monophyletic in the *bd*-type protein tree including Aenigmarchaeota, Thermococci and Chlamydiota (Fig. 5a, Supplemental Table 4). The HCO reductase protein tree has only one monophyletic group corresponding to the taxon Thermococci (Fig. 5b, Supplemental Table 4), which however contains a maximum of five strains, permitting no strong inference about monophyly.

Despite the small number of genomes and groups in the AOX protein tree, no monophyletic group is found (Fig. 5c). This suggests that the AOX reductase is also transferred via LGT in prokaryotes. However, the transfer of genes is restricted to Pseudomonadota. PTOX reductase do not seem to be affected by LGT. They are found only in Cyanobacteriota, making the protein tree a single monophyletic group (Fig. 5d). The current sample and analysis confirms previous reports for the massive role of LGT in the evolution of *bd*-type, HCO and AOX oxygen reductases [40,41,66,68,69]. One origin and 40 subsequent transfers for *bd* oxidases and one origin plus 32 transfers for HCOs inferred from the species trees (Figs. 2, 3) might seem like a large amount of LGT for oxygen reductases, but the number of transfers inferred from the enzyme phylogenies themselves (Fig. 5.ab) are 124 and 121 respectively, vastly exceeding the bare minimum of 40 (*bd*) or 32 (HCO) transfers needed to account for the lineage distribution of the enzymes.

4. Discussion

There is widespread agreement that the Great Oxidation Event (GOE) marked the persistent accumulation of O_2 in Earth's atmosphere, as documented by several lines of geologic evidence [1,36]. In particular, the onset of the GOE is temporally constrained to ca. 2.32–2.22 based on the irreversible disappearance of mass-independently fractionated sulfur isotopes from the sedimentary record [101–103], interpreted as signaling a rise in atmospheric $O_2 > 10^{-6}$ of present atmospheric levels

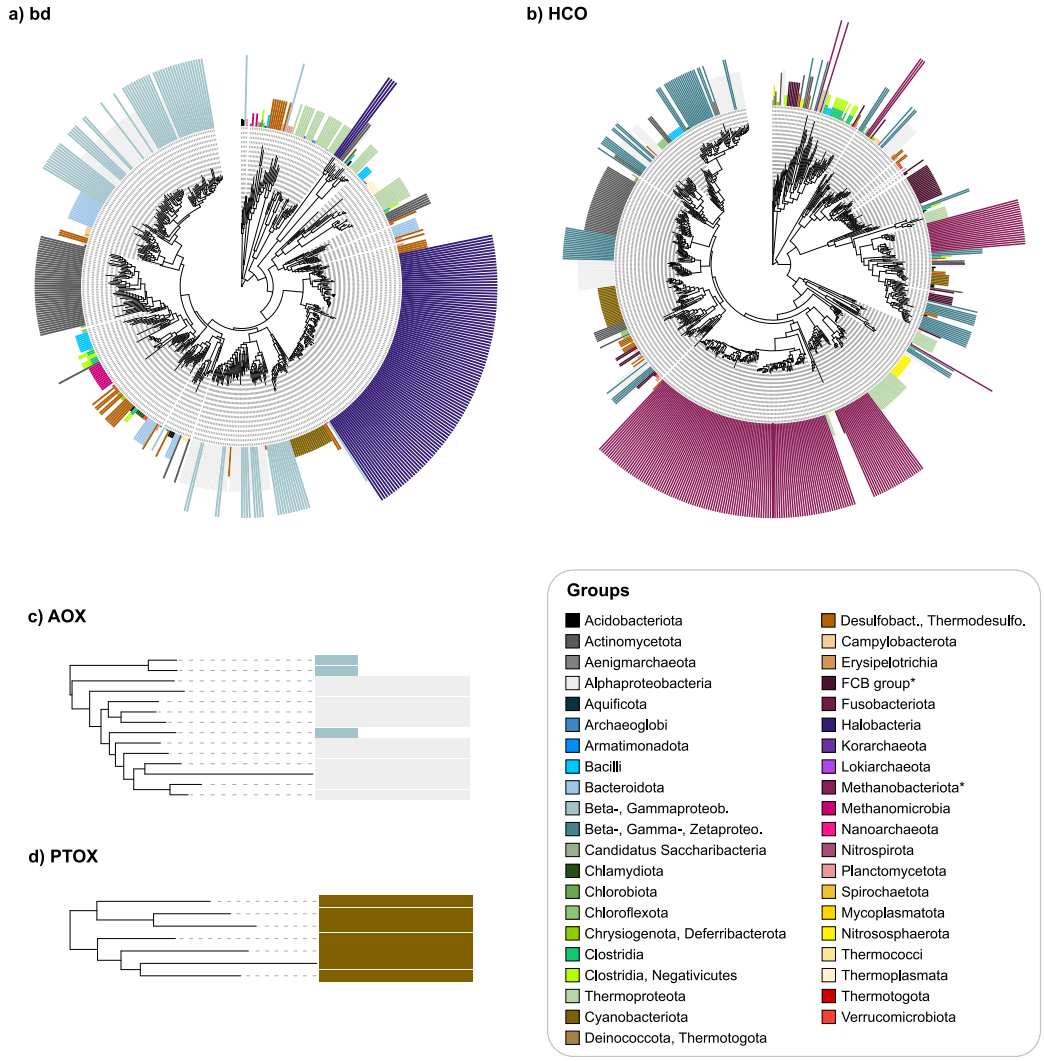


Fig. 5. Size and included taxa of groups within the protein trees of bd-type, HCO, AOX and PTOX oxygen reductase. The leaves of the protein trees for bd-type (a), HCO (b), AOX (c) and PTOX (d) oxygen reductases are colored based on their affiliation to groups, found in the dated tree of life. The sizes of the colored strokes represent the number of strains present in the group. The corresponding taxa included in every group are shown in the right bottom box. *FCB group includes Bacteroidota, Balneolota, Chlorobiota, Rhodothermota and *Methanobacteriota includes Methanomicrobia, Methanonatronarchaeia, Halobacteria

(PAL) [36]. While oxygenic photosynthesis necessarily evolved prior to the GOE, the oldest body fossils interpreted as Cyanobacteria only appear ca. 1.9 Ga [104], leaving geochemical reduction-oxidation (redox) proxies as the primary tools for resolving when environmental O_2 – and, by extension, oxygenic phototrophs – first appeared in Earth's surface environment [1].

Numerous geochemical studies reporting the concentrations of redox sensitive metal concentrations and metal isotope ratios of sedimentary rocks have inferred that oxygenic photosynthesis predated the GOE by up to ca. 600 million years [29,105–107]. Geochemical and mineralogical data associated with the morphology of lacustrine stromatolites have also been used as evidence for oxygenic photosynthesis by ca. 2.7 Ga [108,109]. The conclusion that oxygenic photosynthesis significantly predated the GOE has inspired numerous efforts to explain how photosynthetic O_2 production could have operated on Earth for hundreds of millions of years without oxygenating the atmosphere [110]. The proposed mechanisms vary, but tend to emphasize either enhanced O_2 sinks, such as O_2-consuming reactions with marine and atmospheric reductants [2,36], or diminished O_2 sources, namely extrinsic or intrinsic caps on cyanobacterial primary production, from

phosphorus limitation [111], Fe^{2+} toxicity [112], nitrogenase inhibition by O_2 pre-GOE [113], to low metabolic efficiencies [114]. Despite the ever-growing list of these proposed mechanisms, no clear consensus exists on which one (or combination) of these—if any— actually works as an explanatory platform for advocating for an early origin of oxygenic photosynthesis relative to the GOE.

Although a minority view [36], the simplest explanation for why the GOE happened when it did and not earlier is that oxygenic photosynthesis originated in cyanobacteria only shortly before the GOE [1], and that the rapid rise in O_2 at the GOE simply reflects the rapid (initially exponential) growth of cyanobacteria subsequent to their origin [59]. Collectively, geochemical evidence for free O_2 before the GOE has been criticized as reflecting post-depositional alteration with oxic waters [9,30], and as involving light-driven redox reactions that occurred in the absence of free O_2 [37,39]. Other geochemical evidence from shallow-water banded iron formations has been used to argue that the marine surface and atmosphere contained $<10^{-6}$ PAL O_2 ca. 2.45 Ga, implying that oxygenic photosynthesis had not yet evolved by this time [115]. According to a simple box model, photosynthetic oxygen production could have potentially overwhelmed atmospheric and marine O_2-sinks (e.g., atmospheric H_2 and marine Fe^{2+}) within ca. 100,000 years of its origin [116].

Together, the idea that oxygenic photosynthesis originated only shortly before the GOE arguably represents the simplest and most straightforward reading of the geologic record in the absence of 1) unequivocal evidence for free O_2 (and oxygenic phototrophs) prior to the GOE, and 2) a satisfying explanation for how photosynthetic O_2 production could have operated for over a half-billion years with oxygenating the atmosphere.

Many reports infer the presence of oxygen in earth history from molecular phylogenetic studies [13,19–23], starting with the early study by [117]. Inferences of oxygen in Earth history from gene trees remain contentious because the use of molecular clocks is inapplicable if the gene in question has been affected by lateral gene transfer. All prokaryotic genes have been affected by LGT [26], in particular genes involved in oxygen metabolism [16]. In a molecular clock study, LGT systematically pushes the age of the gene in question artefactually deep, towards the root of the tree. Here we have taken the converse approach in that we allow LGT freely, we use geochemical evidence for the global appearence of oxygen at the GOE as a calibration point for the age of oxygen-dependent respiration, and we plot the appearance of oxygen reductases on a phylogenetic tree constructed from the ATP synthase, a largely vertically inherited gene [81]. The tree that we have used for plotting oxygen reductases was constructed by others as a general timeline reference for prokaryotic evolution, independent of oxygen reductase evolution.

As outlined before, there are isolated reports that trace amounts of oxygen might be synthesized from various reactions prior to the GOE, but these reports are controversial and do not mesh with the evidence for the existence of the GOE [12–16,18]. There are also claims for the occurrence of whiffs of oxygen prior to the GOE [7,8], but the samples in question could have been oxidized post-sedimentation [9], a finding that was rebutted [10] with rebuttal [11] in return. Our reading of the geochemical record is consent with the conservative and straightforward interpretation that the GOE represents the first global appearance of oxygen in Earth history [1,9,102]. We thus interpret the GOE as the earliest time point at which functional O_2 reductases could have arisen. We also assume that LGT occurred freely in the evolution of oxygen reductase genes, consistent with earlier studies [40–42,66,68,69,97] and with the trees of oxygen reductases presented here (Fig. 5). With these simple premises, we find that *bd* oxidase and HCO gene evolution fit more or less perfectly with an origin of oxygen reductases at the GOE, followed by subsequent transfers to different lineages throughout the low oxygen phase of evolution called the Pasteurian billion, because Earth's atmospheric O_2 content was close to the Pasteur point (the O_2 concentration at which facultative anaerobes switch to O_2 respiration) during that time (Fig. 4). The present data do not indicate which lineage invented *bd* oxidases (or HCO), but given the number of subsequent transfers involved, the identity of the *bd*- and HCO-inventing lineages does not impact our findings.

One could argue that Cyanobacteria were the first organisms to evolve oxygen reductases, because they were the first to be confronted with O_2, namely that produced by water-splitting photosynthesis [71]. However, O_2 diffuses out of the cyanobacterial cell faster than it is produced, such that the O_2 concentration in cyanobacterial cells generated by de novo O_2 production is 0.25 μM to 0.025 μM [118]. The O_2 from endogenous production is thus roughly 1000 fold lower than modern concentrations, and well within the Km range of *bd* and HCO enzymes (10 nM to 10 μM, [79]), and sufficient to support the origin of oxygen reductases in cells other than cyanobacteria in Earths' gradually oxygen-accruing environment. As a result, oxygen reductases could have arisen, in principle, in any heme-producing lineage with a preexisting anaerobic respiratory chain.

Prior to the GOE, Earth was inhabited by anaerobes [119]. O_2 is inhibitory for many anaerobes in that it is a stable diradical that can, however, readily accept single electrons from one-electron donors such as quinols, flavins and in particular FeS clusters to generate the $O_2^{\bullet-}$ superoxide radical, a highly reactive oxidant and toxic reactive oxygen species (ROS) [61,120–122]. While flavins, quinols and other cofactors including thiamin [123] generate toxic ROS, they remain active as cofactors upon contact with O_2. By contrast, many FeS clusters undergo oxidative damage upon contact with O_2, such that O_2 inactivates enzymes with surface accessible FeS clusters [61]. Note, however, that many FeS clusters are stable in the presence of O_2, for example the eight FeS clusters in complex I of the mammalian respiratory chain [124]. It has been suggested that the initial function of oxygen reductases, especially *bd*-type oxidases, was to keep the cytosol free of O_2 [125,126], yet for O_2 detoxification, most cells possess dedicated, soluble oxygen-removing and ROS detoxification enzymes, including NADH oxidases and superoxide dismutases [16,19,121,127,128]. In the wake of the GOE, *bd*-type and HCO oxidases could assume their roles in energy conservation, functioning in aerobic respiration in some lineages, in denitrification in others, and in some cases, functioning in biosynthetic pathways [40–42,67].

4.1. The Lomagundi (or Lomagundi-Jatuli) excursion

An aspect of O_2 history that has not been previously addressed by molecular studies is the Lomagundi excursion. More or less concomitant with the GOE, there is a ^{13}C isotope anomaly in the geochemical record called the Lomagundi or Lomagundi-Jatuli excursion [3,48] that designates a ^{13}C enriched marine dissolved inorganic carbon (DIC) pool, which is the sum of dissolved CO_2, HCO_3^- and CO_3^{2-} (Fig. 6). This increase in ^{13}C in the DIC pool indicates increased primary production by oxygenic photosynthesizers, because Rubisco discriminates against $^{13}CO_2$, preferentially incorporating $^{12}CO_2$ into biomass [129], leaving excess ^{13}C in the atmosphere and hence in the DIC pool. Forests during the Carboniferous, for example, deposited CO_2 as photosynthate that became rapidly buried and thus became our modern coal reserves, generating atmosphere O_2 levels on the order of 150 % PAL, which is reflected in high ^{13}C vaules in DIC of the Caboniferous. Today, photosynthetic CO_2 fixation and O_2 respiration take place are roughly equal rates, such that atmospheric O_2 levels are stable [64,65]. It is now agreed that the high ^{13}C at the Lomagundi excursion need not reflect O_2 levels vastly exceeding the present value of 21 % *v*/v [48], but the causes for the appearance and disappearance of the Lomagundi are still debated. Very complicated, multifactorial whole-ecosystem models have been proposed as a cause of the LJE [130] but without identification of specific processes underlying the isotopic excursion. Recent studies have investigated the possibility that Rubisco ^{13}C discrimination might have been higher in the ancient past [131,132] by investigating the discrimination properties of ancestral Rubisco enzymes, but the

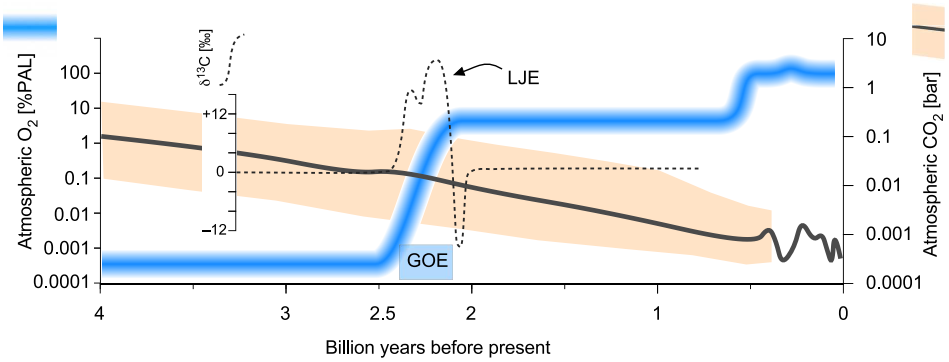


Fig. 6. Atmospheric O_2 and CO_2 during the last 4 billion years in comparison to $\delta^{13}C$ values including the Lomagundi-Jatuli excursion (LJE) and the Great oxidation event (GOE). Comparison of the evolution of $\delta^{13}C$ values (dashed line, [3]), O_2 values (blue line, [60]) and CO_2 values (gray line, [36]) during the last 4 billion years. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

measured effects were small, also in the presence atmospheres containing 2–5 % CO_2, which likely existed around the time of the GOE [36]. Altered properties of ancient Rubisco enzymes are, in principle, a possible cause of the LJE, as are a number of other factors, as outlined by Prave [133].

We consider a sequence of simple processes with few variables at the origin of the LJE, as outlined in Fig. 7. Reading the geochemical record with Occam's razor, there was no cyanobacterial O_2 production prior to the GOE. With the origin of water-splitting photosynthesis, cyanobacteria produced an atmosphere of roughly 2 % oxygen by the end of the LJE and the end of the GOE. There is no explanation in the geochemical record why oxygen stayed flat during the Pasteurian era and nothing existed that limited cyanobacterial growth. However oxygen accumulation ceased at ~2 % and did not exceed ~2 % because nitrogenase is inhibited by 2 % O_2, and without nitrogenase, no net CO_2 fixation (cyanobacterial cell synthesis) is possible [58,59].

Note that nitrogenase is not inhibited by endogenous O_2 production, because O_2 rapidly diffuses out of the oxygen-producing cell, such that endogenous O_2 synthesis generates intracellular O_2 levels of 0.25 μM to 0.025 μM [118], 10 to 100 times lower than that required to inhibit nitrogenase [59]. In oxygenic photosynthesis, one CO_2 is consumed for every O_2 produced. The GOE would have consumed all CO_2 contained in a 2 % CO_2 atmosphere. Even with a modern Rubisco, that CO_2 depletion would be expected to generate a very substantial alteration in the ^{13}C isotope record reflecting high carbonate ^{13}C simply as evidence of increased carbon burial [48,129]. If the atmosphere contained less than 0.02 atm CO_2 at the time of the LJE (Fig. 7) [36], the GOE (which generated 0.02 atm O_2 in the atmosphere) would have essentially scrubbed the atmosphere free of CO_2, bringing O_2 production to a halt, which apparently did not happen (Fig. 7). A 5 % CO_2 atmosphere would have been depleted in CO_2 roughly by half.

One could argue that respiratory processes were replenishing atmospheric CO_2 levels as soon as carbon burial at the GOE commenced. But according to the age of oxygen reductases that we have estimated here, oxygen respiration had either not yet evolved at all at the GOE or had not yet become widespread among bacterial lineages (Fig. 4). In the

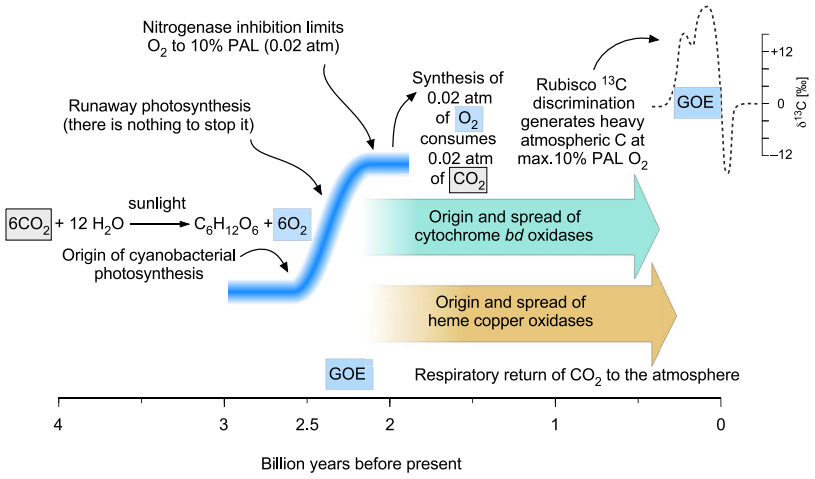


Fig. 7. Model for the causes of Lomagundi-Jatuli excursion (LJE) in connection with the evolution of atmospheric gases as O_2 and CO_2.

absence of *bd* oxidases or HCO in respiratory chains, anaerobic respirations could have returned some CO_2 to the atmosphere. But by the measure of modern CO_2 cycling, the contribution of anaerobic respirations (SO_2, Fe^{3+}) or fermentations would have been modest [64,65], because more than 99 % of biological CO_2 production today comes from O_2 respiration.

The end of the LJE is marked by a sharp spike of low ^{13}C, suggesting, in standard models, rapid release to the DIC pool of sequestered ^{12}C-rich organic material—derived from cells of the newly arisen cyanobacterial lineage in this model. We propose that this rapid release of sequestered organic carbon at the end of the LJE corresponds to the origin of *bd* and heme-copper oxygen reductases and the respiration of a substantial portion of light carbon buried during the GOE. Oxygen levels did not react to the origin and spread of oxygen reductases because nitrogenases imposed an upper on O_2-levels independent of oxygen consumption [58,59].

In this proposal, the LJE indicates a sharp increase in carbon burial at a level sufficient to generate a ^{13}C enrichment in the marine DIC pool, but at no more than 2 % O_2 in the atmosphere, because of nitrogenase inhibition. Furthermore, this proposal entails neither massive export of the greenhouse gas methane to the atmosphere [130], nor does it entail the formation of an ozone layer [130], which under standard models arose long after the GOE, about 600 MY ago [39,134]. Our model requires no attributes of oxygenic photosynthesis or cyanobacterial Rubisco that differ from modern. It does however require an atmospheric CO_2 level (0.02 atm) sufficient to support the synthesis of 0.02 atm of O_2. Following the origin of oxygen reductases at the end of the LJE and the GOE, CO_2 production through respiration and O_2 production through cyanobacterial photosynthesis could have fallen into quantitative balance, as in the modern carbon cycle [64], but in an atmosphere of constant ~2 % O_2 for almost 2 billion years until the origin of land plants [49], because of nitrogenase inhibition [58,59] by O_2.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbabio.2025.149575.

CRediT authorship contribution statement

Katharina Trost: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation. **Robert B. Gennis:** Writing – review & editing. **John F. Allen:** Writing – review & editing. **Daniel B. Mills:** Writing – review & editing, Writing – original draft. **William F. Martin:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Funding acquisition, Conceptualization.

Funding sources

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (grant agreement number 101018894 to W.F.M.) The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf. We thank Nico Bremer for doing the diamond *blastp* analysis on heme-copper oxidase data, Loraine Schwander for providing the clustering of prokaryotes and Natalia Mrnjavac and Ranjani Murali for constructive suggestions.

Data availability

Data will be made available on request.

References

[1] W.W. Fischer, J. Hemp, J.E. Johnson, Evolution of oxygenic photosynthesis, Annu. Rev. Earth Planet. Sci. 44 (2016) 647–683.
[2] H.D. Holland, Volcanic gases, black smokers, and the great oxidation event, Geochim. Cosmochim. Acta 66 (2002) 3811–3826.
[3] T.W. Lyons, C.T. Reinhard, N.J. Planavsky, The rise of oxygen in earth's early ocean and atmosphere, Nature 506 (2014) 307–315.
[4] J.A. Karhu, H.D. Holland, Carbon isotopes and the rise of atmospheric oxygen, Geology 24 (1996) 867–870.
[5] L.R. Kump, M.A. Arthur, Interpreting carbon-isotope excursions: carbonates and organic matter, Chem. Geol. 161 (1999) 181–198.
[6] H. Wang, et al., Two-billion-year transitional oxygenation of the earth's surface, Nature 645 (2025) 665–671.
[7] A.D. Anbar, et al., A whiff of oxygen before the great oxidation event? Science 317 (2007) 1903–1906.
[8] A.J. Kaufman, et al., Late Archean biospheric oxygenation and atmospheric evolution, Science 317 (2007) 1900–1903.
[9] S.P. Slotznick, et al., Reexamination of 2.5-Ga "whiff" of oxygen interval points to anoxic ocean before GOE, Sci. Adv. 8 (2022) eabj7190.
[10] A.D. Anbar, et al., Technical comment on "reexamination of 2.5-Ga 'whiff' of oxygen interval points to anoxic ocean before GOE", Sci. Adv. 9 (2023) eabq3736.
[11] S.P. Slotznick, et al., Response to comment on "reexamination of 2.5-Ga 'whiff' of oxygen interval points to anoxic ocean before GOE", Sci. Adv. 9 (2023) eadg1530.
[12] H. He, et al., An abiotic source of Archean hydrogen peroxide and oxygen that pre-dates oxygenic photosynthesis, Nature 12 (2021) 6611.
[13] H. He, et al., A mineral-based origin of earth's initial hydrogen peroxide and molecular oxygen, Proc. Natl. Acad. Sci. USA 120 (2023) e2221984120.
[14] J. Stone, et al., Tectonically-driven oxidant production in the hot biosphere, Nat. Commun. 13 (2022) 4529.
[15] W.H. Koppenol, H. Sies, Was hydrogen peroxide present before the arrival of oxygenic photosynthesis? The important role of iron(II) in the Archean Ocean, Redox Biol. 69 (2024) 103012.
[16] N. Mrnjavac, et al., Three enzymes governed the rise of O_2 on earth, Biochim. Biophys. Acta Bioenerg. 1865 (2024) 149496.
[17] W.D. Frasch, R. Mei, Hydrogen peroxide as an alternate substrate for the oxygen-evolving complex, Biochim. Biophys. Acta 891 (1987) 8–14.
[18] A.K. Sweetman, et al., Evidence of dark oxygen production at the abyssal seafloor, Nat. Geosci. 17 (2024) 737–739.
[19] A. Bafana, et al., The basic and applied aspects of superoxide dismutase, J. Mol. Catal. B Enzym. 68 (2011) 129–138.
[20] J.S. Boden, et al., Timing the evolution of antioxidant enzymes in cyanobacteria, Nat. Commun. 12 (2021) 4742.
[21] C. Brochier-Armanet, E. Talla, S. Gribaldo, The multiple evolutionary histories of dioxygen reductases: implications for the origin and evolution of aerobic respiration, Mol. Biol. Evol. 26 (2009) 285–297.
[22] A.A. Davín, et al., Geological timescale for bacterial evolution and oxygen adaptation, Science 388 (6742) (2025).
[23] J. Jabłońska, D.S. Tawfik, The evolution of oxygen-utilizing enzymes suggests early biosphere oxygenation, Nat. Ecol. Evol. 5 (2021) 422–448.
[24] F.E. Elling, et al., A novel quinone biosynthetic pathway illuminates the evolution of aerobic metabolism, Proc. Natl. Acad. Sci. USA 122 (2025) e2421994122.
[25] B. Schoepp-Cothenet, et al., Menaquinone as pool quinone in a purple bacterium, Proc. Natl. Acad. Sci. USA 106 (2009) 8549–8554.
[26] F.S.P. Nagies, et al., A spectrum of verticality across genes, PLoS Genet 16 (2020) e1009200. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009200.
[27] A.M. Satkoski, et al., A redox-stratified ocean 3.2 billion years ago, Earth Planet. Sci. Lett. 430 (2015) 43–53.
[28] A.M. Satkoski, et al., Corrigendum to "A redox-stratified ocean 3.2 billion years ago" [Earth Planet. Sci. Lett. 430 (2015) 43–53], Earth Planet. Sci. Lett. 460 (2017) 317–319.
[29] S.A. Crowe, et al., Atmospheric oxygenation three billion years ago, Nature 501 (2013) 535–538.
[30] G. Albut, et al., Modern rather than Mesoarchaean oxidative weathering responsible for the heavy stable Cr isotopic signatures of the 2.95 Ga old Ijzermijn iron formation (South Africa), Geochim. Cosmochim. Acta 228 (2018) 157–189.
[31] J.J. Brocks, et al., Archean molecular fossils and the early rise of eukaryotes, Science 285 (1999) 1033–1036.
[32] B. Rasmussen, et al., Reassessing the first appearance of eukaryotes and cyanobacteria, Nature 455 (2008) 1101–1104.
[33] G.P. Fournier, et al., The Archean origin of oxygenic photosynthesis and extant cyanobacterial lineages, Proc. R. Soc. B 288 (2021) 20210675.
[34] J.W. Schopf, B.M. Packer, Early Archean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona group, Australia, Science 237 (1987) 70–73.
[35] M.D. Brasier, et al., Questioning the evidence for earth's oldest fossils, Nature 416 (2002) 76–81.
[36] D.C. Catling, K.J. Zahnle, The Archean atmosphere, Sci. Adv. 6 (2020) eaax1420.

[37] M. Daye, et al., Light-driven anaerobic microbial oxidation of manganese, Nature 576 (2019) 311–314.
[38] J.E. Johnson, et al., Manganese-oxidizing photosynthesis before the rise of cyanobacteria, Proc. Natl. Acad. Sci. USA 110 (2013) 11238–11243.
[39] W. Liu, et al., Anoxic photogeochemical oxidation of manganese carbonate yields manganese oxide, Proc. Natl. Acad. Sci. USA 117 (2020) 22698–22704.
[40] R. Murali, R.B. Gennis, J. Hemp, Evolution of the cytochrome *bd* oxygen reductase superfamily and the function of CydAA' in Archaea, ISME J. 15 (2021) 3534–3548.
[41] R. Murali, J. Hemp, R.B. Gennis, Evolution of quinol oxidation within the heme-copper oxidoreductase superfamily, Biochim. Biophys. Acta Bioenerg. 1863 (2022) 148907.
[42] R.M. Soo, et al., On the origins of oxygenic photosynthesis and aerobic respiration in Cyanobacteria, Science 355 (2017) 1436–1440.
[43] A. Bekker, et al., Dating the rise of atmospheric oxygen, Nature 427 (2004) 117–120.
[44] M. Schidlowski, R. Eichmann, C.E. Junge, Carbon isotope geochemistry of the Precambrian Lomagundi carbonate province, Rhodesia, Geochim. Cosmochim. Acta 40 (1976) 449–455.
[45] A. Bekker, H.D. Holland, Oxygen overshoot and recovery during the early Paleoproterozoic, Earth Planet. Sci. Lett. 317-318 (2012) 295–304.
[46] M.S.W. Hodgskiss, P.W. Crockford, A.V. Turchyn, Deconstructing the Lomagundi-Jatuli carbon isotope excursion, Annu. Rev. Earth Planet. Sci. 51 (2023) 301–330.
[47] K.B. Mayika, et al., The Paleoproterozoic Francevillian succession of Gabon and the Lomagundi-Jatuli event, Geology 48 (2020) 1099–1104.
[48] A.R. Prave, et al., The grandest of them all: the Lomagundi-Jatuli event and earth's oxygenation, J. Geol. Soc. Lond. 179 (2022) jgs2021-036.
[49] T.M. Lenton, et al., Earliest land plants created modern levels of atmospheric oxygen, Proc. Natl. Acad. Sci. USA 113 (2016), 9704–0709.
[50] I. Mukherjee, et al., The boring billion, a slingshot for complex life on earth, Sci. Rep. 8 (2018) 4432.
[51] D.A. Stolper, C.B. Keller, A record of deep-ocean dissolved O2 from the oxidation state of iron in submarine basalts, Nature 553 (2018) 323–327.
[52] L.J. Alcott, B.J. Mulls, S.W. Poulton, Stepwise earth oxygenation is an inherent property of global biogeochemical cycling, Science 366 (2019) 1333–1337.
[53] A.D. Anbar, A.H. Knoll, Proterozoic Ocean chemistry and evolution: a bioinorganic bridge? Science 297 (2002) 1137–1142.
[54] G.L. Arnold, et al., Molybdenum isotope evidence for widespread anoxia in mid-Proterozoic oceans, Science 203 (2004) 87–90.
[55] D.E. Canfield, A new model for Proterozoic Ocean chemistry, Nature 396 (1998) 450–453.
[56] J.M. Klatt, et al., Possible link between earth's rotation rate and oxygenation, Nat. Geosci. 7 (2021) 1–7.
[57] S.W. Poulton, P.W. Ralick, D.E. Canfield, The transition to a sulphidic ocean similar to ~ 1.84 billion years ago, Nature 431 (2004) 173–177.
[58] J.F. Allen, A proposal for formation of Archaean stromatolites before the advent of oxygenic photosynthesis, Front. Microbiol. 7 (2016) 1784.
[59] J.F. Allen, B. Thake, W.F. Martin, Nitrogenase inhibition limited oxygenation of earth's Proterozoic atmosphere, Trends Plant Sci. 24 (2019) 1022–1031.
[60] D.B. Mills, et al., Eukaryogenesis and oxygen in earth history, Nat. Ecol. Evol. 6 (2022) 520–532.
[61] N. Mrnjavac, et al., The radical impact of oxygen on prokaryotic evolution – enzyme inhibition first, uninhibited essential biosynthesis second, aerobic respiration third, FEBS Lett. 598 (2024) 1692–1714.
[62] P. John, F.R. Whatley, Paracoccus denitrificans and the evolutionary origin of the mitochondrion, Nature 254 (1975) 495–498.
[63] N. Lane, W. Martin, The energetics of genome complexity, Nature 467 (2010) 929–934.
[64] P.A. del Giorgio, C.M. Duarte, Respiration in the open ocean, Nature 420 (2002) 379–385.
[65] P. Van Cappellen, E.D. Ingall, Redox stabilization of the atmosphere and oceans by phosphorus-limited marine productivity, Science 271 (1996) 493–496.
[66] V.B. Borisov, et al., The cytochrome bd respiratory oxygen reductases, Biochim. Biophys. Acta 1807 (2011) 1398–1413.
[67] M. Kuntz, Plastid terminal oxidase and its biological significance, Planta 218 (2004) 896–899.
[68] R. Pennisi, et al., Molecular evolution of alternative oxidase proteins: a phylogenetic and structure modeling approach, J. Mol. Evol. 82 (2016) 207–218.
[69] M.M. Pereira, M. Santane, M. Teixeira, A novel scenario for the evolution of haem-copper oxygen reductases, Biochim. Biophys. Acta 1505 (2001) 185–208.
[70] F.L. Sousa, et al., The superfamily of heme-copper oxygen reductases: types and evolutionary considerations, Biochim. Biophys. Acta 1817 (2012) 629–637.
[71] R.M. Soo, J. Hemp, P. Hugenholtz, Evolution of photosynthesis and aerobic respiration in the cyanobacteria, Free Radic. Biol. Med. 140 (2019) 200–205.
[72] A.E. McDonald., G.C. Vanlerberghe, Alternative oxidase and plastoquinol terminal oxidase in marine prokaryotes of the Sargasso Sea, Gene 11 (2005) 15–24.
[73] A.E. McDonald, G. Vanlerberghe, Branched mitochondrial electron transport in the Animalia: presence of alternative oxidase in several animal phyla, IUBMB Life 56 (2004) 333–341.
[74] A.E. McDonald, et al., Flexibility in photosynthetic electron transport: the physiological role of plastoquinol terminal oxidase (PTOX), Biochim. Biophys. Acta 1807 (2011) 054–967.
[75] J.F. Allen, Photosynthesis of ATP-electrons, proton pumps, rotors, and poise, Cell 110 (2002) 273–276.
[76] Z. Jiang, et al., Mitochondrial AOX supports redox balance of photosynthetic electron transport, primary metabolite balance, and growth in *Arabidopsis thaliana* under high light, Int. J. Mol. Sci. 20 (2019) 3067.
[77] D. Wang, A. Fu, The plastid terminal oxidase is a key factor balancing the redox state of thylakoid membrane, Enzymes 40 (2016) 143–171.
[78] R. Murali, et al., Diversity and evolution of nitric oxide reduction in bacteria and archaea, Proc. Natl. Acad. Sci. USA 121 (2024) e2316422121.
[79] M. Degli Esposti, et al., Oxygen reductases in alphaproteobacterial genomes: physiological evolution form low to high oxygen environments, Front. Microbiol. 20 (2019) 499.
[80] V. Sharma, M. Wikström, A structural and functional perspective on the evolution of the heme-copper oxidases, FEBS Lett. 588 (2014) 3787–3792.
[81] T.A. Mahendrarajah, et al., ATP synthase evolution on a cross-braced dated tree of life, Nat. Commun. 14 (2023) 7456.
[82] N.A. O'Leary, et al., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, Nucleic Acids Res. 44 (2016) D733–D745.
[83] F.P. Rosenbaum, V. Müller, Energy conservation under extreme energy limitation: the role of cytochromes and quinones in acetogenic bacteria, Extremophiles 25 (2021) 413–424.
[84] R.J. Weaver, A.E. McDonald, Mitochondrial alternative oxidase across the tree of life: presence, absence, and putative cases of lateral gene transfer, Biochim. Biophys. Acta Bioenerg. 1864 (2023) 149003.
[85] M. Blum, et al., InterPro: the protein sequence classification resource in 2025, Nucleic Acids Res. 53 (2025) D444–D456.
[86] B. Buchfink, C. Xie, D.H. Huson, Fast and sensitive protein alignment using DIAMOND, Nat. Methods 12 (2015) 59–60.
[87] I. Letunic, P. Bork, Interactive tree of life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool, Nucleic Acids Res. 52 (2024) W78–W83.
[88] J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: reconstruction, analysis, and visualization of Phylogenomic data, Mol. Biol. Evol. 33 (2016) 1635–1638.
[89] K. Katoh, et al., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (2002) 3059–3066.
[90] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313.
[91] C. Johnson, et al., Pathways of iron and sulfur acquisition, cofactor assembly, destination, and storage in diverse archaeal methanogens and alkanotrophs, J. Bacteriol. 203 (2021) e0011721.
[92] Y. Ou, et al., Expanding the phylogenetic distribution of cytochrome b-containing methanogenic archaea sheds light on the evolution of methanogenesis, ISME J. 16 (2022) 2373–2387.
[93] R.K. Thauer, et al., Methanogenic archaea: ecologically relevant differences in energy conservation, Nat. Rev. Microbiol. 6 (2008) 579–591.
[94] J. Hemp, R.B. Gennis, Diversity of the heme-copper superfamily in archaea: insights from genomics and structural modeling, Results Probl. Cell Differ. 45 (2008) 1–31.
[95] D. Wu, et al., The IMMUTANS variegation locus of Arabidopsis defines a mitochondrial alternative oxidase homolog that functions during early chloroplast biogenesis, Plant Cell 11 (1999) 43–55.
[96] M.S. Albury, C. Elliott, A.L. Moore, Towards a structural elucidation of the alternative oxidase in plants, Physiol. Plant. 137 (2009) 316–327.
[97] S. Nelson-Sathi, et al., Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea, Proc. Natl. Acad. Sci. USA 109 (2012) 20537–20542.
[98] M. Degli Esposti, On the evolution of cytochrome oxidases consuming oxygen, Biochim. Biophys. Acta Bioenerg. 1861 (2020) 148304.
[99] M.D. Brasier, J.F. Lindsay, A billion years of environmental stability and the emergence of eukaryotes: new data from northern Australia, Geology 26 (1998) 555–558.
[100] R. Buik, D.J. Des Marais, A.H. Knoll, Stable isotopic compositions of carbonates from the Mesoproterozoic Bangemall group, northwestern Australia, Chem. Geol. 123 (1995) 153–171.
[101] J. Farquhar, H. Bao, M. Thiemens, Atmospheric influence of earth's earliest sulfur cycle, Science 289 (2000) 756–759.
[102] G. Luo, et al., Rapid oxygenation of earth's atmosphere 2.33 billion years ago, Sci. Adv. 2 (2016) 5.
[103] S.W. Poulton, et al., A 200-million-year delay in permanent atmospheric oxygenation, Nature 592 (2021) 232–236.
[104] H. Hofmann, Precambrian microflora, Belcher Islands, Canada; significance and systematics, J. Paleontol. 50 (1976) 1040–1073.
[105] B. Eickmann, et al., Isotopic evidence for oxygenated Mesoarchaean shallow oceans, Nat. Geosci. 11 (2018) 133–138.
[106] L.A. Patry, et al., Dating the evolution of oxygenic photosynthesis using La-Ce geochronology, Nature 642 (2025) 99–104.
[107] N.J. Planavsky, et al., Evidence for oxygenic photosynthesis half a billion years before the great oxidation event, Nat. Geosci. 7 (2014) 283–286.
[108] D.T. Wilmeth, et al., Evidence for benthic oxygen production in Neoarchean lacustrine stromatolites, Geology 50 (2022) 907–911.
[109] D.T. Wilmeth, et al., Neoarchean (2.7 Ga) lacustrine stromatolite deposits in the Hartbeesfontein Basin, Ventersdorp Supergroup, South Africa: Implications for oxygen oases, Precambrian Res. 320 (2019) 291–302.
[110] J.F. Kasting, What caused the rise of atmospheric O_2? Chem. Geol. 362 (2013) 13–25.
[111] C.J. Bjerrum, D.E. Canfield, Ocean productivity before about 1.9 Gyr ago limited by phosphorus adsorption onto iron oxides, Nature 417 (2002) 159–162.

[112] E.D. Swanner, et al., Modulation of oxygen production in Archaean oceans by episodes of Fe(II) toxicity, Nat. Geosci. 8 (2015) 126–130.
[113] J.F. Kasting, D.E. Canfield, The Global Oxygen Cycle, in: Fundamentals of Geobiology, John Wiley & Sons, Ltd, Chichester, UK, 2012, pp. 93–104.
[114] C.L. Grettenberger, D.Y. Sumner, Physiology, not nutrient availability, may have limited primary productivity after the emergence of oxygenic photosynthesis, Geobiology 22 (2024) e12622.
[115] B. Rasmussen, et al., Evidence for anoxic shallow oceans at 2.45 Ga: Implications for the rise of oxygenic photosynthesis, Geology 47 (2019) 622–626.
[116] L.M. Ward, J.L. Kirschvink, W.W. Fischer, Timescales of oxygenation following the evolution of oxygenic photosynthesis, Orig. Life Evol. Biosph. 46 (2016) 51–65.
[117] J. Castresana, et al., Evolution of cytochrome oxidase, an enzyme older than atmospheric oxygen, EMBO J. 13 (1994) 2516–2525.
[118] S. Kihara, D.A. Hartzler, S. Savikhin, Oxygen concentration inside a functioning photosynthetic cell, Biophys. J. 106 (2014) 1882–1889.
[119] W.F. Martin, F.L. Sousa, Early microbial evolution: the age of anaerobes, Cold Spring Harb. Perspect. Biol. 8 (2015) a018127.
[120] J.A. Imlay, Pathways of oxidative damage, Ann. Rev. Microbiol. 57 (2003) 395–418.
[121] M. Khademian, J.A. Imlay, How microbes evolved to tolerate oxygen, Trends Microbiol. 29 (2021) 428–440.
[122] N. Pan, J.A. Imlay, How does oxygen inhibit central metabolism in the obligate anaerobe Bacteroides thetaiotaomicron, Mol. Microbiol. 39 (2001) 1562–1571.
[123] S.W. Ragsdale, Pyruvate ferredoxin oxidoreductase and its radical intermediate, Chem. Rev. 103 (2003) 2333–2346.
[124] C. Wirth, et al., Structure and function of mitochondrial complex I, Biochim. Biophys. Acta Bioenerg. 1857 (2016) 902–914.
[125] A.D. Baughn, M.H. Malamy, The strict anaerobe Bacteroides fragilis grows in and benefits from nanomolar concentrations of oxygen, Nature 427 (2004) 221–222.
[126] E. Forte, et al., The terminal oxidase cytochrome bd promotes sulfide-resistant bacterial respiration and growth, Sci. Rep. 6 (2016) 23788.
[127] L.E. Khmelevtsova, et al., Prokaryotic peroxidases and their application in biotechnology (review), Appl. Biochem. Microbiol. 56 (2020) 373–380.
[128] H.S. Tehrani, A.A. Moosavi-Movahedi, Catalase and its mysteries, Prog. Biophys. Mol. Biol. 140 (2018) 5–12.
[129] J.M. Hayes, Factors controlling 13C contents of sedimentary organic compounds: principles and evidence, Mar. Geol. 113 (1993) 111–125.
[130] D.Y. Sumner, Oxygenation of earth's atmosphere induced metabolic and ecologic transformations recorded in the Lomagundi-Jatuli carbon isotopic excursion, Appl. Environ. Microbiol. 90 (6) (2024).
[131] M. Kędzior, et al., Resurrected rubisco suggests uniform carbon isotope signatures over geologic time, Cell Rep. 39 (2022) 110726.
[132] R.Z. Wang, et al., Carbon isotope fractionation by an ancestral rubisco suggest that biological proxies for CO_2 through geologic time should be reevaluated, Proc. Natl. Acad. Sci. USA 120 (2023) e2300466120.
[133] A.R. Prave, Environmental microbiology explains the largest positive carbon isotope excursion in earth history, the Lomagunid-Jatuli event, Appl. Environ. Microbiol. 90 (2024) e00936–24.
[134] S. Maruyama, et al., Initiation of leaking earth: an ultimate trigger of the Cambrian explosion, Gondwana Res. 25 (2014) 910–944.

6 Zusammenfassung der Ergebnisse

Prokaryotische Genome werden stark durch LGT und Genverlust beeinflusst, wodurch ein Prozess des Genflusses entsteht (Mira *et al.* 2001, Arnold *et al.* 2022). Dieser Genfluss generiert die Pangenom Struktur, welche ein konserviertes Kerngenom beinhaltet, das von einem variablen akzessorischen Genom umrandet wird (Tettelin *et al.* 2005, Medini *et al.* 2005, Tettelin *et al.* 2008, Vernikos *et al.* 2015, Brockhurst *et al.* 2019). Genflussraten wurden bis heute hauptsächlich auf der Spezies- und Genusebene untersucht, wobei eine starke Beziehung zwischen phylogenetischer Distanz und Menge an Genunterschieden entdeckt wurde (Hao & Golding 2006, Marri *et al.* 2006, Marri *et al.* 2007, Nowell *et al.* 2014, Wolf *et al.* 2016, Touchon *et al.* 2009, Wielgoss *et al.* 2016, Andreani *et al.* 2017, Rocha 2018, Haudiquet *et al.* 2022). Jedoch sind diese Studien nicht direkt vergleichbar, da entweder die Berechnungen der phylogenetischen Distanz auf unterschiedlichen Methoden basieren oder auf Kerngenen beruhen, welche stark von der Zusammensetzung des Pangenoms abhängig sind (Vernikos *et al.* 2015).

In **Publikation I** wurden prokaryotische Genflussraten anhand von 5.655 Genomen und 2.872 MAGs berechnet und über verschiedene Taxa hinweg miteinander verglichen. Die Genflussraten wurden basierend auf der Beziehung zwischen Sequenzdivergenz universeller und konservierter Kerngene sowie Unterschieden im Geninhalt prokaryotischer Genompaare berechnet. Wie in vorherigen Studien zeigt sich auch in dieser Publikation eine stark positive Beziehung zwischen beiden Messwerten (Touchon *et al.* 2009, Wolf *et al.* 2016, Wielgoss *et al.* 2016, Andreani *et al.* 2017, Rocha 2018, Haudiquet *et al.* 2022, Trost *et al.* 2024). Die Genflussraten höherer prokaryotischer Taxa wiesen nahezu identische Werte auf, mit einer durchschnittlichen bakteriellen Rate von 2,9 % und eine archaellen Rate von 2,57 % Geninhalt-Unterschieden pro 1 % Aminosäuresequenzdivergenz der Kerngene. Dieses Muster wird auch in den Analysen metagenomisch-assemblierter Genome bestätigt, sofern die Qualität der einzelnen MAGs hoch genug ist (> 80 %).

Auf der Genus- bis zur Phylum-Ebene bleiben diese Genflussraten fast konstant und sinken nur leicht, mit steigendem taxonomischem Rang. Wohingegen die Werte auf der Spezies-Ebene stark variieren. Die ähnlichen Genflussraten in höheren prokaryotischen Taxa sowie über verschiedene taxonomische Ebenen hinweg deuten grundsätzlich auf ein uhrähnliches Verhalten der Veränderung des Geninhalts relativ zu Aminosäuresequenzdivergenz in den universellsten und am vertikalsten vererbten Genen in

Prokaryoten. Die Genflussraten moderner prokaryotischer Genome könnten somit genauso alt sein wie Prokaryoten selber. Des Weiteren unterstreicht dies, dass LGT eine natürliche Komponente der Evolution von Prokaryoten ist (Trost *et al.* 2024).

Anhand der Genflussraten höherer Taxa können auch Schätzungen für den durchschnittlichen Anteil akzessorischer Gene in aktuellen Genomen gemacht werden. Dies kann als Hinweis gesehen werden, dass es Pangenomstrukturen mit stabilem Kerngenom und variablen akzessorischem Genom in Prokaryoten schon seit der Entstehung der bakteriellen und archaellen Linie gibt (Trost *et al.* 2024).

Der zweite Teil der Arbeit (**Publikation II und III**) geht näher auf die Evolution von atmosphärischem Sauerstoff ein. Klare Erkenntnisse zeigen, dass der Sauerstoffgehalt vor ca. 2,4 Milliarden Jahren stark anstieg, das GOE (Holland 2002, Gumsley *et al.* 2017). Einige Studien untersuchen die Möglichkeit, dass Sauerstoff bereits kurz vor dem GOE in geringen Konzentrationen in der Atmosphäre auftrat (Anbar *et al.* 2007, Kaufman *et al.* 2007, Czaja *et al.* 2012, Crowe *et al.* 2013, Meixnerová *et al.* 2021, He *et al.* 2021, Stone *et al.* 2022, He *et al.* 2023, Sweetman *et al.* 2024). Diese Studien, die oft auf molekulare Phylogenien von O_2-metabolisierenden Enzymen basieren (Brochier-Armanet *et al.* 2009, Boden *et al.* 2021, Jablońska & Tawfik 2021, Bafana *et al.* 20200, He *et al.* 2023, Davin *et al.* 2025, Elling *et al.* 2025), nehmen durch ihre Methodik automatisch an, dass die Enzyme nicht durch LGT beeinflusst wurden. Heute ist jedoch klar, dass fast alle Gene prokaryotischer Genome lateral transferiert werden (Nagies *et al.* 2020, Dagan & Martin 2007, Dagan *et al.* 2008, Trost *et al.* 2024). Eine Analyse aus **Publikation II** untersucht ebenfalls, inwiefern sich der Einfluss von LGT auf sauerstoffabhängige und sauerstoffunabhängige Gene unterscheidet. Dazu wurden Vertikalitätsmesswerte einzelner Gene genutzt, welche beschreiben, wie häufig ein Gen zwischen Abstammungslinien transferiert wurde (Nagies *et al.* 2020). Es zeigt sich deutlich, dass sauerstoffabhängige Gene mehr durch LGT beeinflusst werden, als sauerstoffunabhängige Gene. Dieses Ergebnis bleibt über verschiedene funktionelle Kategorien hinweg konsistent. Weitere Ergebnisse aus **Publikation II** deuten darauf hin, dass O_2-abhängige Enzyme einen physiologischen Vorteil mit Zunahme der Sauerstoffkonzentration vor ca. 2,4 Milliarden Jahren hatten. Im Gegensatz zu traditionellen Annahmen, nach denen O_2-abhängige Enzyme hauptsächlich mit aerober Atmung und damit einhergehender Energiegewinnung verknüpft werden (Rytkönen 2018), deuten Analysen aus **Publikation II** darauf hin, dass prokaryotische Zellen zunächst in der Lage sein mussten in sauerstoffreichen Umgebungen zu überleben bevor sie O_2 zur Steigerung der Energieeffizienz nutzen konnten. Somit bestand die entscheidende physiologische Anpassung, die durch O_2-abhängige Enzyme bewirkt wurde, zunächst in der

Resistenz gegenüber Sauerstofftoxizität. Diese Resistenz entstand, indem zusätzliche O_2-abhängige Enzyme in prokaryotische Genome integriert oder O_2-sensitive Enzyme durch funktionelle, O_2-abhängige Analoge ausgetauscht wurden (Mrnjavac *et al.* 2024).

Anders als molekulare Methoden setzen geochemische Befunde den Ursprung von relevanten Mengen an Sauerstoff in der Atmosphäre mit dem GOE gleich (Holland 2002, Gumsley *et al.* 2017), was bedeuten würde, dass auch Sauerstoff-metabolisierende Enzyme wie terminale Oxidasen ihren Ursprung mit oder kurz nach dem GOE haben. In **Publikation III** wurde der Ursprung von terminalen Oxidasen sowie ihre Verteilung in prokaryotischen Abstammungslinien untersucht, indem ihr Auftreten in verschiedenen Taxa auf einen zeit-kalibrierten phylogenetischen Baum (Mahendrarajah *et al.* 2023) kartiert wurde. Dieser Ansatz unterscheidet sich von vorherigen phylogenetischen Studien, da (i) das GOE als frühste Möglichkeit einer Entstehung der terminalen Oxidasen akzeptiert wurde, (ii) die Existenz von LGT bei prokaryotischen Genen, besonders sauerstoffabhängigen Genomen akzeptiert wurde und (iii) ein unabhängiger phylogenetischer Baum zur Datierung genutzt wurde, welcher nicht auf Sauerstoff-metabolisierenden Enzymen basiert, sondern auf Basis der hauptsächlich vertikal vererbten ATP-Synthase generiert wurde.

Die Daten aus **Publikation III** weisen darauf hin, dass Cytochrom-*bd* Oxidasen, Häm-Kupfer-Oxidasen und alternative Oxidasen (AOX, PTOX) im Zuge des GOE vor etwa 2,4 Milliarden Jahren entstanden sind, woraufhin diese Gene stark von LGT beeinflusst wurden. Außerdem unterstreichen die Ergebnisse die Physiologie rund um das GOE und eröffnen ein biologisches Modell, dass die ^{13}C-Isotpenanomalie der Lomagundi-Jatuli Exkursion (LJE) als Produkt eines einzigen cyanobakteriellen Enzyms direkt erklären. Diese besagt, dass zunächst vor dem GOE den Vorläufern der Cyanobakterien ein Sauerstoffentwicklungs-Komplex fehlte, sodass Cyanobakterien keinen Beitrag zum atmosphärischen O_2 leisten konnten. Mit Beginn des GOE wurden Cyanobakterien sauerstoffproduzierend. Ihr ungebremstes Wachstum führte zu einer exponentiellen Anreicherung von O_2 in der Atmosphäre, die jedoch bei 2 % [v/v] atmosphärischem O_2 begrenzt war, da die Stickstofffixierung der Cyanobakterien, speziell die Nitrogenasen, durch O_2 oberhalb dieser Schwelle gehemmt werden. Da das atmosphärische CO_2 zum Zeitpunkt des GOE auf etwa 0,02 atm geschätzt wird, verbrauchte die Produktion von 0,02 atm O_2 durch Kohlenstoffbindung mittels des Enzyms RuBisCo fast den gesamten atmosphärischen CO_2-Vorrat. Dies führte zu einer extremen Anreicherung von ^{13}C in der Atmosphäre, der LJE, verursacht durch die Isotopendiskriminierung von RuBisCo. Die LJE fand jedoch bei einem O_2-Gehalt von 2 % statt, da die durch Nitrogenase auferlegte strenge Obergrenze für die O_2-Anreicherung bis zum Entstehen der Landpflanzen in Kraft blieb. Der

hohe ^{12}C-Gehalt in der Atmosphäre am Ende der Lomagundi-Jatuli-Exkursion markiert den Ursprung der Sauerstoffreduktasen, ihre rasche Verbreitung durch ihre Funktion der Freisetzung von CO_2 bei der Atmung und den Beginn des Gleichgewichts zwischen photosynthetischer O_2-Produktion und respiratorischem O_2-Verbrauch, zunächst bei einem atmosphärischen O_2-Gehalt von 2 % (Trost *et al.* 2026).

7 Literaturverzeichnis

Abe, K., Nomura, N. & Suzuki, S. (2020). Biofilms: hot spots of horizontal gene transfer (HGT) in aquatic environments, with a focus on a new HGT mechanism. *FEMS Microbiol Ecol.* **96:**fiaa031.

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

Alcott, L.J., Mulls, B.J. & Poulton, S.W. (2019). Stepwise earth oxygenation is an inherent property of global biogeochemical cycling. *Science.* **366:**1333–1337.

Allen, J.F., Thake, B. & Martin, W.F. (2019). Nitrogenase inhibition limited oxygenation of Earth's Proterozoic atmosphere. *Trends Plant Sci.* **24:**1022–1031.

Almeida, A., Nayfach, S., Boland, M. *et al.* (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol.* **39:**105–114.

Anbar, A.D. & Knoll, A.H. (2022) Proterozoic Ocean chemistry and evolution: a bioinorganic bridge? *Science.* **297:**1137–1142.

Andreani, N.A., Hesse, E. & Vos, M. (2017) Prokaryote genome fluidity is dependent on effective population size. *ISME J.* **11:**1719–1721.

Anbar, A.D., Duan, Y., Lyons, T.W. *et al.* (2007) A whiff of oxygen before the great oxidation event? *Science.* **317:**1903–1906.

Arnold, B.J., Huang, I. & Hanage, W.P. (2022) Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol.* **20:**206–218.

Atteia, A., van Lis, R., van Hellemond, J.J. *et al.* (2004) Identification of prokaryotic homologues indicates an endosymbiotic origin for the alternative oxidase of mitochondria (AOX) and chloroplasts (PTOX). *Gene.* **330:**143–148.

Barrick, J.E., Yu, D.S., Yoon, S.H. *et al.* (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature.* **461:**1243–1247.

Baumdicker, F. & Kupczok, A., (2023) Tackling the pangenome dilemma requires the concerted analysis of multiple population genetic processes. *Genome Biol Evol.* **15:**evad067.

Bekker, A., Holland, H.D., Wang, P-L. *et al.* (2004) Dating the rise of atmospheric oxygen. *Nature.* **427:**117–120.

Bekker, A. & Holland, H.D. (2012) Oxygen overshoot and recovery during the early Paleoproterozoic. *Earth and Planet Sci Lett.* **317-318:**295–304.

Berthold, D.A. & Stenmark, P. (2003) Membrane-bound diiron carboxylate proteins. *Annu Rev Plant Biol.* **54:**497–517.

Borisov, V.B., Gennis, R.B., Hemp, J. *et al.* (2011) The cytochrome *bd* respiratory oxygen reductases. *Biochim Biophys Acta.* **1807:**1398–1413.

Borisov, V.B., Gennis, R.B., Hemp, J. *et al.* (2015) The cytochrome *bd* respiratory oxygen reductases. *Biochim Biophys Acta.* **1807:**1398–1413.

Boughner, L.A. & Singh, P. Microbial Ecology: Where are we now? *Postdoc J.* **4:**3–17.

Brockhurst, M.A., Harrison, E., Hall, J.P.J. *et al.* (2019) The Ecology and Evolution of Pangenomes. *Curr Biol.* 29:1094–1103.

Canfield, D.E. (1998) A new model for Proterozoic Ocean chemistry. *Nature.* **396:**450–453.

Chain, P.S.G., Grafham, D.V., Fulton, R.S. *et al.* (2009) Genomics. Genome project standards in a new era of sequencing. *Scienc.e* **326:**236–237.

Chen, I., Christie, P.J. & Dubnau, D. (2005). The Ins and Outs of DNA Transfer in Bacteria. *Science.* **310:**1456–1460.

Chklovski, A., Parks, D.H., Woodcroft, B.J. *et al.* (2023) CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods.* **20:**1203–1212.

Crowe, S.A., Døssing, L.N., Beukes, N.J. *et al.* (2013). Atmospheric oxygenation three billion years ago. *Nature.* **40:**535–538.

Czaja, A.D., Johnson, C.M., Roden, E.E. *et al.* (2012) Evidence for free oxygen in the Neoarchean ocean based on coupled iron-molybdenum isotope fractionation. *Geochim Cosmochim Acta.* **86:**118–137.

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

Dagan, T., Artzy-Randrup, Y., & Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci U S A.* **105:**1099–1108.

Darwin, C. (1860) Über die Entstehung der Arten im Thier- und Pflanzen-Reich durch natürliche Züchtung. Stuttgart: Schweizerbart.

Davín, A.A., Woodcroft, B.J., Soo, R.M. *et al.* (2025) A geological timescale for bacterial evolution and oxygen adaptation. *Science.* **388:**6742.

Degli Esposti, M., Rosas-Pérez, T., Servín-Garcidueñas, L.E. *et al.* (2015) Molecular evolution of cytochrome *bd* oxidases across proteobacterial genomes. *Genome Biol Evol.* **7:**801–820.

Demoulin, C.F., Lara, Y.J., Lambion, A. *et al.* (2024) Oldest thylakoids in fossil cells directly evidence oxygenic photosynthesis. *Nature.* **625:**529–534.

Doolittle, W.F. (1999). Phylogenetic classification and the universal tree. *Science.* **284:**2124–2129.

Dubey, G.P. & Ben-Yehuda, S. (2011). Intercellular nanotubes mediate bacterial communication. *Cell.* **144:**590–600.

Dubnau, D. (1999) DNA Uptake in Bacteria. *Annu Rev Microbiol.* **53:**217–244.

Fischer, W.W., Hemp, J. & Johnson, J.E. (2016). Evolution of oxygenic photosynthesis. *Annu Rev Earth Planet Sci.* **44:**647–683.

Frauenstein, F., Veizer, J., Beukes, N. *et al.* (2009) Transvaal Supergroup carbonates: Implications for Paleoproterozoic $\delta^{18}O$ and $\delta^{13}C$ records. *Precambrian Res.* **174:**149–160.

Garza, D.R. & Dutilh, B.E. (2015) From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. *Cell Mol Life Sci.* **72:**4287–4308.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A. *et al.* (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J of Syst Evol Microbiol.* **57:**81–91.

Gumsley, A.P., Chamerlain, K.R., Bleeker, W. *et al.* (2017) Timing and tempo of the Great Oxidation Event. *Procs Nat Acad Sci U S A.* **114:**1811–1816.

Han, K., Li, Z., Peng, R. *et al.* (2013) Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. *Sci Rep.* **3:**2101.

Hao, W. & Golding, G.B. (2006) The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res.* **16:**1655–1663.

Haudiquet, M., De Sousa, J.M., Touchon, T. *et al.* (2022) Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Philos Trans R Soc Lond B Biol Sci.* **377:**20210234.

Hayashi, T., Makino, K., Ohnishi, M. *et al.* (2001) Complete genome sequence of enterohemorrhagic, Escherichia coli O157:H7 and genomic comparison with laboratory strain K-12. *DNA Res.* **8:**11–22.

Hayes, J.M. (1993) Factors controlling ^{13}C contents of sedimentary organic compounds: Principles and evidence. *Mar Geol.* **113:**111–125.

He, H., Wu, X., Xian, H. *et al.* (2021). An abiotic source of Archean hydrogen peroxide and oxygen that pre-dates oxygenic photosynthesis. *Nature.* **12:**6611.

He, H., Wu, Y., Zhu, J. *et al.* (2023). A mineral-based origin of Earth's initial hydrogen peroxide and molecular oxygen. *Proc Nat Acad Sci U S A.* **120:**e2221984120.

Hodgskiss, M.S.W., Crockford, P..W., Peng, Y. *et al.* (2019) A productivity collapse to end Earth's Great Oxidation. *Proc Nat Acad Sci U S A.* **116:**17207–17212.

Hogg, J.S., Hu, F.Z., Janto, B. *et al.* (2007). Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol.* **8:**R103.

Holland, H.D. (2002) Volcanic gases, black smokers, and the great oxidation event. *Geochim Cosmochim Acta*. **66:**3811–3826.

Holland, H.D. (2006) The oxygenation of the atmosphere and oceans. *Philos Trans R Soc Lond B Biol Sci.* **361:**903–915.

Hu, Z., Cheng, L. & Wang, H. (2015) The Illumina-solexa sequencing protocol for bacterial genomes. *Methods Mol Biol.* **1231:**91–97.

Huson, D.H. & Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol***. 23:**254–267.

Jablońska, J. & Tawfik, D.S. (2021) The evolution of oxygen-utilizing enzymes suggests early biosphere oxygenation. *Nat Ecol Evol.* **5:**442–448.

Jiang, S.C. & Paul, J.H. (1998) Gene transfer by transduction in the marine environment. *Appl Environ Microbiol.* **64:**2780–2787.

Karhu, J.A. & Holland, H.D. (1996) Carbon isotopes and the rise of atmospheric oxygen. *Geology.* **24:**867-870.

Kato, K., Miyazaki, N., Hamaguchi, T. *et al.* (2021). High-resolution cryo-EM structure of photosystem II reveals damage from high-dose electron beams. *Commun Biol.* **4**:382.

Kaufman, A.J., Johnston, D.T., Farquhar, J. *et al.* (2007). Late Archean Biospheric Oxygenation and Atmospheric Evolution. *Science.* **317:**1900–1903.

Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature.* **217:**624–626.

Kitahara, K. & Miyazaki, K. (2013) Revisiting bacterial phylogeny: Natural and experimental evidence for horizontal gene transfer of 16S rRNA. *Mob Genet Elements* **2:**e24210.

Klatt, J.M., Chennu, A., Arbic, B.K. *et al.* (2021). Possible link between Earth's rotation rate and oxygenation. *Nat Geosci.* **7:**1–7.

Konstantinidis, K.T. & Tiedje, J.M. (2004) Genomic insights that advance the species definition for prokaryotes. *Proc Nat Acad Scie U S A.* **102:**2567–2572.

Koonin, E.V., Makarova, K.S. & Aravind, L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol.* **55:**709–742.

Koppenol, W.H. & Sies, H. (2024). Was hydrogen peroxide present before the arrival of oxygenic photosynthesis? The important role of iron(II) in the Archean Ocean. *Redox Biol.* **69:**103012.

Kunin, V. & Ouzounis, C.A. (2003) The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* **13**:1589–1594.

Kunin, V., Goldovsky, L., Darzentas, N. *et al.* (2005). The net of life: reconstructing the microbial phylogenetic network. *Genome Res.* **15:**954–959.

Kuntz, M. (2004) Plastid terminal oxidase and its biological significance. *Planta* **218:**896–899.

Lang, A.S. & Beatty, J.T. (2006). Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol.* **15:**54–62.

Lawrence, J.G. & Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A.* **95:**9413–9417.

Lenton, T.M., Dahl, T.W., Daines, S.J. *et al.* (2016). Earliest land plants created modern levels of atmospheric oxygen. *Proc Natl Acad Sci U S A.* **113:**9704–0709.

Li, C., Huang, J., Ding, L. *et al.* (2021) Estimation of oceanic and land carbon sinks based on the most recent oxygen budget. *Earth's Future* **9:**e2021EF002124.

Lok, C. (2015) Mining the microbial dark matter. *Nature.* **522:**9413–9417.

Lyons, T.W., Reinhard, C.T. & Planavsky, N.J. (2014) The rise of oxygen in Earth's early ocean and atmosphere. *Nature.* **506:**307–315

Mahendrarajah, T.A., Moody, E.R.R., Schrempf, D. *et al.* (2023) ATP synthase evolution on a cross-braced dated tree of life. *Nat Commun.* **14:**7456.

Maiden, M.C.J., Bygraves, J.A., Feil, E. *et al.* (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A.* **95:**3140–3145.

Manni, M., Berkeley, M.R., Seppey, M. *et al.* (2021) BUSCO: Assessing Genomic Data Quality and Beyond. *Curr Protoc.* **1:**e323.

Marcy, Y., Ouverney, C., Bik, E.M. *et al.* (2007) Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A.* **104:**11889–11894.

Mardis, E., McPherson, J., Martienssen, R. *et al.* (2002) What is finished, and why does it matter. *Genome Res.* **12:**669–671.

Margulies, M., Egholm, M., Altman, W.E. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* **437:**376–380.

Marreiros, B.C., Calisto, F., Castro, P.J. *et al.* (2016) Exploring membrane respiratory chains. *Biochim Biophys Acta.* **1857:**1039–1067.

Marri, P.R., Hao, W. & Golding, G.B. (2006) Gene gain and gene loss in *Streptococcus*: is it driven by habitat? *Mol Biol Evol.* **23:**2379-2391.

Marri, P.R., Hao, W. & Golding, G.B. (2007) The role of laterally transferred genes in adaptive evolution. *BMC Evol Biol.* **7:**1–14.

Martin, A.P., Condon, D.J., Prave, A.R. *et al.* (2013) A review of temporal constraints for the Palaeoproterozoic large, positive carbonate carbon isotope excursion (the Lomagundi-Jatuli Event) *Earth Sci Rev.* **127:**242–261.

Martin, W. (1999) Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays* .**21:**99–104.

Martin, W.F., Tielens, A.G.M. & Mentel, M. (2020) Mitochondria and Anaerobic Energy Metabolism in Eukaryotes: Biochemistry and Evolution. De Gruyter, Berlin.

Matthews, C.A., Watson-Haigh, N.S., Burton, R.A. *et al.* (2024) A gentle introduction to pangenomics. *Brief Bioinform.* **25:**bbae588.

Maxwell, D.P, Wang, Y. & McIntosh, L. (1999) The alternative oxidase lowers mitochondrial reactive oxygen production in plant cells. *Proc Nat Acad Sci U S A.* **96:**8271–8276.

McCutcheon, J.P. & Moran, N.A. (2012) Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol.* **10:**13–26.

McDonald. A.E. & Vanlerberghe, G.C. (2005) Alternative oxidase and plastoquinol terminal oxidase in marine prokaryotes of the Sargasso Sea. *Gene.* **11:**15–24.

McInerney, J.O., McNally, A. & O'Connel, M.J. (2017) Why prokaryotes have pangenomes. *Nat Microbiol.* **2:**17040.

McKernan, K.J., Peckham, H.E., Costa, G.L. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **18:**1527–1541.

Medini, D., Donati, C., Tettelin, H. *et al.* (2005). The microbial pan-genome. *Curr Opin Genet Dev.* **15:**589–594.

Meixnerová, J., Blum, J.D., Johnson, M.W. *et al.* (2021). Mercury abundance and isotopic composition indicate subaerial volcanisms prior to the end-Archean "whiff" of oxygen. *Proc Nat Acad Sci U S A.* **118:**e2107511118.

Melezhik, V.A., Fallick, A.E., Hanski, E.J. *et al.* (2005) Emergence of the aerobic biosphere during the Archean-proterozoic transition: Challenges of future research. *GSA Today.* **15:**4–11.

Mills, D.B., Boyle, R.A., Daines, S.J. *et al.* (2022). Eukaryogenesis and oxygen in earth history. *Nat Ecol Evol.* **6:**520–532.

Mira, A., Ochman, H. & Moran, N.A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17:**589–596.

Mrnjavac, N., Nagies, F.S.P., Wimmer, J.L.E. *et al.* (2024). The radical impact of oxygen on prokaryotic evolution – enzyme inhibition first, uninhibited essential biosynthesis second, aerobic respiration third. *FEBS Lett.* **598:**1692–1714.

Mrnjavac, N., Degli Esposti, M., Mizrahi, I. *et al.* (2024). Three enzymes governed the rise of O_2 on Earth. *Biochim Biophys Acta.* **1865:**149496.

Mukherjee, I., Large, R.R., Corkrey, R. *et al.* (2018) The boring billion, a slingshot for complex life on earth. *Sci. Rep.* **8:**4432.

Murali, R., Gennis, R.B. & Hemp, J. (2021) Evolution of the cytochrome bd oxygen reductase superfamily and the function of Cyd AA' in Archaea. *ISME J.* **15:** 3534–3548.

Murali, R., Hemp, J. & Gennis, R.B. Evolution of quinol oxidation within the heme-copper oxidoreductase superfamily. (2022) *Biochim. Biophys. Acta.* **1863**:148907.

Nagies, F.S.P., Brueckner, J., Tria, F.D.K. *et al.* (2020) A spectrum of verticality across genes. *PLOS Genet.* **16:**e1009200.

Nayfach, S., Roux, S., Seshadri, R. *et al.* (2021) A genomic catalog of Earth's microbiomes. *Nat Biotechnol.* **39:**499–509.

Nowell, R.W., Green, S., Laue, B.E. *et al.* (2014) The extent of genome flux and its role in the differentiation of bacterial lineages. *Genome Biol Evol.* **6:**1514–1529.

Ochman, H., Lawrence, J.G. & Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*. **405:**299–304.

Pallen, M.J. & Wren, B.W. (2007) Bacterial pathogenomics. *Nature.* **449:**835–842.

Parks, D.H., Imelfort, M., Skennerton, C.T. *et al.* (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25:**1043–1055.

Park, D.H., Rinke, C., Chuvochina, M. *et al.* (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol.* **2:**1533–1542.

Parks, D.H., Chuvochina, M., Waite, D.W. *et al.* (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol.* **36:**996–1004.

Pasolli, E., Asnicar, F., Manara, S. *et al.* (2019) Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell.* **176:**649–662.

Pedersen, G.B., Blaschek, L, Frandsen, K.E.H. *et al.* (2023) Cellulose synthesis in land plants. *Mol Plant.* **16:**206–231.

Pennisi, R., Salvi, D., Brandi, V. *et al.* (2016) Molecular Evolution of Alternative Oxidase Proteins: A Phylogenetic and Structure Modeling Approach. *J. Mol. Evol.* **82:**207–218.

Pereira, M.M., Santane, M. & Teixeira, M. (2001) A novel scenario for the evolution of haem-copper oxygen reductases. *Biochim. Biophys. Acta.* **1505**:185–208.

Perna, N.T., Plunkett 3rd, G., Burland, V. *et al.* (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature.* **409:**529–533.

Popa, O. & Dagan, T. (2011). Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol.* **14:**615–635.

Poulton, S.W., Ralick, P.W. & Canfield, D.E. (2004). The transition to a sulphidic ocean approximately 1.84 billion years ago. *Nature.* **431:**173–177.

Prave, A.R., Kirsimäe, K., Lepland, A. *et al.* (2022) The grandest of them all: the Lomagundi-Jatuli Event and Earth's oxygenation. *J. Geol. Soc. Lond.* **179:**jgs2021–036.

Puigbò, P., Lobkovsky, A.E., Kristensen, D.M. *et al.* (2014). Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biology.* **12:**66.

Rappé, M.S. & Giovannoni, S.J. (2003) The uncultured microbial majority. *Annu Rev Microbiol.* **57:**369–394.

Rinke, C., Schwientek, P., Sczyrba, A. *et al.* (2013) Insight into the phylogeny and coding potential of microbial dark matter. *Nature.* **499:**431–437.

Rocha, E.P.C. (2018) Neutral theory, microbial practice: challenges in bacterial population genetics. *Mol Biol Evol.* **35:**1338–1347.

Raymond, J. & Segrè (2006) The effect of oxygen on biochemical networks and the evolution of complex life. *Science.* **311:**1764–1767.

Rytkönen, K.T. (2018) Evolution: Oxygen and early animals. *Elife* **7:**e34756.

Schidlowski, M., Eichmann, R. & Junge, C.E. (1976) Carbon isotope geochemistry of the Precambrian Lomagundi carbonate province, Rhodesia. *Geochim. Cosmochim. Acta.* **40:**449–455.

Schidlowski, M. (1988) A 3,800-million-year isotopic record of life from carbon in sedimentary rocks. *Nature.* **333:**313–318.

Segerman, B. (2012) The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. *Front Cell Infect Microbiol.* **2:**116.

Setubal, J.C. (2021) Metagenome-assembled genomes: concepts, analogies, and challenges. *Biophys Revi.* **13:**905–909.

Sharma, V. & Wikström, M. (2014) A structural and functional perspective on the evolution of the heme-copper oxidases. *FEBS Lett.* **588:**3787–3792.

Shen, J.-R. (2015). The structure of photosystem II and the mechanisms of water oxidation in photosynthesis. *Annu Rev of Plant Biol.* **66:**23–48.

Simão, F.A., Waterhouse, R.M., Ioannidis, P. *et al.* (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31:**3210–3212.

Slotznick, S.P., Johnson, J.E., Rasmussen, B. *et al.* (2022) Reexamination of 2.5-Ga "whiff" of oxygen interval points to anoxic ocean before GOE. *Sci Adv.* **8:**eabj7190.

Sonnenberg, C.B., Kahlke, T. & Haugen, P. (2020) Vibrionaceae core, shell and cloud genes are non-randomly distributed on Chr 1: An hypothesis that links the genomic location of genes with their intracellular placement. *BMC Genomics.* **21:**695.

Soo, R.M., Hemp, J., Parks, D.H. *et al.* (2019) On the origins of oxygenic photosynthesis and aerobic respiration in Cyanobacteria. *Science.* **355:**1436–1440.

Sousa, F.L., Alves, R.J., Ribeiro, M.A. *et al.* (2012) The superfamily of heme-copper oxygen reductases: types and evolutionary considerations. *Biochim. Biophys. Acta.* **1817:**629–637.

Stolper, D.A. & Keller, C.B. (2018) A record of deep-ocean dissolved O_2 from the oxidation state of iron in submarine basalts. *Nature.* **553:**323–327.

Stewart, W.D. & Lex, M. (1970) Nitrogenase activity in the blue-green alga Plectonema boryanum strain 594. *Arch Microbiol.* **73:**250–260.

Stone, J., Edgar, J.O., Gould, J.A. *et al.* (2022) Tectonically-driven oxidant production in the hot biosphere. *Nat Commun.* **13:**4529.

Stull, G.W., Qu, X., Parins-Fukuchi, C. *et al.* (2021) Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. *Nat Plants.* **7:**1015–1025.

Sweetman, A.K., Smith, A.J., de Jonge, D.S.W. *et al.* (2024) Evidence of dark Oxygen production at the abyssal seafloor. *Nat Geosci.* **17:**737–739.

Tettelin, H., Masignani, V., Cieslewicz, M.J. *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome *Proc Nat Acad Sci U S A*. **102:**13950–13955.

Tettelin, H., Riley, D., Cattuto, C. *et al.* (2008) Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol.* **11:**472–477.

Touchon, M., Hoede, C., Tenaillon, O. *et al.* (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLOS Genet.* **5:**e1000344.

Touchon, M., Perrin, A., Moura de Sousa, J.A. *et al.* (2020) Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLOS Genet***. 16:**e1008866.

Tourova, T.P., Kuznetzov, B.B., Novikova, E.V. *et al.* (2001) Heterogeneity of the Nucleotide Sequences of the 16S rRNA genes of the Type Strain of *Desulfotomaculum kuznetsovii*. *Microbiol.* **70:**678–684.

Treangen, T.J. & Rocha, E.P.C. (2011) Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PloS Genet.* **7:**e1001284.

Tria, F.D.K. & Martin, W.F. (2021) Gene duplications are at least 50 time less frequent than gene transfers in prokaryotic genomes. *Genome Biol Evol.* **13:**evab224.

Trost, K., Knopp, M.R., Wimmer, J.L.E. *et al.* (2024) A universal and constant rate of gene content change traces pangenome flux to LUCA. *FEMS Microbiol Lett.* **371:**fnae068.

Trost, K., Gennis, R.B., Allen, J.F. *et al.* (2026) Oxygen reductase origin followed the great oxidation event and terminated the Lomagundi excursion. *BBA – Bioenergetics* **1867:**149575.

Vernikos, G., Medini, D., Riley, D.R. *et al.* (2015) Ten years of pan-genome analyses. *Curr Opin Microbiol.* **23:**148–154.

Wayne, L.G., Brenner, D.J., Colwell, R.R. *et al.* (1987) Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *IJSB.* **37:**463–464.

Whittaker, R.H. (1969) New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. *Science.* **163:**150–160.

Wielgoss, S., Didelot, X., Chaudhuri, R.R. *et al.* (2016) A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium *Myxococcus xanthus*. *ISME J.* **10:**2468–2477.

Wikstrom, M.K.F. (1977) Proton pump coupled to cytochrome c oxidase in mitochondria. *Nature.* **266:**271–273.

Woese, C.R. & Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Nat Acad Sci U S A.* **74:**5088–5090.

Woese, C.R. (1987) Bacterial evolution. *Microbiol Mol Biol Rev.* **51:**221-271.

Wolf, Y.I., Makarova, K.S., Lobkovsky, A.E. *et al.* (2016) Two fundamentally different classes of microbial genes. *Nat Microbiol.* **2:**1–6.

Yang, C., Chowdhury, D., Zhang, Z. *et al.* (2021) A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput Struct Biotechnol J.* **19:**6301–6314.

Yarza, P., Richter, M., Peplies, J. *et al.* (2008) The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol.* **31**:241–250.

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

Danke

Zunächst gilt mein besonderer Dank Prof. Dr. William F. Martin, der mir bereits während meines Bachelors die Möglichkeit gegeben hat, an seinem Institut zu arbeiten, und mich auch während meines Masters sowie meiner Promotion stets unterstützt und gefördert hat. Ein herzliches Dankeschön gilt zudem meinem Zweitgutachter Prof. Dr. Sven B. Gould.

Ein weiterer Dank geht auch an alle Koautor:innen, mit denen ich das Glück hatte, zusammenzuarbeiten. Außerdem möchte ich mich bei meinen Korrekturleser:innen Nico Bremer, Luca Modjewski, Natalia Mrnjavac, Nils Kapust, Loraine Schwander und Nadja Hoffmann bedanken, deren sorgfältige Anmerkungen und Hinweise wesentlich zur Verbesserung dieser Arbeit beigetragen haben.

Ein besonderer Dank gilt außerdem dem Spaßbüro und dem gesamten MolEvol-Team. Danke für die gute Stimmung, den Zusammenhalt und für viele Grillpartys bei Wind und Wetter. Ohne Euch wäre die Zeit im Institut nur halb so schön gewesen.

Zuletzt danke ich meinem Ehemann Rico, meiner Mutter Annette und meiner Schwester Charlotte, die mich stets unterstützt und motiviert haben während meiner ganzen Schul-, Universitäts- und Promotionszeit. Euer Rückhalt hat diese Arbeit überhaupt erst möglich gemacht.