Phylogenetische Analyse prokaryotischer Genome zur Rekonstruktion des letzten gemeinsamen Vorfahrens der Prokaryoten

Inaugural - Dissertation

zur Erlangung des Doktorgrades der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Madeline Chantal Weiß

aus Geseke

Düsseldorf, Februar 2020

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

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

Berichterstatter:

1. Prof. Dr. William F. Martin

2. Prof. Dr. Sven Gould

Tag der mündlichen Prüfung: 28.05.2020

Diese Arbeit ist meinem im Mai 2013 verstorbenen Vater
Diplom Informatiker Thomas Christian Weiß
gewidmet sowie meiner im Dezember 2018 verstorbenen Großmutter Anna

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

Publikationen in Fachzeitschriften

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

Weiss MC, Neukirchen S, Roettger M, Mrnjavac N, Nelson-Sathi S, Martin WF, Sousa FL: Reply to 'Is Luca a thermophilic progenote?' Nat Microbiol 1: 16230 (2016).

Martin WF, **Weiss MC**, Neukirchen S, Nelson-Sathi S, Sousa FL: Physiology, phylogeny and LUCA. Microb Cell 3: 582–587 (2016).

Martin WF, Zimorski V, **Weiss MC**: Wo lebten die ersten Zellen – und wovon? Biologie in unserer Zeit 47: 186–192 (2017).

Weiss MC, Preiner M, Xavier JC, Zimorski V, Martin WF: The last universal common ancestor between ancient Earth chemistry and the onset of genetics. PLoS Genet 14: e1007518 (2018).

Barth C, **Weiss MC**, Roettger M, Martin WF, Unden G: Origin and phylogenetic relationships of [4Fe-4S]-containing O_2-senors of bacteria. Environ Microbiol 20:4567–4586 (2018).

[*]Der Beitrag beider Autoren ist gleichwertig

Posterpräsentationen

Madeline C. Weiss, Natalia Mrnjavac, Filipa L. Sousa, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi und William F. Martin. Luca's informational core. European Bioenergetics Conference (EBEC), 2016. Riva del Garda, Italien.

Natalia Mrnjavac, Madeline C. Weiss, Filipa L. Sousa, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi und William F. Martin. Looking into Luca's cofactors. European Bioenergetics Conference (EBEC), 2016. Riva del Garda, Italien.

Madeline C. Weiss, Natalia Mrnjavac, Filipa L. Sousa, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi und William F. Martin. Luca's informational core. Molecular Biology of Archaea 5 (MAB5), 2016. London, Vereinigtes Königreich.

Madeline C. Weiss, Natalia Mrnjavac, Filipa L. Sousa, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi und William F. Martin. Luca's informational core. Black Forest Summer School (BFSS), 2016. Herzogenhorn, Deutschland.

Madeline C. Weiss und William F. Martin. Charting the gene set of the last universal common ancestor. Society for Molecular Biology and Evolution (SMBE), 2018. Yokohama, Japan.

Madeline C. Weiss und William F. Martin. Charting the gene set of the last universal common ancestor. VII Congress and Associate Symposiums of Vavilov Society of Geneticists and Breeders, 2019. Sankt Petersburg, Russland.

Vorträge

Madeline C. Weiss, Filipa L. Sousa, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi und William F. Martin. Hydrothermal vents, the origin of life, and the physiology and habitat of Luca.Workshop "Astrobiology – Life in the Context of Cosmic Evolution", 2016. Berlin, Deutschland.

Madeline C. Weiss, Filipa L. Sousa, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi und William F. Martin. Physiology and habitat of the last universal common ancestor (Luca). Gastvorträge, 2017. Palmerston Nord und Dunedin, Neuseeland.

Inhaltsverzeichnis

1 Zusammenfassung

Die Idee phylogentische Methoden zur Identifikation von evolutionären Zusammenhängen zu verwenden ist bereits sehr alt. Bereits Darwin (1837) und Haeckel (1866) verwendeten Stammbäume, um Zusammenhänge zwischen Organismen und sogar Reichen herzuleiten. Durch die Entdeckung der DNA-, RNA- und Proteinstrukturen (Zuckerkandl und Pauling, 1962; Woese und Fox, 1977) konnten genauere Stammbäume auch auf Molekularebene berechnet werden. Die Biologie wurde durch die neuen Analysemethoden in unterschiedliche Lager gespalten, zum einen die Verfechter des Drei-Domänen-Baumes (Woese und Fox, 1977), welche einen gemeinsamen Vorfahren für die Prokaryoten, also Bakterien und Archaeen, sowie für die Eukaryoten postulieren. Dem gegenüber stehen die Befürworter des Zwei-Domänen-Baumes (Williams *et al.*, 2013). Letzterer impliziert, dass es einen gemeinsamen Vorfahren für die Prokaryoten gibt und die Eukaryoten einen eigenen Vorfahren besitzen. Die Eukaryoten sind laut dieser These aus einer Symbiose zwischen einem Bakterium und einem Archaeum entstanden (Mereschkowsky, 1905; Martin und Kowallik, 1999; Martin und Russell, 2003).

Durch weiterentwickelte phylogenomische Analysemethoden und größere Datenbanken (Weiss *et al.*, 2016a) wird die Theorie des Zwei-Domänen-Baumes weiter unterstützt. Basierend auf den Erkenntnissen aus dem Zwei-Domänen-Baum, wurde für die Rekonstruktion des Proteoms des letzten gemeinsamen Vorfahren (engl. *last universal common ancestor*, LUCA) nur prokaryotische Proteine verwendet.

Es ist interessant zu wissen, wie und wo LUCA gelebt hat, sowie welche Physiologie LUCA hatte. Dazu muss zunächst geklärt werden, wo LUCA evolviert ist. Eine mögliche Erklärung ist die Hydrothermalquellen Theorie von Martin und Russell (2003). Demnach ist LUCA möglicherweise in den Wasserkammern einer 70-90°C heißen, alkalischen Hydrothermalquelle (engl. *hydrothermal vent*) entstanden. Die Wände der Hydrothermalquelle besitzen eine poröse Struktur und befinden sich in der Nähe von Magmarkammern. Durch Serpentinisierung wird der für chemische Energie benötigte Wasserstoff generiert (Martin und Russell, 2007; Wächtershäuser, 2006).

Die Genzusammensetzung von LUCA anhand von molekularen Informationen zu rekonstruieren, wurde öfter von unterschiedlichen Wissenschaftlern durchgeführt. Demnach besitzt LUCA einen Kern an Proteinen, der für die Informationsweitergabe zuständig ist. Zu diesen Proteinen gehören Ribosomen. In unterschiedlichen Analysen wurde jeweils nach den Proteinen gefiltert, welche in allen Prokaryoten des jeweiligen Datensatzes auftraten. Hierbei konnten etwa 30 (Charlebois und Doolittle, 2004) bis 100 (Puigbò *et al.*, 2010) Proteine identifiziert werden. In einer weiteren Analyse wurde festgestellt, dass LUCA tatsächlich der Vorfahre der Archaeen und Bakterien sein kann (Williams *et al.*, 2013). Allerdings wird dadurch, dass nur nach universell oder nahezu universell auftretenden Proteinen

in den bisherigen Untersuchungen gesucht wurde, nur 1% eines durchschnittlichen Genoms für die Analyse verwendet (Dagan und Martin, 2006). Die Informationen über die Physiologie eines Organismus sind in den übrigen 99% enthalten (Dagan und Martin, 2006). Des Weiteren können die 1% Proteine nichts über die Proteine einer ganzen prokaryotischen Gruppe aussagen (Nelson-Sathi *et al.*, 2015).

Um weitreichendere Informationen über die Physiologie von LUCA zu erhalten, wurde in dieser Arbeit ein Datensatz von 1.981 prokaryotischen Genomen untersucht. Es wurden Proteinfamilien und phylogenomische Stammbäume für die 6,1 Millionen Protein kodierenden Gene der prokaryotischen Genome erstellt, um LUCA zu rekonstruieren. Durch Anwendung von phylogenomischen Filtern konnten 355 Proteinfamilien (ungefähr 0,1% der Gesamtmenge) identifiziert werden, welche auf LUCA zurück zu führen sind.

Neben den bekannten Proteinen für die Informationsverarbeitung, wie Ribosomen, wurden auch Proteine identifiziert, die zu Stoffwechselprozessen, zum Transport und zu zellbezogenen Prozessen gehören. Des Weiteren wurden auch sauerstoffbezogene Proteine gefunden. Diese können darauf hinweisen, dass die Filter nicht alle LGT-Events rausgefiltert haben oder die Annotation falsch ist. Weiterhin wurden Hinweise auf Cofaktoren wie S-adenosyl-methionin (SAM), Eisen-Sulfur-Cluster (FeS-Cluster) und Adenosintriphosphat (ATP) in den Daten gefunden. Diese Cofaktoren verhelfen Proteinen der RNA-Modifikation, Energiestoffwechsel, Kohlenstoffassimilierung und Stickstoffassimilierung zur richtigen Funktion (Weiss *et al.*, 2016b). Der Cofaktor SAM ist bekannt dafür, dass er eine zentrale Rolle in den frühen Stoffwechselprozessen besitzt und wichtig für die Interaktion zwischen tRNA-mRNA-rRNA ist (Broderick *et al.*, 2014; Weiss *et al.*, 2016b). Weiterhin wird aus den Daten ersichtlich, dass Methylgruppen einen sehr frühen Ursprung haben. Daher ist davon auszugehen, dass die Ribosomenmodifikation ebenfalls sehr ursprünglich ist. Neben Dehydrogenasen und Synthetasen, welche wichtig für die Kohlenstoffassimilation sind, wurden auch Proteine identifiziert, die Teil des Wood-Ljungdahl Stoffwechselweges sind. Ableitend von den analysierten Daten kann davon ausgegangen werden, dass LUCA ein anaerober und autotropher Organismus war, welcher von Wasserstoff, Kohlenstoffdioxid und Schwefelwasserstoffen abhängig war. Durch die Analyse der Distanzen in den erstellten Stammbäumen zwischen den Archaeen und Bakterien konnte festgestellt werden, dass die ursprünglichsten Prokaryoten acetogene Clostridien und methanogene Archaeen sind (Weiss *et al.*, 2016b). Diese Organismen haben die geringste Distanz zueinander und liegen am nächsten zu der Wurzel im Stammbaum.

Die Ergebnisse dieser Analyse können auch für die Bestimmung ursprünglicher Proteine bestimmter Domänen verwendet werden. Demnach sind die FeS-Cluster ein Indikator für Sauerstoffsensoren in Bakterien. Durch eine Analyse von ausgewählten Proteinen der Sensoren in Bezug der gefundenen LUCA Proteine konnte festgestellt werden, dass die Sensoren FNR, NreB und WhiB3 keine Homologie aufweisen und daher wahrscheinlich durch LGT verbreitet wurden (Barth *et al.*, 2018).

2 | Summary

The idea of using phylogenetic methods for the identification of evolutionary relationships between organisms is relatively old. The pioneers in this field are Darwin (1837) and Haeckel (1866). They were the first ones to use trees to determine a context between organisms and kingdoms. Due to the discovery of DNA-, RNA- and protein structures (Zuckerkandl und Pauling, 1962; Woese und Fox, 1977), more detailed trees could be calculated based on molecular structures. Thus, different hypotheses that describe the correlation between organisms based on diverse trees and rooting positions were introduced by biologists. On the one hand there were the scientists that wanted to describe life based on the three domain tree (Woese und Fox, 1977). In this tree, the root which can be seen as the ancestor, is placed between bacteria, archaea and eukaryotes. On the other hand there were the supporter of the two domain tree (Williams *et al.*, 2013). Here, the root is placed between archaea and bacteria. Hence, the two domain tree hypothesis, the eukaryotes arose from a symbioses between an archaeon and a bacterium (Mereschkowsky, 1905; Martin und Kowallik, 1999; Martin und Russell, 2003). Based on the interpretation of the two domain tree, only prokaryotic proteins were used to recreate the proteom of the last universal common ancestor (LUCA) of all prokaryotes.

The habitat and physiology of LUCA were also of interest. At first, before starting the analysis, the setting where LUCA arose had to be determined. One possible explanation is the hydrothermal vent theory by Martin und Russell (2003). They proposed that LUCA originated in the 70-90 °C water contained in the chimneys of alkalic hydrothermal vents. Those vents consist of porose rocky structures. The required hydrogen results from serpentinisation (Martin und Russell, 2007; Wächtershäuser, 2006).

The identification of LUCA's genetic composition based on molecular information was performed several times. Thus, LUCA contained proteins for informational processes. Those proteins are mainly ribosomal proteins. In different approaches, scientists have already looked at universal or nearly universal proteins in prokaryotes. They found between 30 (Charlebois und Doolittle, 2004) and 100 (Puigbò *et al.*, 2010) proteins that probably belong to LUCA. Another analysis showed that LUCA is truly the ancestor of bacteria and archaea (Williams *et al.*, 2013). However, the disadvantages of the previous analysis are that only around 1% of the average genome was analysed (Dagan und Martin, 2006). But if the analysis of the physiology is the main goal, the other 99% are important. Furthermore, there is no information about an entire prokaryotic group contained in the 1% (Nelson-Sathi *et al.*, 2015).

To get a more in-depth view on what LUCA's physiology was like, 1,981 prokaryotic genomes were analyzed. Protein families and phylogenetic trees were calculated for 6.1 million prokaryotic proteins of coding genes to reconstruct LUCA. After filtering based on

phylogenomic criteria 355 protein families (0.1%) were identified that trace back to LUCA.

An addition to already known proteins for informational processes like ribosomes, new proteins from several different pathways, for example transport, cellular prozesses, RNA-modification, carbon assimilation and many more were found. Some proteins are related to oxygen. Some proteins, which are related to oxygen, are caviates and can be explained by LGT events, that couldn't be filtered out, or misleading annotations. Furthermore, there is evidence for ancient cofactor related proteins, that need S-adenosyl-methionin (SAM), iron-sulfur cluster (FeS-cluster) and adenosintriphosphate (ATP). Those cofactors are known to play an important role in RNA-modification, energy metabolism, carbon assimilation and nitrogen assimilation (Weiss *et al.*, 2016b). It is also known that SAM plays a central role in ancient pathways and is important for the tRNA-mRNA-rRNA (Broderick *et al.*, 2014; Weiss *et al.*, 2016b) interactions. Furthermore methyl groups are ancient and are part of ribosome modifications. Next to dehydrogenases and synthetases, which are important for carbon assimilation, other proteins that are part of the ancient Wood-Ljungdahl pathways could be identified. In summary, there is evidence that LUCA was an anaerobic, autotrophic organism that could live from hydrogen, carbondioxid and hydrogen sulfide. On top of this, the most ancient prokaryotes were acetogenic clostridia and methanogenic archaea (Weiss *et al.*, 2016b).

The results of the described analysis can be used to identify ancient proteins in different domains. Therefore, FeS-clusters are indicators for oxygen sensing proteins in bacteria and can be used to identify the relationship between different proteins from a variety of organisms. Using the proteins that are present in LUCA and proteins that are related to oxygen sensing, it is possible to show that there were multiple origins or LGT events and the proteins FNR, NreB and WhiB3 are not homologous (Barth *et al.*, 2018).

3 Einleitung

3.1 Ursprung des Lebens und die Entstehung der Prokaryoten

Alle Lebewesen, ob Eukaryoten oder Prokaryoten, haben eine genetische Schnittmenge, die auf einen evolutionären Zusammenhang hinweist. Durch die Endosymbionten-Theorie (Mereschkowsky, 1905; Martin und Kowallik, 1999; Martin und Russell, 2003) kann der Ursprung der Eukaryoten, sowie der evolutionäre Zusammenhang von Eukaryoten und Prokaryoten beschrieben werden. Der Ursprung der Prokaryoten ist bis heute jedoch noch nicht vollständig geklärt. Da es keine fossilen Rückstände der ersten Zellen gibt, wird angenommen, dass heutzutage auftretende Organismen Rückschlüsse auf die ersten Zellen geben können.

Bereits seit mehreren Jahrzehnten werden Hypothesen aufgestellt wie die ersten Zellen entstanden sein könnten. Eine der ersten Hypothesen wurde von Oparin (1952) und Haldane (1929) unabhängig voneinander aufgestellt. Demnach sind die ersten Organismen in einer Flüssigkeit, der Ursuppe, entstanden (Oparin, 1952; Haldane, 1929), welche aus Gasen wie Methan und Ammonium, sowie Wasser bestand (Lane *et al.*, 2010). Später wurde vermutet, dass diese Stoffe aus dem Weltall stammten und durch einen Meteoriten auf die Erde kamen (Napier, 2004). Durch UV-Strahlung wurden die sich im Wasser befindenden Stoffe energetisch angeregt und zu organischen Komponenten umgewandelt. Diese Komponenten reagierten weiter miteinander und bildeten die ersten Moleküle, welche sich wiederum zu Makromolekülen zusammenschlossen und später die ersten Zellen bildeten. Einen experimentellen Beweis für diese Hypothese lieferte Miller (1953). Während dieser Experimente wurde eine elektrische Spannung auf ein Methan (CH_4), Ammoniak (NH_3) und Wasserstoff (H_2) Gasgemisch ausgeübt und so organische Komponenten wie HCN, Aldehyde, Aminosäuren, Öle und Teer generiert. Durch spätere Experimente konnte nachgewiesen werden, dass diese Stoffe Teil der Synthese von Aminosäuren sind (Bada, 2004; Martin *et al.*, 2008).

Die Ursuppen-Hypothese geht davon aus, dass es zunächst genetisches Material gab, welches der heutigen RNA oder Proteinen ähnelte (Orgel, 2004). Daher wird die Zeit als RNA-Welt (engl. *RNA world*) bezeichnet. Dem gegenüber wird angenommen, dass zunächst energiegewinnende Prozesse entstanden sind und sich danach die Zellbestandteile für die Informationsspeicherung entwickelt haben (engl. *metabolism first* (Oparin, 1952)). Eine Hypothese, die energiegewinnende Prozesse als ursprünglich ansieht, beschreibt den Ursprung der ersten Zellen an Hydrothermalquellen. Dazu muss zunächst zwischen zwei Arten von Hydrothermalquellen unterschieden werden. Zum einen die schwarzen Raucher (engl. *Black smokers*) und zum anderen die Lost City Systeme (Kelley *et al.*, 2002). Die schwarzen Raucher befinden sich meist direkt über Magmakammern, welche ein bis drei

Kilometer unter dem Meeresboden liegen. Das Wasser in den schwarzen Rauchern kann bis zu 405 °C heiß werden, da es in direktem Kontakt zu den Magmakammern steht. Das aus den Kaminen der schwarzen Raucher kommende Wasser ist sehr sauer (pH 2–3), reich an gelösten Metallen und enthält eine hohe Konzentration an Wasserstoff (Martin *et al.*, 2008). Die Lost City Systeme liegen mehrere Kilometer von den Magmakammern entfernt. Das Wasser in den Kaminen der Systeme hat Temperaturen von 70–90 °C und ist alkalisch (pH 9–10). Diese Art der Hydrothermalquellen hat eine längere Lebensspanne als die schwarzen Raucher. Sie können bis zu 100.000 Jahre aktiv sein (Martin *et al.*, 2014). Die Konzentration von Wasserstoff liegt in den Lost City Systemen bei ca. 10 mM und ist somit geringer als in den schwarzen Rauchern (Martin *et al.*, 2008; Sousa *et al.*, 2013). Der Wasserstoff, welcher in den schwarzen Rauchern und Lost Cities zu finden ist, stammt von einer Reaktion, welche Serpentinisierung genannt wird. Dabei reagieren die aus der Kruste der Hydrothermalquellen gelösten Metalle, wie zum Beispiel Eisen- und Magnesium-Silikate mit dem Meerwasser und dem darin gelösten Kohlenstoffdioxid (Sousa *et al.*, 2013; Martin *et al.*, 2014).

Die Hypothese der Hydrothermalquellen beschreibt den Vorfahren der heutigen Prokaryoten als chemiosmotischen Organismus, welcher das aus der Serpentinisierung stammende H_2 zur Energiegewinnung nutzte (Martin und Russell, 2007, 2003; Wächtershäuser, 2006). Bei der Serpentinisierung reagiert das Meerwasser mit Metallen im Meeresgrund und wird zu Wasserstoff reduziert. Des Weiteren entsteht bei dieser Reaktion der Stoff Serpentinit, welcher für die Namensgebung der Reaktion dient (Russell, 2007). Die Protozelle, also die ursprünglichste Zelle (de Duve, 2005), hatte die Fähigkeit zu wachsen, sich zu multiplizieren und gegen andere Protozellen zu konkurieren. Durch diese Konkurrenz zwischen den Protozellen fand wahrscheinlich eine Selektion statt (de Duve, 2005). Diese Selektion könnte zu der Entwicklung unterschiedlicher biochemischer Stoffwechselwege in Bakterien und Archaeen geführt haben. Des Weiteren weisen die unterschiedlichen Zellwandreaktionen sowie die konservierende Redoxchemie an der Zelloberfläche auf unabhängige Entwicklungen hin (Schleifer und Kandler, 1972).

Die ältesten bekannten Protozellen sind die Methanbilder (engl. *methanogens* (Balch *et al.*, 1979)) und Acetogenen (Sousa *et al.*, 2013). Die Methanbilder sind der Ursprung der Archaeen und reduzieren in der Methanogenese Kohlenstoffdioxid durch Wasserstoff zu Methan (Thauer, 1998). Die Acetogenen werden als der Ursprung der Bakterien angesehen und nutzen Wasserstoff und Kohlenstoffdioxid in mehreren Reaktionen, um Acetat als Energiequelle zu bilden (Poehlein *et al.*, 2012). Neben den unterschiedlichen Stoffwechselwegen zur Energiegewinnung unterscheiden sich die Archaeen und Bakterien auch in der Zellwand-Struktur. Die meisten Bakterien haben eine Zellwand aus Peptidoglycan (Kandler, 1995), wohingegen die Zellwand bei Archaeen sehr divers sein kann (Kandler, 1995). Die Zellwand der Archaeen kann aus Glycoprotein Untereinheiten sowie aus Lipoglycanen in unterschiedlichen Zusammensetzungen bestehen.

3.2 Der Baum des Lebens und seine Entstehung

Die Klassifizierung von Lebewesen hat eine jahrtausendalte Entstehungsgeschichte. Neben der bis jetzt noch nicht geklärten Frage „Woher stammt das Leben?“ ist die zweite wichtige Frage „Wie sind Lebewesen miteinander verwandt?“. Es gibt zwei Ansätze diese Fragen zu beantworten. Zum einen kann das Bestimmen von morphologischen Merkmalen, wie Aristoteles es gemacht hat, Verwandtschaften zwischen Lebewesen deutlich machen, oder evolutionäre Zusammenhänge von Lebewesen, wie Darwin es erkannte (Wheeler, 2012).

Der Philosoph Aristoteles stellte in seiner Schrift „*Kategorien*“ die These auf, dass Lebewesen in unterschiedlichen Gruppen eingeteilt werden können. Er klassifizierte die ihm bekannten Lebewesen anhand von morphologischen Merkmalen, zum Beispiel in Gangtiere, Zweifüßer, Flugtiere und Wassertiere. Neben seiner Art der Klassifizierung führte Aristoteles auch ein hierarchisches System ein, welches die Klassifikation nochmal verfeinerte. Die sehr einfache Methode von Aristoteles, Lebewesen in unterschiediche Gruppen zu ordnen, hat sich bis heute bewahrheitet, nur die Methode diese Gruppierungen zu erstellen hat sich verändert. Das hierarchische System von ihm hat sich bis heute in gewissen Zügen gehalten (Wheeler, 2012). Seine Arbeiten werden als Beginn der mordernen Klassifizierung angesehen (Wheeler, 2012).

Viele Wissenschaftler folgten der Theorie von Aristoteles und nutzten morphologische Merkmale sowie Habitate, um Lebewesen und Pflanzen zu klassifizieren. Zu diesen Wissenschaftlern zählte auch Carl von Linné. Er nutzte die Idee von Aristoteles und erschuf ein System, um Lebewesen und Pflanzen in Gruppen einzuordnen, sowie diese zu benennen. Dieses System von Carl von Linné hatte mehr als 200 Jahre Bestand in der Biologie (Wheeler, 2012). In seiner Publikation „*Systema Naturae*“ von 1758 stellte Carl von Linné zudem die bis heute genutzte hierarchische Nomenklatur vor:

- Imperium ()
- Regnum (Reich)
- Classis (Klasse)
- Ordo (Ordnung)
- Genus (Genus)
- Species (Spezies)
- Varietas ()

Jean-Baptiste Lamarck war nicht überzeugt von der Klassifikation nach Aristoteles und Carl von Linné. Er nutzte allerdings das Wissen aus den bekannten Methoden und modifizierte diese, um den ersten Stammbaum zu erstellen. Dieser wurde 1809 veröffentlicht und stellt eine Tabelle dar, welche Verbindungen zwischen den Organismen aufweist und somit wie ein Baum aussieht.

Charles Darwin (1859) war der erste Wissenschaftler, der die biologische Varianz durch evolutionäre Zusammenhänge beschrieb. In seinen Aufzeichnungsbüchern von 1837 skizzierte er einen Entwurf wie in Abbildung 3.1 zu sehen ist, der Verwandtschaften zwischen Lebewesen darstellte.

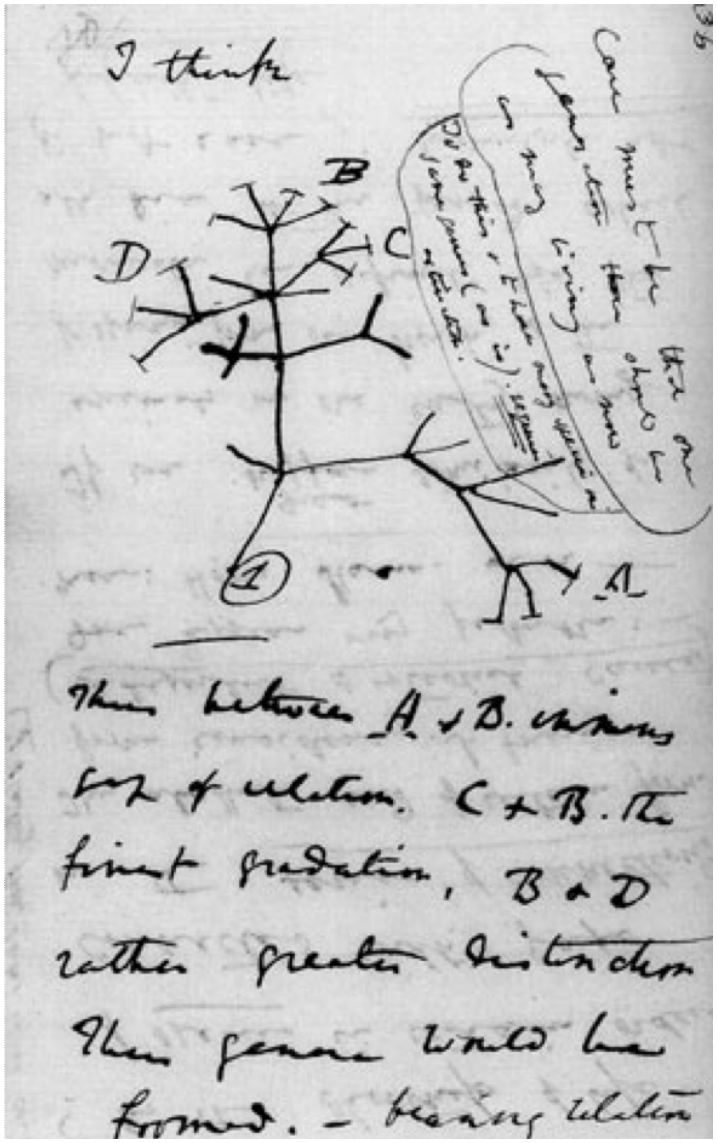


Abbildung 3.1 Skizze eines Stammbaumes mit evolutionären Zusammenhängen, gezeichnet von Charles Darwin in sein Notizbuch „B„ 1837. Zwischen den Ästen A und B ist keine Verwandtschaft zu sehen, die Äste C und B besitzen den höchsten Verwandtschaftsgrad und zwischen B und D ist auch ein eher großer Unterschied. Die 1 beschreibt die Wurzel des Stammbaumes (Charles Darwin, Notizbuch „B“ ,1837).

In seinem Werk „*The Origin of Species*“ veröffentlichte Darwin einen schematischen Stammbaum, welcher evolutionäre Zusammenhänge wiederspiegelte.

Ernst Haeckel (1866) erstellte einen Stammbaum, der evolutionäre Zusammenhänge und Aristoteles' natürliche Hierarchie vereinigte (Abbildung 3.2). Die Äste des Baumes sollen die genealogischen Verwandtschaften darstellen und die Distanz zur Wurzel spiegelt die

natürliche Varianz wieder. In der Wurzel befinden sich die Monera, welche den heutigen Prokaryoten ähneln, und in der Spitze wird der Mensch aufgeführt. Dazwischen befinden sich die Würmer, Mollusken, Tetrapoden, Mammalia und Primaten (Wheeler, 2012).

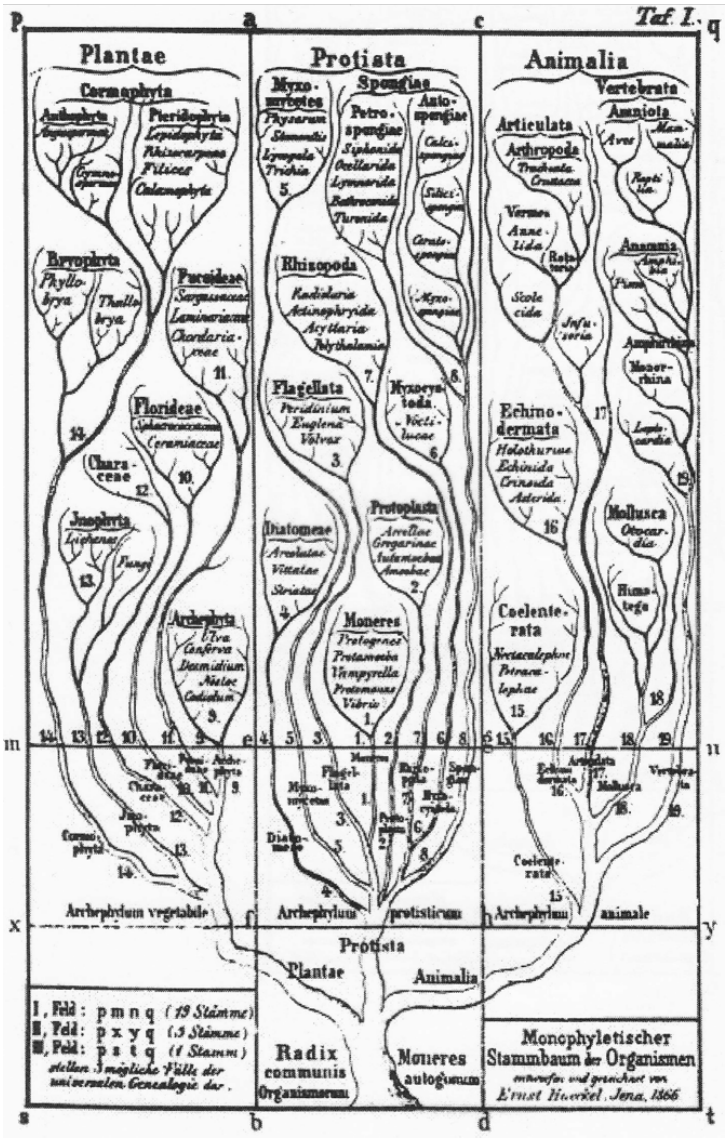


Abbildung 3.2 Zeichnung eines Stammbaumes von Haeckel, welcher die evolutionären Zusammenhänge sowie die natürliche Hierarchie verbindet. Als Wurzel werden die Monera beschrieben. Der Stammbaum ist in drei große Äste, Plantae, Protista und Animalia unterteilt („Monophyletischer Stammbaum der Organismen“, Haeckel 1866).

Im Zuge seines publizierten Baumes prägte Haeckel den Begriff der Monophylie. Seiner Ansicht nach werden alle Organismen, die einen gemeinsamen Vorfahren haben, als taxonomische Gruppe definiert, die monophyletisch ist. Nach späterer Kenntnis würden auch paraphyletische Organismen in diese taxonomischen Gruppen fallen. Laut Hennig (1950, 1966) müssen die Organismen einer taxonomischen Gruppe einen direkten, gemeinsamen Vorfahren haben, um als monophyletisch angesehen zu werden.

Das Klassifizieren von Organismen durch morphologische Merkmale stieß bei vielen Gruppen von Lebewesen schnell an seine Grenzen. Speziell im Bereich der Mikroorganismen waren morphologische Merkmale nur schwer zu unterscheiden, da prokaryotische Organismen keine komplexen intrazellulären Strukturen aufweisen und sich morphologisch nicht genug voneinander unterscheiden (Graur und Li, 2000). Daher führte die Entdeckung der DNA und ihrer Struktur als Träger der Erbinformationen aller Lebewesen (Watson und Crick, 1953), als neuer Ansatz zur Klassifizierung von Organismen, zu neuen Möglichkeiten. Durch die Untersuchung bezüglich der am besten geeigneten Moleküle als Basis für die Rekonstruktion molekularer Stammbäume (Zuckerkandl und Pauling, 1962) wurde es möglich, verwandtschaftliche Beziehungen durch Gemeinsamkeiten und Unterschiede in vererbten Molekülen zu berechnen.

Ein weiterer Schritt zur Entschlüsselung der Verwandtschaftsverhältnisse waren die Entwicklung der Proteinsequenzierungsmethoden und die immer schneller werdende Sequenzierung von Nukleinsäuren (Sanger *et al.*, 1977). Durch diese Methoden wurden immer mehr Sequenzen generiert, die einen großen Einfluss auf die molekulare Phylogenetik hatten (Graur und Li, 2000). Kurze Zeit nach der Entdeckung der neuen Strukturen nutzten Woese und Fox (1977) ribosomale Ribonukleinsäuren (rRNA) von Eukaryoten und Prokaryoten und berechneten einen Baum des Lebens. Der Stammbaum von Woese und Fox (1977) ist drei geteilt. Zum einen werden die Prokaryoten in zwei einzelne Äste (Archaebakterien und Eubakterien) geteilt und zum anderen bilden die Eukaryoten den dritten Ast. Alle drei Gruppen haben demnach einen prokaryotischen Ursprung. Neben der rRNA wurden viele andere einzelne Proteine analysiert und Bäume generiert (Delsuc *et al.*, 2005). Diese Analysen führten zu sehr vielen Konflikten innerhalb der Stammbäume. Besonders bei der Betrachtung von einzelnen Proteinen bei den Prokaryoten können falsche Schlüsse gezogen werden, da Gene nicht nur vertikal vererbt werden, sondern auch horizontal (lateral) (Wolf *et al.*, 2002). Durch diese Organismen-spezifische-Eigenheit können spezifische Genverluste oder Zugewinne an Genen durch lateralen Gentransfer (LGT) bei den Prokaryoten auftreten. Daher ist es ratsam nicht nur einzelne Proteine aus einzelnen Organismen für die Rekonstruktion des „Baum des Lebens“ zu nutzen, sondern Proteine, die in allen Genomen vertreten sind oder zumindest in der Mehrheit der Genome (Wolf *et al.*, 2002).

Durch die Weiterentwicklung der Methode zur Rekonstruktion des Baum des Lebens und der Einbeziehung von mehr Daten konnte der Drei-Domänen-Baum von Woese und Fox (1977) weitestgehend widerlegt und als Zwei-Domänen-Baum neu dargestellt werden (Williams *et al.*, 2013). Demnach sind die Archaeen und Bakterien Schwestergruppen und haben einen gemeinsamen Vorfahren (engl. *last universal common ancestor*), jedoch nicht Archaeen und Eukaryoten wie vorherige Studien zeigten (Raymann *et al.*, 2015; Woese und Fox, 1977). Diese Vermutung stützt auch die Endosymbionten-Theorie, nach der die Eukaryoten durch eine Symbiose aus einem Archaeon und einem oder mehr Bakterien entstanden sind (Martin und Russell, 2003).

3.3 Phylogenomische Stammbaumanalysen und Methodik

Phylogenomische Analysen gehören zu dem Gebiet der vergleichenden Biologie (Miyamoto und Cracraft, 1991). Sie wurden eingeführt, da die phänotypische Einteilung der Organismen in unterschiedliche Reiche, Stämme und Ordnungen möglich war, allerdings nichts über ihren evolutionären Hintergrund aussagte. Um diesen Zusammenhang ermitteln zu können, wurden die Genotypen der Organismen miteinander verglichen. Durch diese Art der Analysen können auch ältere Sequenzen mit neueren verglichen und so Aussagen über genetische Veränderungen getroffen werden. Zu den Analyseverfahren gehören zum einen Sequenzvergleiche auf lokaler und globaler Ebene, sowie Clusteringverfahren, Alignierungsmethoden und Algorithmen zur Rekonstruktion von Stammbäumen.

3.3.1 Sequenzvergleiche

Durch die immer größer werdende Anzahl an Nukleotid- und Proteinsequenzen und den damit neu generierten Datenbanken, war es nicht mehr möglich die Sequenzen per Hand zu vergleichen. Daher wurden Computeralgorithmen entwickelt, die es ermöglichten, große Datenbanken an Sequenzen zu vergleichen und Homologien zwischen den Sequenzen festzustellen.

Needleman und Wunsch (1970) entwickelten einen Algorithmus, der zwei Sequenzen über die gesamte Länge miteinander verglich, um die genetische Ähnlichkeit dieser Sequenzen zueinander zu definieren. Dazu wurde eine iterative Matrixmethode zur Berechnung der Ähnlichkeiten verwendet (Smith und Waterman, 1981). Zunächst mussten unterschiedliche Ebenen von möglichen Vergleichen von Needleman und Wunsch (1970) festgelegt werden. Die kleinste Einheit der Vergleiche war demnach ein Paar von Aminosäuren. Dabei wurde je eine Aminosäure von jeder zu vergleichenden Sequenz miteinander verglichen. Demnach wird die größte Übereinstimmung zwischen zwei Sequenzen erzielt, wenn viele Aminosäuren einer Sequenz in der gleichen Reihenfolge, unter Berücksichtigung möglicher Deletionen, in der anderen Sequenz auftreten. Der beste Treffer beim Vergleich von Sequenzen untereinander wird als der Treffer mit der größten Übereinstimmung definiert und kann aus dem erstellten zweidimensionalen Array (Matrix) abgelesen werden (Needleman und Wunsch, 1970).

Während des Vergleiches werden die Sequenzen Aminosäure für Aminosäure beim N-Terminus beginnend verglichen und für eine Übereinstimmung wird eine 1 in der Matrix zugewiesen, ansonsten eine 0 (Abbildung 3.3a). Um die Sequenzen mit der höchst möglichen Übereinstimmung zu alignieren, können auch Lücken (engl. *gaps*) eingefügt werden. Durch mögliche Strafpunkte (engl. *penalty point*) soll verhindert werden, dass das Ergebnis zu viele oder zu lange Lücken aufweist. Zur Bestimmung der Stellen an denen eine Lücke eingefügt werden muss, wird die Matrix vom C-Terminus beginnend, Richtung N-Terminus durchlaufen und der maximale Wert bestimmt (Abbildung 3.3b). Durch mehrmals vorkommende Aminosäuren ist es möglich, mehrere Wege durch die Matrix mit den gleichen Werten zu finden. Bei diesen Ereignissen wird ein Weg zufällig vom Algorithmus gewählt.

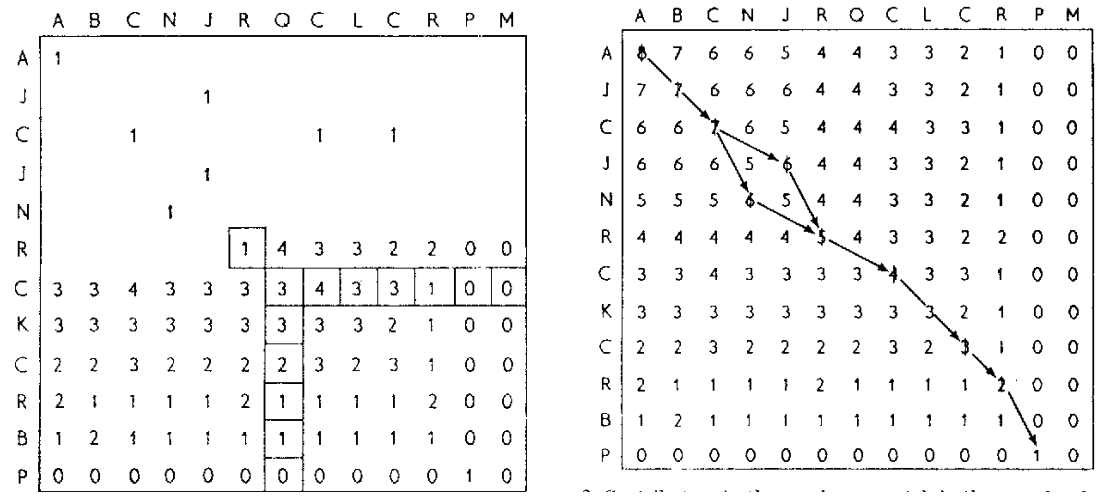


Abbildung 3.3 Beispiel für die Bestimmung eines Alignments mit Hilfe des Needleman-Wunsch-Algorithmus. (a) Die Zahl in jeder Zelle der Matrix stellt die größte Anzahl an identischen Paaren dar, wenn die Zelle als Ursprung eines Weges durch die Matrix angesehen wird und ansteigenden Indices. Identische Aminosäuren erhalten den Wert eins. Wenn keine Übereinstimmungen vorhanden sind, wird den leeren Zellen der Wert Null zugeordnet. Nachdem alle Zellen mit 0 und 1 befüllt sind, werden sie aufsummiert, beginnend mit der letzten Reihe. Die Bestimmung der Reihe R ist noch nicht ganz abgeschlossen. Durch die Übereinstimmung von dem Aminosäure Paar R-R wird eine 1 zu der aus der vorherigen stammenden Reihe vier addiert und der neue Wert fünf eingetragen. (b) Das Alignment der beiden Sequenzen wird durch alternative Wege in der Matrix dargestellt. Beginnend beim C-Terminus wird vom größten Wert ausgehend das Alignment durch die Matrix bestimmt. Wenn in einer Spalte nicht der nächst kleinere Wert vorzufinden ist, wird diese übersprungen und eine Lücke in das Alignment eingefügt. Die Abbildungen stammen aus Needleman und Wunsch (1970).

Eine Weiterentwicklung des Needleman und Wunsch Algorithmus' zur Berechnung von globalen Identitäten wurde von Smith und Waterman (1981) vorgenommen. Demnach muss nicht mehr die gesamte Länge der zu vergleichenden Sequenzen homolog sein, um feststellen zu können, dass sie die gleichen Funktionen ausführen, sondern nur bestimmte Abschnitte der Sequenzen. Lokale Bereiche der zwei zu vergleichenden Sequenzen müssen laut Smith und Waterman (1981) eine höhere Ähnlichkeit haben als eine der Sequenzen zu einer anderen.

Durch die immer größer werdenden Sequenzdatenbanken war es notwendig, die bekannten Sequenzvergleichsalgorithmen anzupassen. Der Needleman-Wunsch-Algorithmus für globale Vergleiche sowie der Smith-Waterman-Algorithmus für lokale Vergleiche waren nur für Vergleiche von zwei Sequenzen miteinander programmiert. Daher publizierten Wilbur und Lipman (1983) einen Algorithmus, der auf dem Needleman-Wunsch-Algorithmus, sowie dem Smith-Waterman-Algorithmus basierte, allerdings für Datenbanken modifiziert und FASTA genannt wurde. Um den Algorithmus zu beschleunigen, werden die Sequenzen nicht mehr Paar für Paar überprüft, sondern über eine bestimmte Wortlänge, den k-Tupeln (Dumas und Ninio, 1982). Durch das Verwenden von k-Tupeln werden die Sequenzen in mehrere Teilstücke unterteilt und anhand der ersten Aminosäuren entschieden, ob die Ami-

nosäuren in der Sequenz vorkommen. Bei einer Übereinstimmung wird ein Punkt in der Matrix gesetzt. Danach werden Diagonalen mit dem höchsten Wert in der Matrix gesucht und wenn nötig über Lücken miteinander verbunden. Als letzter Schritt wird ein lokales Alignment für den Bereich mit den Diagonalen durch dynamische Programmierung erstellt (Abbildung 3.4).

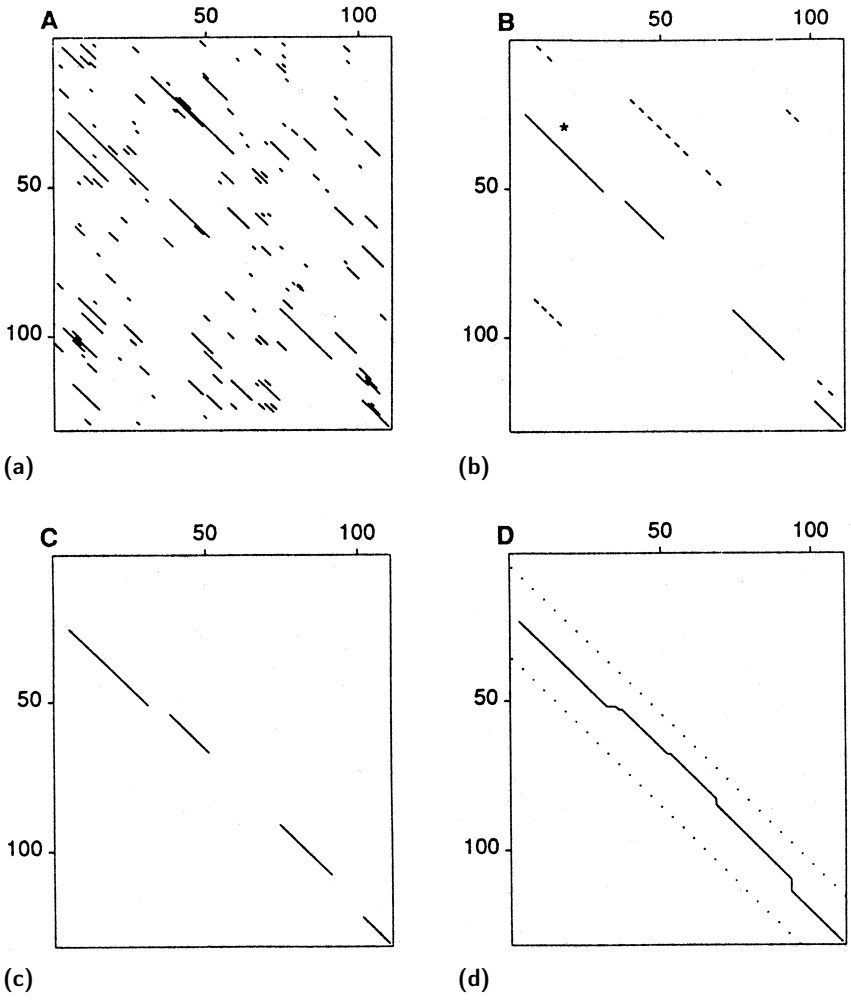


Abbildung 3.4 Beispiel für die Identifikation von Sequenzähnlichkeiten mit FASTA. Die vier Matrizen zeigen die Berechnung eines optimalen Alignments zwischen zwei Sequenzen. (a) Identische Regionen der zwei Sequenzen werden durch schwarze Punkte in der Matrix gekennzeichnet. (b) Durch eine Bewertungsmatrix werden die Regionen mit der höchsten Identität in der Matrix identifiziert und die Region mit dem höchsten Wert gekennzeichnet (*). Die Regionen, die den Schwellenwert nicht erreichen, werden gestrichelt dargestellt. (c) Regionen, welche unter dem Schwellenwert liegen, werden aus der Matrix entfernt. Aus den übergebliebenen Regionen wird das Alignment erstellt. (d) Die Regionen mit den besten Werten werden miteinander verbunden und bilden das optimale Alignment. Die gepunktete Linie stellt die Grenzen des optimalen Alignments dar. Die Abbildungen stammen aus Pearson und Lipman (1988).

Der FASTA-Algorithmus wurde von Lipman und Pearson (1985) für Proteinsequenzen modifiziert. Wenige Jahre später wurde der Algorithmus auch für Vergleiche von Nukleotidsequenzen zu Proteindatenbanken und umgekehrt angepasst (Pearson, 1990).

Die Anpassungen, die bei diesem lokalen Vergleichsalgorithmus vorgenommen wurden, sind sehr stringent. Als Treffer werden nur Übereinstimmungen gewertet, wenn die zu vergleichenden Aminosäuren oder Nukleotide identisch sind. Dadurch können biologisch relevante Proteine nicht richtig miteinander aligniert werden, denn Proteine können unterschiedliche Aminosäuren enthalten, allerdings trotzdem homolog sein. Um diesen Aspekt zu minimieren wurde das Basic Local Alignment Search Tool (BLAST) entwickelt (Altschul *et al.*, 1990).

Bei dem BLAST-Algorithmus handelt es sich um einen heuristischen Sequenzvergleichs-Algorithmus. Er kann wie FASTA für Nucleotidsequenz- sowie Proteinsequenz-Suchen in Datenbanken verwendet werden und wird für Motivsuchen, Genidentifikationen und in der Analyse von mehrfach auftretenden Regionen in ähnlichen sowie langen Nukleotidsequenzen verwendet (Altschul *et al.*, 1990). Genau wie bei dem FASTA-Algorithmus wird die Sequenz in k-Tupel geteilt und mit der Datenbank verglichen. Bei der BLAST-Suche werden nicht nur identische Aminosäuren gewertet, sondern auch biologisch ähnliche Aminosäuren können den Trefferwert beeinflussen. Des Weiteren arbeitet der BLAST-Algorithmus nach der Zwei-Treffer-Methode. Das bedeutet, dass Treffer nur gewertet werden, wenn sich ein weiterer Treffer auf der Diagonalen befindet (Altschul *et al.*, 1997). Eine weitere Berechnung, die vom BLAST-Algorithmus durchgeführt wird und ausschlaggebend für die Identifikation von Proteinen ist, ist der Erwartungswert (engl. *e-value*). Dieser Wert gibt an, ob die Sequenzen nur Zufallstreffer sind oder wirkliche Treffer. Je kleiner der Erwartungswert ist, desto wahrschenlicher ist es ein echter Treffer und kein zufälliger.

3.3.2 Clustering

Die Analyse von homologen Proteinen kann am besten durchgeführt werden, wenn die durch die globalen und lokalen Alignmentmethoden identifizierten homologen Proteine zu Proteinfamilien zusammengefasst werden. Dazu werden Clusteringmethoden verwendet.

Eine bekannte Clusteringmethode ist die graphenbasierte-Clustering-Methode (Jain *et al.*, 1999). Ein Graph besteht aus Vertices (Knoten), welche durch Kanten miteinander verbunden sind. Beim Graphenclustering werden die Vertices eines Graphen mit Berücksichtigung der Kanten zusammengefasst. Dabei sollen sich viele Kanten innerhalb des Clusters befinden und verhältnismäßig wenige Verbindungen zwischen den Clustern vorhanden sein (Schaeffer, 2007b).

Bei der graphenbasierten-Methode wird meistens zunächst ein Baum auf Grundlage der Daten designt, der die geringste Distanz zwischen den Datenpunkten aufweist. Häufig werden die lokalen beziehungsweise globalen Identitäten oder die Erwartungswerte von Proteinpaaren als Datenpunkte verwendet. Der so generierte Baum wird als *minimal spanning tree* (MST) bezeichnet. Durch das Löschen der längsten Verbindungen werden unterschiedliche Cluster erstellt (Zahn, 1971) (Abbildung 3.5).

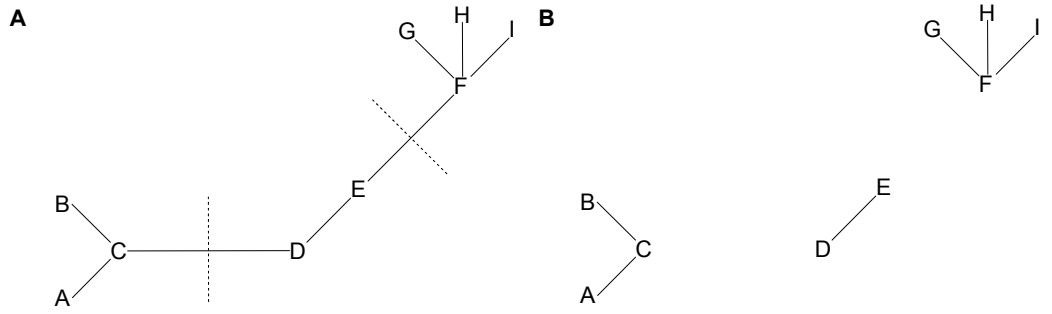


Abbildung 3.5 (A) Clusterbildung mit Hilfe eines MST. Die gestrichelten Linien zeigen an, wo Cluster gebildet werden können. Die längste Verbindung befindet sich zwischen C,D. (B) Dadurch werden A,B,C zu einem Cluster zusammengefasst. Ein zweites Cluster besteht aus D,E und das letzte Cluster setzt sich aus F,G,H,I zusammen. Die Abbildung ist modifiziert nach Jain *et al.* (1999).

Ein weiterer in der Biologie häufig verwendeter Clusteralgorithmus ist das agglomerative, hierarchische Clusterverfahren. Die Datenpunkte werden dabei so in einzelne Cluster aufgeteilt, dass ein gewurzelter, binärer Baum (Dendogramm) entsteht. Das hierarchische Clusterverfahren (Abbildung 3.6) kann auch für die Stammbaumanalyse verwendet werden (Jain *et al.*, 1999).

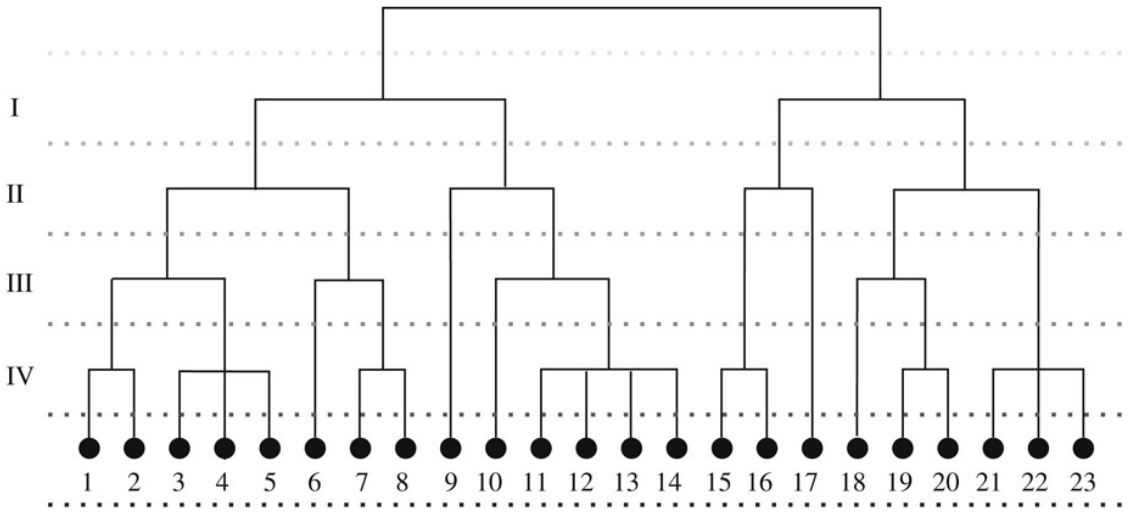


Abbildung 3.6 Dargstellt ist ein Dendogramm eines hierarchischen Cluster-Verfahrens. Die 23 Elemente stellen einzelne Blätter dar und wurden über vier Schritte in Cluster aufgeteilt. Die unterschiedlichen Schritte sind durch gepunktete Linien gekennzeichnet. In jedem Schritt werden die Blätter weiter zusammen gruppiert und neue Cluster gebildet. Diese Abbildung ist ein Beispiel für den nicht so häufig genutzten divisiven Algorithmus. Die Abbildung stammt aus Schaeffer (2007a).

Bei dem agglomerativen oder "*bottom-up*" Cluster-Verfahren, werden die Datenpunkte

auf der untersten Ebene des Baumes als *Sequenz* dargestellt. Diese Datenpunkte oder auch Blätter genannt, sind bereits eigenständige Cluster und werden als "*Singletons*" bezeichnet. Im ersten Schritt werden alle Blätter mit der geringsten Unterscheidung zueinander zu neuen Clustern zusammengefügt. Dieser Schritt wird so lange wiederholt, bis das Wurzelcluster entsteht. Dieses letzte Cluster umfasst alle Datenpunkte (Jain *et al.*, 1999). Der in der Biologie selten genutzte divisive Algorithmus oder "*top-down*" Algorithmus, beginnt bei dem Wurzelcluster. Dieses umfasst alle Daten und jeder Datenpunkt ist ein eigenständiges Cluster (Schaeffer, 2007a). Das Wurzelcluster wird in mehreren Durchgängen immer weiter geteilt, bis die Cluster nicht weiter logisch geteilt werden können.

Bei beiden Ansätzen müssen die Distanzen zwischen den Clustern berechnet werden, um logische Cluster zu finden. Um diese Berechnungen durchführen zu können, gibt es verschiedene Ansätze:

- *single linkage,*
- *complete linkage,*
- *average linkage.*

Das einfachste Verfahren ist das *single linkage* Cluster-Verfahren (Florek *et al.*, 1951), welches auch als Nächste-Nachbar-Methode bekannt ist. Die Einordnung von zwei Clustern A und B wird anhand der minimalsten Distanz zwischen zwei Elementen a und b der Cluster bestimmt:

$$dist(A, B) = \min_{a \in A, b \in B} \{d(a, b)\} \tag{1}$$

Da immer nur ein Element pro Cluster für die Einteilung der Cluster benötigt wird, ist dieses Verfahren anfällig für Ausreißer. Des Weiteren bilden sich bei diesem Verfahren häufiger "kettenförmige" Cluster, da nur wenige Elemente benötigt werden, um zwei Cluster zusammen zu fassen (Jain *et al.*, 1999).

Bei dem *complete linkage* Verfahren sollen die negativen Effekte vom *single linkage* Verfahren vermindert werden, indem der Durchmesser der Cluster minimiert wird (Sørensen, 1948). Daher wird die Distanz zwischen zwei Clustern A und B durch die maximale Distanz zwischen zwei Elementen a und b der Cluster bestimmt:

$$dist(A, B) = \max_{a \in A, b \in B} \{d(a, b)\} \tag{2}$$

Durch dieses Verfahren können viele kleine Cluster mit gleichem Durchmesser entstehen. Das am häufigsten verwendete und rechenintensivste Verfahren ist das *average linkage* Verfahren. Dieses Verfahren ist auch als WPGMA (engl. *weighted pair-group method arithmetic average*) bekannt (Sokal und Michener, 1958). Durch das Einbeziehen aller Elemente der Cluster zur Berechnung des Abstandes und das daraus errechnete Mittel ist dieses Verfah-

ren genauer, aber auch zeitaufwendiger:

$$dist(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A, b \in B} d(a, b) \tag{3}$$

Durch das *average linkage* Verfahren werden weder "kettenförmige", noch kleine Cluster wie bei den beiden anderen Verfahren erzeugt. Allerdings ist dieses Verfahren für große Datenmengen nicht geeignet, da alle Verbindungen zwischen den Datenpunkten der Cluster berechnet werden.

Durch die Bestimmung von Distanzfunktionen und Ableitungen von Clustern aus Dendogrammen, sind die hierarchischen Clusteringverfahren ungenauer, da die Distanzfunktionen die Cluster beeinflussen sowie Schwellenwerte selbst gewählt werden. Um genauere Cluster erstellen zu können, wurde der Marcov-Cluster Algorithmus entwickelt.

Der Markov-Cluster Algorithmus wurde im Jahr 2000 von Stijn van Dongen entwickelt (Dongen, 2000) und basiert auf der Markov-Kette. Die Markov-Kette (nach Andrei Andrejewitsch Markov, 1856–1922) ist ein diskreter stochastischer Prozess (Schaeffer, 2007a), welcher den zukünftigen Zustand eines Datenpunktes nur abhängig vom derzeitigen Zustand ableitet (Schaeffer, 2007a). Demnach ist der Zustand der Markov-Kette zum Zeitpunkt t+1 nur von dem aktuellen Zustand zum Zeitpunkt t abhängig, die vorherigen Zustände haben keinen Einfluss.

Für die Biologie ist der Markov-Cluster Algorithmus bei der Generierung von Proteinfamilien und Genfamilien von großer Bedeutung (Enright *et al.*, 2002). Gene und Proteine, die aufgrund von Sequenzähnlichkeiten zueinander gehören, können durch den Algorithmus zusammengefasst und für Sequenzalignments und Stammbaumrekonstruktionen gesammelt werden. Dabei stellen die Proteine die Vertices in einem Graphen dar. Außerdem bilden die Treffer der lokalen oder globalen Sequenzanalysen (Abschnitt 3.3.1) die Kanten sowie die dazu gehörigen Identitätspaare das Ähnlichkeitsmaß. Durch die verwendeten Identitäten können sich innerhalb des Graphen Regionen bilden, in denen sich Vertices mit stärkeren Verbindungen, also kurze Kanten, befinden und auch Regionen mit losen Verbindungen zwischen den Vertices (Schaeffer, 2007b).

Das Durchlaufen eines Graphs kann einfach beschrieben werden: Zunächst wird der Vertex x zum Zeitpunkt 0 zufällig gewählt. Der nächste Vertex y wird abhängig vom Ähnlichkeitsmaß ausgewählt, das bedeutet, dass der Vertex y der dem Vertex x am ähnlichsten ist, am wahrscheinlichsten verbunden wird. An dem neuen Vertex y angelangt, werden alle Verbindungen mit dem Vertex x nicht weiter beachtet und nur noch der Zustand des Vertex y betrachtet (Abbildung 3.7). Dieser Vorgang wird so lange durchgeführt, bis keine offenen Kanten mehr vorhanden sind (Abbildung 3.7).

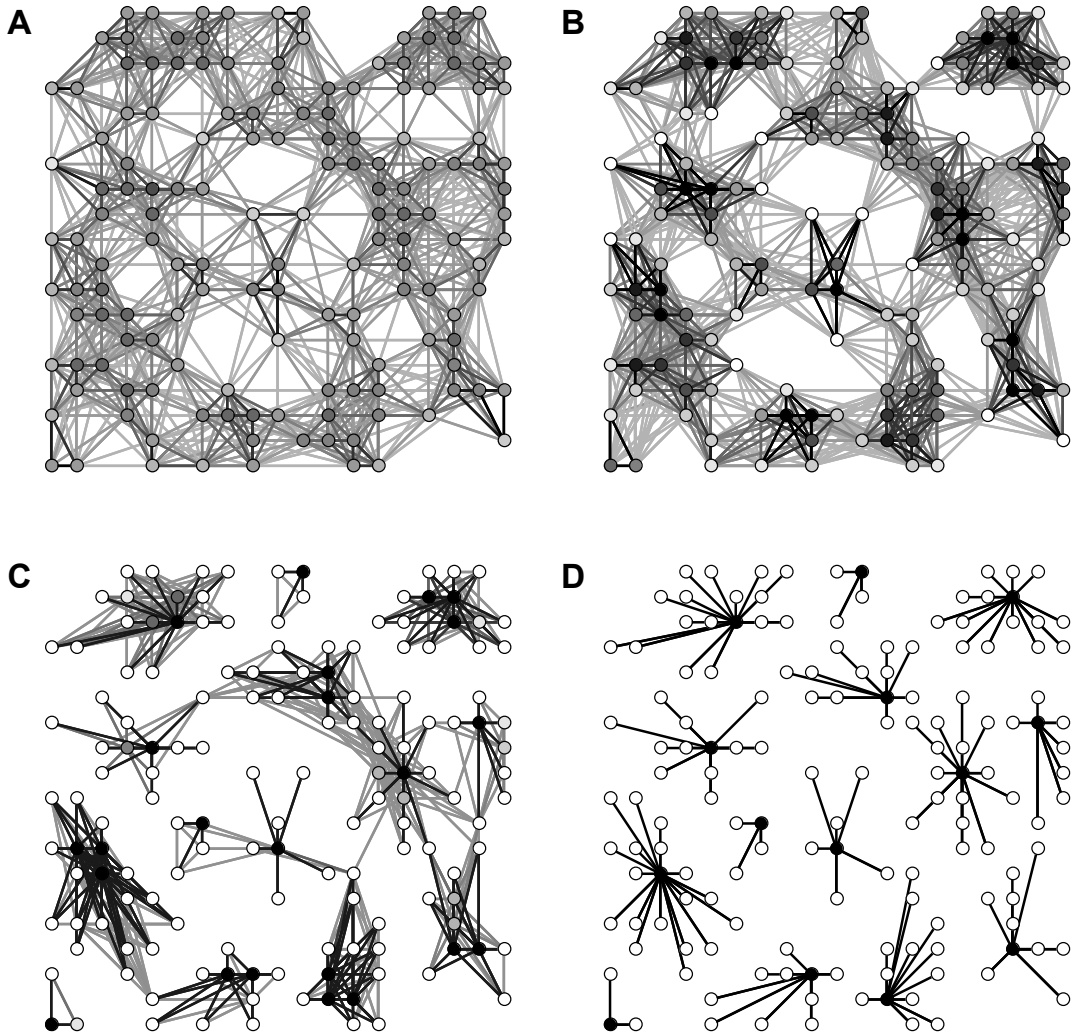


Abbildung 3.7 Dargestellt sind die unterschiedlichen Stadien während des MCL Prozesses. Am Anfang werden alle Datenpunkte mit Verbindungen dargestellt, wie sie aus der gegebenen Matrix abgelesen werden (A). Die unterschiedlichen Grautöne der Knoten (Kreise) geben die Anzahl der Verbindungen an. Je dunkler das Grau ist, desto mehr Verbindungen sind vorhanden. Die Stärke der Verbindungen wird ebenfalls durch unterschiedliche Grautöne angegeben. Starke Verbindungen werden dunkler angezeigt, als schwache. Durch das Durchlaufen der Graphen werden zufällige Knoten (schwarze Kreise), die kürzeste Verbindung zu den nächsten Knoten ausgewählt und verstärkt (B). Alle weiteren Verbindungen werden aus den Graphen gelöscht (C). Dieser Vorgang wird wiederholt, bis keine schwachen Verbindungen mehr vorhanden sind (D). Die Abbildung stammt aus van Dongen (2000).

3.3.3 Alignments

Um die Proteinfamilien graphisch darstellen zu können, müssen zunächst die Sequenzen miteinander verglichen und für die Stammbaumrekonstruktion vorbereitet werden. Der Vergleich von zwei Sequenzen für ein optimales Alignment kann durch den Needleman-Wunsch-Algorithmus durchgeführt werden. Ein Alignment für mehr als zwei Sequenzen ist komplizierter und rechenintensiver. Das Problem wächst exponentiell mit der Anzahl der gleichzeitig miteinander zu vergleichenden Sequenzen (Sievers *et al.*, 2011). Als mögliche Lösung stellt sich das progressive Alignment (Feng und Doolittle, 1987) dar.

3.3.4 Phylogenomische Stammbäume

Ein weiterer Schritt in der phylogenomischen Analyse eines Proteins, bzw. einer Proteinfamilie, ist die Erstellung eines Stammbaumes. Dargestellt werden die Proteine als operationale taxonomische Einheiten (engl. *operational taxonomic unit*, OTU (Sneath und Sokal, 1973)). Verbunden werden die OTUs durch Äste (engl. *branch*). Die Verbindung zwischen den Ästen wird als Knoten bezeichnet und stellt eine hypothetische taxonomische Einheit dar (engl. *hypothetical taxonomic unit*, HTU). Ein Knoten stellt eine Spezies dar, welche sich in zwei weitere teilt (De Soete, 1984). Das durch die Äste gebildete Muster wird als Topologie des phylogenomischen Stammbaumes bezeichnet. Durch diese Topologie können evolutionäre Zusammenhänge festgestellt werden. Des Weiteren kann an der Topologie abgelesen werden, ob Proteine homolog oder paralog entstanden sind. Die meisten Baumrekonstruktionsmethoden erstellen ungewurzelte Stammbäume. Mit der Wurzel wird der Punkt der Entwicklung des Proteins festgelegt. Es gibt zwei Zustände eines phylogenomischen Stammbaumes, zum einen der gewurzelte oder zum anderen der ungewurzelte.

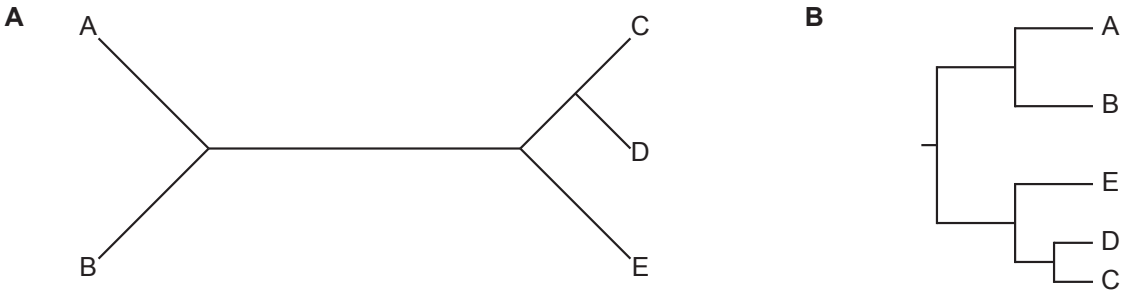


Abbildung 3.8 Zwei Zustände eines phylogenomischen Stammbaumes. (A) Der phylogenomische Stammbaum ist ungewurzelt. Die Wurzel kann an jedem Ast platziert werden. (B) Der phylogenomische Stammbaum ist gewurzelt. Die Wurzel wurde zwischen (A,B) und (E,(D,C)) eingefügt. Die Astlängen haben sich im Vergleich zu dem gewurzelten Stammbaum nicht geändert.

Um einen ungewurzelten Stammbaum zu wurzeln, können unterschiedliche Methoden verwendet werden. Des Weiteren kann auch berechnet werden, wie viele unterschiedliche Wurzeln es gibt. Für gewurzelte Bäume wird die Anzahl der möglichen Bäume (N_R) durch

unterschiedliche Wurzeln von n OTUs wie folgt berechnet, wenn $n \geq 2$:

$$N_{\mathrm{R}} = \frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{4}$$

Für ungewurzelte Bäume wird wie folgt die Anzahl an möglichen Bäumen (N_U) durch unterschiedliche Wurzeln für n OTUs berechnet, wenn $n \geq 3$ (Cavalli-Sforza und Edwards, 1967):

$$N_{\mathrm{U}} = \frac{(2n-5)!}{2^{n-3}(n-3)!} \tag{5}$$

Bei dem Mittelpunkt-Wurzeln (engl. *midpoint rooting*) wird die Mitte zwischen den beiden längsten Ästen als Wurzel festgelegt (Farris, 1972). Eine weitere Methode, die häufig für das Wurzeln von Stammbäumen angewendet wird, ist das Außengruppen-Wurzeln (engl. *outgroup rooting*). Bei dieser Methode wird eine Verzweigung des Stammbaumes gewählt und als Außengruppe verwendet (Maddison *et al.*, 1984). Häufig werden dazu taxonomische Gruppen verwendet, die in keiner direkten Verbindung zu den zu untersuchenden Organismen stehen. Die Wurzel wird an die Stelle gesetzt, an der Außengruppe und Innengruppe miteinander verbunden sind (Boykin *et al.*, 2010). Eine weitere Methode, die für das Wurzeln der Stammbäume verwendet werden kann, ist die Minimale Ursprungsabweichungsmethode (engl. *minimal ancestor deviation*, MAD). Bei dieser Methode wird für jeden Ast die Abweichung zu einer möglichen Wurzel berechnet und die Wurzel an den Ast platziert, welcher die geringste Abweichung aufweist (Tria *et al.*, 2017).

Es gibt zwei Arten von Stammbäumen. Zum einen die ultrametrischen Stammbäume, wozu die ungewichtete Paargruppen Methode mit arithmetischen Mitteln (engl. *unweighted pair group method with aritmetic mean*, UPGMA) zählt, sowie die additiven Stammbäume. Zu der zweiten Gruppe von Stammbäumen zählen die Stammbäume, die auf Distanzen, Wahrscheinlichkeiten und Charakteren basieren. Eine häufig genutzte Charakter basierte Methode ist die maximale-Parsimonie-Methode (engl. *maximum parsimony*, MP). Die bekannteste Wahrscheinlichkeiten basierte Methode ist die maximale-Wahrscheinlichkeitsmethode (eng. *maximum likelihood*, ML) und die geläufigste Distanzbasierte-Methode ist die Nachbarschafts-Verbindungs-Methode (engl. *neighbor-joining*, NJ).

Bei den ultrametrischen Bäumen wird davon ausgegangen, dass der Stammbaum eine molekulare Uhr (engl. *molecular clock*, Zuckerkandl und Pauling (1962)) darstellt. Durch diese Annahme wird allen Knoten ein Alter zugewiesen. Die ultrametrischen Bäume werden immer gewurzelt dargestellt (De Soete, 1984). Die OTUs haben das Alter von 0 und die Wurzel ist dementsprechend am ältesten. Ultrametrische Stammbäume erfüllen immer die drei Punkt Bedingung (engl. *three point condition*, Bruneman (1971)). Demnach wird angenommen, dass für die Einträge x, y, und z einer ultrametrischen Distanzmatrix, von den Distanzen d(x,y), d(x,z) und d(y,z), zwei Distanzen den maximal Werten entsprechen (Wheeler, 2012):

$$d(x, y) < d(x, z) = d(y, z) \tag{6}$$

Bei der UPGMA-Methode wird anhand einer hierarchischen Clustermethode der Stammbaum erstellt. Die Methode basiert wie die NJ-Methode auf Distanzen. In einen Stammbaum, der durch die UPGMA-Methode erstellt wurde, werden alle Distanzen gleichermaßen in den Mittelwert mit einbezogen. Es wird keine Gewichtung vorgenommen, dementsprechend wird der Stammbaum als Dendrogramm dargestellt und enthält keine echten Astlängen (De Soete, 1984).

Bei den additiven Stammbäumen werden die Äste gewichtet. Das bedeutet, dass die Äste unterschiedlich lang sind und der Distanz zwischen den Taxa entsprechen (Wheeler, 2012). Diese Art der Stammbäume wird anhand der vier Punkte Bedingung berechnet (Bruneman, 1971). Wird davon ausgegangen, dass sich in einem Baum die Taxa, x, y, u und v befinden, wird die Distanz (d) zwischen den Taxa wie folgt bestimmt, um einem additiven Stammbaum zu entsprechen:

$$d(x, y) + d(u, v) \leq d(x, u) + d(y, v) = d(x, v) + d(y, u) \tag{7}$$

Durch diese Gleichung kann auch gezeigt werden, dass jeder additive Stammbaum in einen ultrametrischen Stammbaum umgewandelt werden kann, allerdings kann nicht jeder ultrametrische Stammbaum in einen additiven Stammbaum umgewandelt werden (Farris Transformation, Farris *et al.* (1970)).

Bei der MP-Methode wird der Stammbaum als richtig erachtet, welcher die kleinste Anzahl an Charakter-Austauschen (Mutationen) in den gegebenen Sequenzen beschreibt (Felsenstein, 1996). Durch die hohe evolutionäre Variabilität, kann es zu einer Verfälschung der Astlängen (engl. „*long branch attraction*“, Felsenstein (1978)) kommen, da multiple Mutationen auftreten können und diese durch die MP Methode nicht korrigiert werden (Gabaldón, 2005). Des Weiteren stellt die Topologie der MP Stammbäume die geringste Distanz dar (Takahashi und Nei, 2000). Da diese Methode allerdings nicht alle Informationen der Originaldaten im Stammbaum darstellen kann (Penny, 1982), ist die Methode für evolutionäre Stammbäume eher widersprüchlich (Felsenstein, 1981).

Mit der ML Methode werden Stammbäume berechnet, welche die höchste *logarithmische (log)* Wahrscheinlichkeit aufweisen (Takahashi und Nei, 2000). Dazu wird für jeden möglichen Stammbaum die *log* Wahrscheinlichkeit, basierend auf unterschiedlichen Parametern (Topologie, Astlänge) berechnet (Takahashi und Nei, 2000). Ein Vorteil der ML Methode ist, dass keine Konstanz der Sequenzen vorausgesetzt wird (Felsenstein, 1981). Dadurch kann die ML Methode den anderen Stammbaumrekonstruktions-Methoden übergeordnet werden (Hasegawa *et al.*, 1991). Sogar wenn die evolutionäre Varianz zwischen den Sequenzen eines Stammbaumes voneinander abweicht, kann die ML Methode einen korrekten Stammbaum berechnen (Kishino *et al.*, 1990). Durch die Berechnung aller möglichen

Topologien ist der Zeitaufwand für große Datenmengen höher als bei anderen Methoden (Gabaldón, 2005). Aufgrund der Tatsache, dass durch die ML Methode die wahrscheinlichsten Stammbäume berechnet werden können, werden die Programme für die Berechnung stetig weiterentwickelt und die Laufzeit deutlich minimiert (Stamatakis, 2014).

Auch die NJ Methode ist eine weit verbreitete Methode, um Stammbäume zu berechnen. Es ist ein sehr schneller Algorithmus, benötigt allerdings eine sehr akkurate Distanzmatrix, um einen korrekten Stammbaum zu erstellen (Hasegawa und Fujiwara, 1993). Als Nachbar wird ein OTU Paar berechnet, welches nur durch einen HTU in einem ungewurzelten Baum verbunden ist. Das Ziel dieser Methode ist es, einen Stammbaum zu finden, welcher die geringste Astlänge insgesamt hat (Saitou und Nei, 1987).

4 Zielsetzung

Die heutigen Prokaryoten, Archaeen und Bakterien, stammen von einem gemeinsamen Vorfahren (engl.: *last universal common ancestor*, Luca) ab. Die Rekonstruktion des Genoms könnte Informationen über den Ursprungsort sowie die Lebensweise von Luca liefern. Durch das Fehlen von fossilen Spuren müssen andere Methoden angewendet, beziehungsweise bereits bestehende Methoden adaptiert werden, um das Genom zu rekonstruieren, da es nicht wie bei modernen Organismen sequenziert werden kann.

Eine Möglichkeit Lucas genetische Informationen zu gewinnen, ist die Verwendung von prokaryotischen Genomen. Anhand von Sequenzähnlichkeiten und Proteinfamilien, sowie Stammbaumrekonstruktionen können mögliche Gene identifiziert werden, die zu Lucas Genom gehören.

Prokaryoten neigen dazu, ihre Gene nicht nur vertikal, sondern auch horizontal/lateral, durch Konjugation, Transformation und Transduktion, sogar über die Artengrenze weiterzugeben. Durch den lateralen Gentransfer (LGT) (Kannan *et al.*, 2013), ist es notwendig, Filterschritte zu entwickeln, durch die nur noch ursprüngliche Gene von Luca identifiziert werden können.

In dieser Arbeit soll mit Hilfe von phylogenomischen Analysen anhand von prokaryotischen Genomen das Genom von Luca entschlüsselt werden. Dazu wurden Filterschritte entwickelt, die den Einfluss von LGT sowie die überrepräsentierten Prokaryoten Gruppen auf die Analyse verringerten. Des Weiteren ist es möglich, die ursprünglichsten Stoffwechselwege und Prokaryoten, also Prokaryoten die eine ähnliche Genzusammensetzung aufweisen wie LUCA, zu detektieren.

Um die Ziele dieser Arbeit zu erreichen, sollen die zu untersuchenden prokaryotischen Proteinsequenzen und die daraus resultierenden Proteinfamilien aligniert und phylogenomische Stammbäume erstellt werden. Durch die Bestimmung der Distanzen innerhalb der Stammbäume können die ursprünglichsten Prokaryoten bestimmt werden. Anhand der Proteinfamilien, die auf Luca deuten, kann ermittelt werden, welche Stoffwechselwege am ursprünglichsten sind.

5 Publikationen

5.1 The physiology and habitat of the last universal common ancestor

Madeline C. Weiss, Filipa L. Sousa, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi und William F. Martin

Affiliations

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

Dieser Artikel wurde am 25 Juli 2016 in *Nature Microbiology* Ausgabe 1 veröffentlicht.

Beitrag von Madeline C. Weiß:

Ich habe die Analyse zu den bereits erstellten Proteinfamilien durchgeführt. Dazu gehörte die Aufteilung der Proteinfamilien in drei Kategorien, ausschließlich bakterielle Sequenzen, ausschließlich archaeelle Sequenzen und archaeelle und bakterielle Sequenzen gemeinsam vorhanden. Für die weitere Analyse wurden nur die Proteinfamilien mit archaeellen und bakteriellen Sequenzen verwendet und Sequenzalignments sowie Maximum Likelihood Bäume berechnet. Des Weiteren wurden von mir die Filter für die Identifikation von möglichen Proteinen, welche zum letzten gemeinsamen Vorfahren gehören, ausgearbeitet und auf den Datensatz angewendet. Weiterhin habe ich die Matrix, welche in Abbildung 2 verwendet wurde, erstellt und am Text mitgearbeitet. Neben Literaturrecherchen und Korrekturlesungen habe ich den Methodentext mitgeschrieben.

nature microbiology

ARTICLES

PUBLISHED: 25 JULY 2016 | ARTICLE NUMBER: 16116 | DOI: 10.1038/NMICROBIOL.2016.116

The physiology and habitat of the last universal common ancestor

Madeline C. Weiss†, Filipa L. Sousa†, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi and William F. Martin*

The concept of a last universal common ancestor of all cells (LUCA, or the progenote) is central to the study of early evolution and life's origin, yet information about how and where LUCA lived is lacking. We investigated all clusters and phylogenetic trees for 6.1 million protein coding genes from sequenced prokaryotic genomes in order to reconstruct the microbial ecology of LUCA. Among 286,514 protein clusters, we identified 355 protein families (~0.1%) that trace to LUCA by phylogenetic criteria. Because these proteins are not universally distributed, they can shed light on LUCA's physiology. Their functions, properties and prosthetic groups depict LUCA as anaerobic, CO_2-fixing, H_2-dependent with a Wood–Ljungdahl pathway, N_2-fixing and thermophilic. LUCA's biochemistry was replete with FeS clusters and radical reaction mechanisms. Its cofactors reveal dependence upon transition metals, flavins, *S*-adenosyl methionine, coenzyme A, ferredoxin, molybdopterin, corrins and selenium. Its genetic code required nucleoside modifications and *S*-adenosyl methionine-dependent methylations. The 355 phylogenies identify clostridia and methanogens, whose modern lifestyles resemble that of LUCA, as basal among their respective domains. LUCA inhabited a geochemically active environment rich in H_2, CO_2 and iron. The data support the theory of an autotrophic origin of life involving the Wood–Ljungdahl pathway in a hydrothermal setting.

The last universal common ancestor (LUCA) is an inferred evolutionary intermediate[1] that links the abiotic phase of Earth's history with the first traces of microbial life in rocks that are 3.8–3.5 billion years of age[2]. Although LUCA was long considered the common ancestor of bacteria, archaea and eukaryotes[3,4], newer two-domain trees of life have eukaryotes arising from prokaryotes[5,6], making LUCA the common ancestor of bacteria and archaea. Previous genomic investigations of LUCA's gene content have focused on genes that are universally present across genomes[4,7,8], revealing that LUCA had 30–100 proteins for ribosomes and translation. In principle, genes present in one archaeon and one bacterium might trace to LUCA, although their phylogenetic distribution could also be the result of post-LUCA gene origin and interdomain lateral gene transfer (LGT)[8], given that thousands of such gene transfers between prokaryotic domains have been detected[9].

To identify genes that can illuminate the biology of LUCA, we took a phylogenetic approach. Among proteins encoded in sequenced prokaryotic genomes, we sought those that fulfil two simple criteria: (1) the protein should be present in at least two higher taxa of bacteria and archaea, respectively, and (2) its tree should recover bacterial and archaeal monophyly (Fig. 1). Genes meeting both criteria are unlikely to have undergone transdomain LGT, and thus were probably present in LUCA and inherited within domains since the time of LUCA. By focusing on phylogeny rather than universal gene presence, we can identify genes involved in LUCA's physiology—the ways that cells access carbon, energy and nutrients from the environment for growth.

Results

Tracing proteins to LUCA by removing transdomain LGTs. Using the standard Markov cluster algorithm (MCL) at a 25% global identity threshold, we sorted all 6,103,411 protein coding genes in 1,847 bacterial and 134 archaeal genomes into 286,514 protein families, or clusters (see Methods), 11,093 of which contained homologues from bacteria and archaea. After alignment and maximum likelihood (ML) tree construction, only 355 clusters preserve domain monophyly while also having homologues in ≥2 archaeal lineages and ≥2 bacterial lineages (see Methods). Encouragingly, 83% (294/355) of LUCA's genes have some functional annotation (Supplementary Tables 1 and 2), with only a minority belonging to translation.

These 355 proteins were probably present in LUCA and thus provide a glimpse of LUCA's genome. Their distribution across prokaryotic higher taxa is presented in Fig. 2 and Supplementary Fig. 1. Clearly, the list of these 355 genes comes with caveats, such as lineage sampling, sequence conservation and the possibility that multiple LGTs might mimic intradomain vertical inheritance. However, there are also quality benchmarks against which to check the list. For example, LUCA's genes encode 19 proteins involved in ribosome biogenesis and eight aminoacyl tRNA synthetases, which are also essential for the genetic code to work (Supplementary Table 2). Thus, our phylogenetic criteria do not miss the informational core, which itself can be affected by LGTs, such that only subsets of even universally present genes will also meet the domain monophyly criterion. As another benchmark, our phylogenetic criteria return a highly non-random sample of genes. The distribution of functional categories represented among the 355 genes tracing to LUCA is significantly different ($P << 1 \times 10^{-16}$) from that represented in the 11,093 cluster sample (Supplementary Table 3; see Methods), with oxygen sensitive enzymes (Supplementary Table 2) and FeS proteins (Supplementary Table 1) overrepresented in LUCA's list.

LUCA's microbial ecology reconstructed from genomes. Reconstructed from genomic data, LUCA emerges as an anaerobic

Institute of Molecular Evolution, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany. †These authors contributed equally to this work. *e-mail: bill@hhu.de

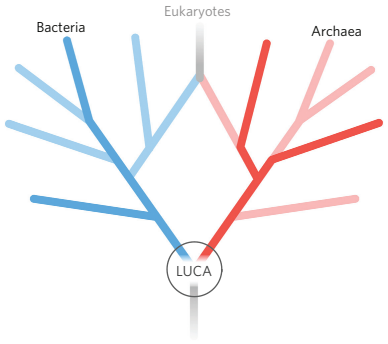


Figure 1 | Phylogeny for LUCA's genes. In the two-domain tree of life[5,6], eukaryotes stem from prokaryotes, so the last universal common ancestor, LUCA, is the ancestor of archaea and bacteria. The tree shows a schematic phylogeny of phyla for a gene present in two archaeal and two bacterial phyla and in which both prokaryotic domains are monophyletic. By applying the criteria—(1) the gene should be present in at least two members each of two bacterial phyla and two archaeal phyla (see Methods) and (2) the protein tree should recover monophyly of bacteria and archaea—355 clusters were identified that trace to LUCA.

autotroph[10] that used a Wood–Ljungdahl (WL) pathway[11] and existed in a hydrothermal setting[12,13], but that was only half-alive and was dependent upon geochemistry, as summarized in Fig. 3. LUCA's genes harbour traces of carbon, energy and nitrogen metabolism. Cells conserve energy via chemiosmotic coupling[14] with rotor–stator-type ATP synthases or via substrate-level phosphorylation (SLP)[15]. LUCA's genes encompass components of two enzymes of energy metabolism: phosphotransacetylase (PTA) and an ATP synthase subunit (Supplementary Table 2). PTA generates acetylphosphate from acetyl-CoA, conserving the energy in the thioester bond as the energy-rich anhydride bond of acetylphosphate, which can phosphorylate ADP or other substrates[15]. The PTA reaction plays a central role in autotrophic theories of microbial origins that focus on thioester-dependent SLP as the ancestral state of microbial energy metabolism[16,17]. The presence of a rotor–stator ATP synthase subunit points to LUCA's ability to harness ion gradients for energy metabolism[17], yet the rotor–stator ATP synthase has undergone transdomain LGT[18], excluding many of its subunits from LUCA's set. Crucially, components of electron-transfer-dependent ion-pumping are altogether lacking among LUCA's genes. LUCA's ATPase was possibly able to harness geochemically derived ion gradients[17] via H^+/Na^+ antiporters[19], which are present among the membrane proteins in the list (Supplementary Table 2). The presence of reverse gyrase, an enzyme specific for hyperthermophiles[20], indicates a thermophilic lifestyle for LUCA.

Enzymes of chemoorganoheterotrophy are lacking, but enzymes for chemolithoautotrophy are present. Among the six known pathways of CO_2 fixation[11], only enzymes of the WL pathway are present in LUCA (Supplementary Table 4). LUCA's WL enzymes are replete with FeS and FeNiS centres[21], indicating transition-metal requirements and also requiring organic cofactors: flavin, F_{420}, methanofuran, two pterins (the molybdenum cofactor MoCo and tetrahydromethanopterin) and corrins (Supplementary Table 4 and Supplementary Fig. 2). Microbes that use the WL pathway obtain their electrons from hydrogen[11], hydrogenases also being present among LUCA's genes. LUCA accessed nitrogen via nitrogenase and via glutamine synthetase. The WL pathway, nitrogenase and hydrogenases are also very oxygen-sensitive. LUCA was an anaerobic autotroph that could live from the gases H_2, CO_2 and N_2.

Several cofactor biosynthesis pathways trace to LUCA, including those for pterins, MoCo, cobalamin, siroheme, thiamine pyrophosphate, coenzyme M and F_{420} (Supplementary Table 5). Many of these enzymes are *S*-adenosyl methionine (SAM)-dependent. A number of LUCA's SAM-dependent enzymes are radical SAM enzymes (Supplementary Table 1), an ancient class of oxygen-sensitive proteins harbouring FeS centres that initiate radical-dependent methylations and a wide spectrum of radical reaction mechanisms[22]. Radical SAM reactions point to a prevalence of one-electron reactions in LUCA's central metabolism, as does an abundance of flavoproteins (Supplementary Table 1), in addition to a prominent role for methyl groups.

FeS clusters, long viewed as relics of ancient metabolism[23,24], are the second most common cofactor/prosthetic group in LUCA's proteins behind ATP (Supplementary Table 1). The abundance of transition metals and FeS as well as FeNiS clusters in LUCA's enzymes indicates that it inhabited an environment rich in these metals. These features of LUCA's environment, in addition to thermophily and H_2, clearly point to a hydrothermal setting[12,13,17]. Selenoproteins, required in glutathione and thioredoxin synthesis and for some of LUCA's RNA modifications, are present, as is selenophosphate synthase (Supplementary Table 2). Like FeS centres, selenium in amino acids and nucleosides is thought to be an ancient trait[25]. Sulfur was involved in ancient metabolism[26] and LUCA was capable of S utilization, as indicated by siroheme, which is specific to redox reactions involving environmental S. Enzymes for sugar metabolism mainly encompass glycosylases, hydrolases and nonoxidative sugar metabolism, possibly reflecting primitive cell wall synthesis.

LUCA's genes point to acetogenic and methanogenic roots. The 355 trees also harbour phylogenetic information about LUCA's descendants, because archaea and bacteria are reciprocally rooted. Clostridia were the most frequently basal-branching bacteria, while methanogens were the most frequently basal-branching archaea (Supplementary Table 6). Clostridia and methanogens use the WL pathway[11]; they are abundant among microbial communities that inhabit the Earth's crust today[27,28], they harbour species that can live from methyl groups[6,27,28] and—like LUCA (Fig. 3)—they depend on H_2.

Today, environmental H_2 has two main sources: geological processes and H_2-producing fermentations. When LUCA existed, biological H_2 production did not exist, because primordial organics delivered from space are non-fermentable substrates[29]. For LUCA, that leaves only geological sources of H_2. The main geological source of H_2, both today and on the early Earth, is serpentinization, a process in which Fe^{2+} in the crust reduces water circulating through hydrothermal systems to produce H_2 at high activities in hydrothermal effluent[30] of up to 26 mmol kg^{-1}. Yet, as well as H_2, methane[31,32] and other reduced C1 compounds[30,33] are synthesized abiotically in hydrothermal systems today.

Hydrothermal vents, methyl groups, and nucleoside modifications. LUCA's genes for RNA nucleoside modification (Supplementary Table 4) indicate that it performed chemical modification of nucleosides in both tRNA and rRNA[34]. Four of LUCA's nucleoside modifications are methylations requiring SAM (Supplementary Table 4). In the modern code, several base modifications are even strictly required for codon–anticodon interactions at the wobble position[35] (Supplementary Fig. 3). Consistent with the recurrent role of methyl groups in LUCA's biology, by far the most common tRNA and rRNA nucleoside modifications that are conserved across the archaeal bacterial divide[36] are methylations (Fig. 4a), although thiomethylations and incorporation of sulfur and selenium are observed.

That LUCA's genetic code involved modified bases in tRNA–mRNA–rRNA interactions attributes antiquity and functional

NATURE MICROBIOLOGY DOI: 10.1038/NMICROBIOL.2016.116 ARTICLES

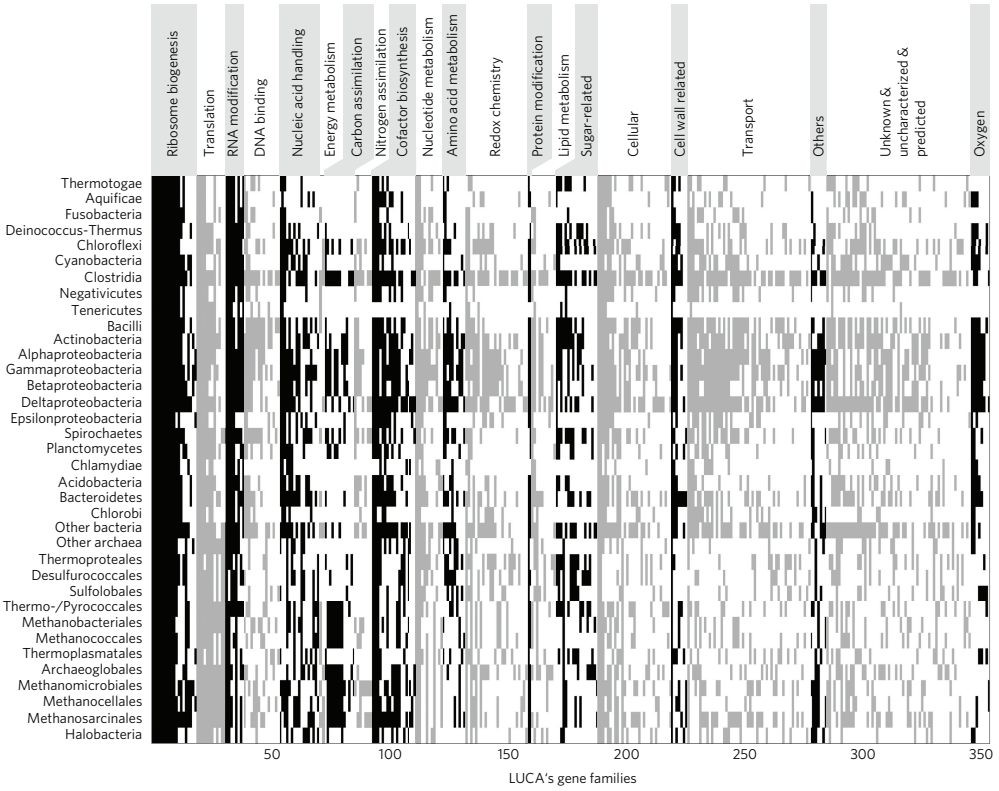


Figure 2 | Taxonomic distribution of LUCA's genes grouped by functional categories. The 355 clusters that fulfil the two criteria in Fig. 1 are grouped into 21 different functional categories (top). For each protein family (columns), ticks indicate the presence (black or grey) or absence (white) of the gene in the corresponding group shown on the left (rows). Within each functional category, clusters were sorted by the total sum of genes present (*x* axis). The distribution of the functional categories represented in the 355 LUCA gene families is significantly different from that represented in the data set as a whole (Supplementary Table 3).

significance to methylated bases in the evolution of the ribosome and the genetic code. It also forges links between the genetic code (Fig. 4a), primitive carbon and energy metabolism (Fig. 4b,c) and hydrothermal environments. How so? At modern hydrothermal vents, reduced C1 intermediates are formed during the abiotic synthesis of methane[30,33]. The intermediates can accumulate in some modern hydrothermal systems[33] and the underlying reactions can be simulated in the laboratory[37–39]. These reactions occur because under the reducing conditions of hydrothermal vents, the equilibrium in the reaction of H_2 with CO_2 lies on the side of reduced carbon compounds[40,41]. That is the reason why methanogens and clostridial acetogens, both of which figure prominently in autotrophic theories for life's origin[17,19], can grow by harnessing energy from the exergonic reactions of hydrogenotrophic methane and acetate synthesis[14,15]. The genes in LUCA's list (Supplementary Table 2) and the basal lineages among the 355 reciprocally rooted trees (Supplementary Table 6) indicate that LUCA lived in an environment where the geochemical synthesis of methane from H_2 and CO_2 was taking place, hence where chemically accessible reduced C1 intermediates existed.

Where LUCA arose, the genetic code arose. Either the chemical modifications of RNA nucleosides were absent in LUCA and were introduced later in evolution by some kind of adaptation, as some have suggested[42], or they are ancient[43]. Conservation of nucleoside modifications across the archaeal bacterial divide (Fig. 4a) indicate the latter. The enzymes that introduce nucleoside methylations are typically SAM enzymes, including members of the radical SAM family, which harbour an FeS cluster that initiates a radical in the reaction mechanism[22] and which are currently thought to be among the most ancient enzymes in metabolism[22]. Both the presence in LUCA's genome (Supplementary Table 4) of several SAM enzymes involved in nucleoside modifications and the presence of the nucleosides themselves (Fig. 4a), sometimes even at conserved positions in tRNA (Supplementary Fig. 3), indicate that these nucleoside methylations were present in LUCA's code, reflecting the code's ancestral state.

In methanogens and acetogenic clostridia, which the trees identified as the closest relatives of LUCA (Supplementary Table 6), methyl groups are central to growth, comprising the very core of carbon and energy metabolism. As shown in Fig. 4b, the methyl group generated by the WL pathway in the energy metabolism of methanogens is transferred from a nitrogen atom in tetrahydromethanopterin to a Co(I) atom in a corrin cofactor of the methyltransferase (MtrA-H) complex, possibly bound by MtrE[44], then subsequently transferred to the thiol sulfur atom of coenzyme M before being transferred to hydride at the methyl–CoM reductase

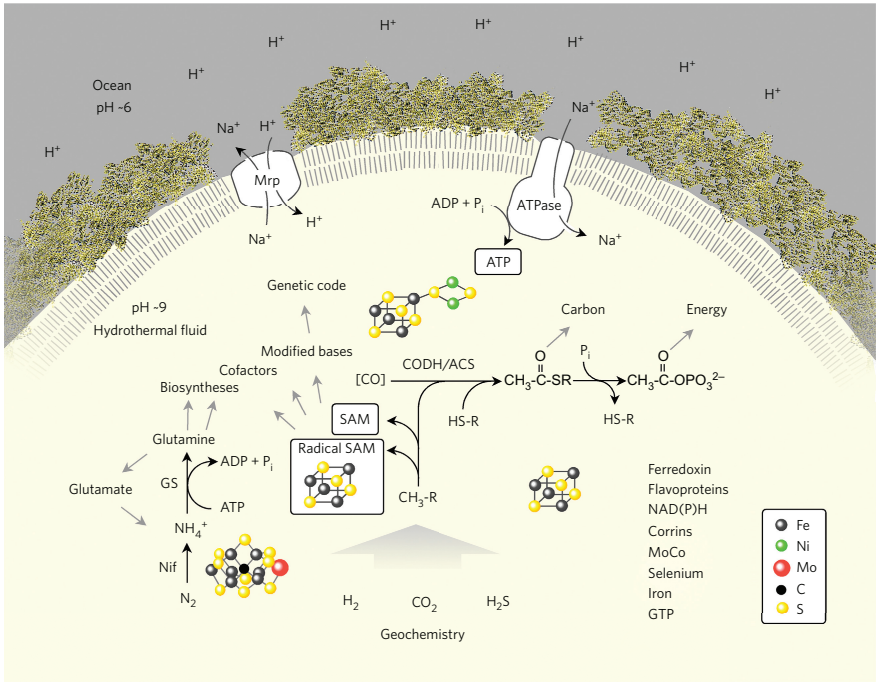


Figure 3 | LUCA reconstructed from genome data. Summary of the main interactions of LUCA with its environment, a vent-like geochemical setting[12,13,17,19], as inferred from genome data (Supplementary Table 2). Abbreviations: CODH/ACS, carbon monoxide dehydrogenase/acetyl CoA-synthase; Nif, nitrogenase; GS, glutamine synthetase; Mrp, MrP type Na^+/H^+ antiporter; CH_3-R, methyl groups; HS-R, organic thiols. The components listed on the lower right are present in LUCA, in addition to the cofactors listed in Supplementary Table 1. In modern CODH/ACS complexes, CO is generated from CO_2 and reduced ferredoxin[21]. The figure does not make a statement regarding the source of CO in primordial metabolism (uncatalysed via the gas water shift reaction or catalysed via transition metals), symbolized by [CO]. A Na^+/H^+ antiporter could transduce a geochemical pH gradient (indicated on the left) inherent in alkaline hydrothermal vents[13,17] into a more stable Na^+ gradient to feed a primordial Na-dependent ATP synthase[19]. LUCA undisputably possessed genes, because it had a genetic code; the question of which genes it possessed has hitherto been more difficult to address. The transition metal catalysts at the nitrogenase active site and the CODH/ACS active site as well as a 4Fe–4S cluster as in ferredoxin are indicated.

reaction. Energy conservation via Na^+ pumping occurs during the N-to-Co(I)-to-S transfer sequences of the MtrA-H complex[14,44]. This is an unusual coupling reaction in that electrons are not transferred; a methyl group is instead transferred, from a N atom to a S atom[14,44]. As shown in Fig. 4c, the methyl transfer chain of acetogens is a bit longer. It starts with a nitrogen-bound methyl moiety in tetrahydrofolate, which is transferred by AscE to a Co(I) atom in a corrin cofactor of the corrinoid FeS protein[45] and onto a Ni atom in the FeNiS cluster of acetyl-CoA synthase in an unusual metal-to-metal methyl transferase reaction[45]. Carbonyl insertion[21] generates a Ni-bound acetyl group that is removed via thiolysis to generate the thioester, which can either be used for carbon assimilation or for energy conservation as acyl phosphate and ATP[11,15]. In methanogens, carbon metabolism follows the same path to the thioester[11,14].

These methyl transfer reactions suggest that the environment where primordial carbon and energy metabolism arose was rich in methyl groups, S and transition metals. The conserved SAM-dependent methylations and S substitutions in modified nucleosides that allow tRNA anticodons to decode mRNA into protein carry the same chemical imprint (Fig. 4a and Supplementary Fig. 3), uncovering a hitherto underappreciated antiquity and significance of methyl groups at the core of biological chemistry. Methyl groups provide previously unrecognized links between carbon and energy metabolism in anaerobic autotrophs (Fig. 4b,c), tRNA–mRNA–rRNA interactions in the genetic code (Fig. 4a and Supplementary Fig. 3) and spontaneous chemistry at hydrothermal vents[30–33].

Spelling out caveats and allowing for some LGT. No approach to the study of early evolution is consummate; there are always caveats. Using our strict phylogenetic criterion, 355 protein families that are present in at least two higher taxa per domain and that preserve interdomain monophyly were identified. Universally distributed genes can be subject to transdomain LGT, yielding false negatives and underestimates of LUCA's gene content, while multiple LGT events might mimic vertical inheritance for some clusters, yielding false positives, or overestimates of LUCA's gene content. As an example of the latter, O_2-dependent enzymes should generally be absent from the list, because in LUCA's day, O_2 did not exist in physiologically relevant amounts[2]. LUCA's list does, however, contain five enzymes that use O_2 as a substrate and three that detoxify O_2 (Supplementary Table 2), functions that cannot be germane to LUCA, hence resulting from multiple transfers that phylogenetically emulate vertical intradomain inheritance. Given the massive influence that O_2 had on the origin and spread of new genes during evolution[46], finding O_2-dependent reactions at a frequency of 2.3% (8/355 proteins) suggests that the list of 355 genes harbours comparatively few multiple transfer cases. At the same time, ecological specialization to oxic niches will induce

NATURE MICROBIOLOGY DOI: 10.1038/NMICROBIOL.2016.116 ARTICLES

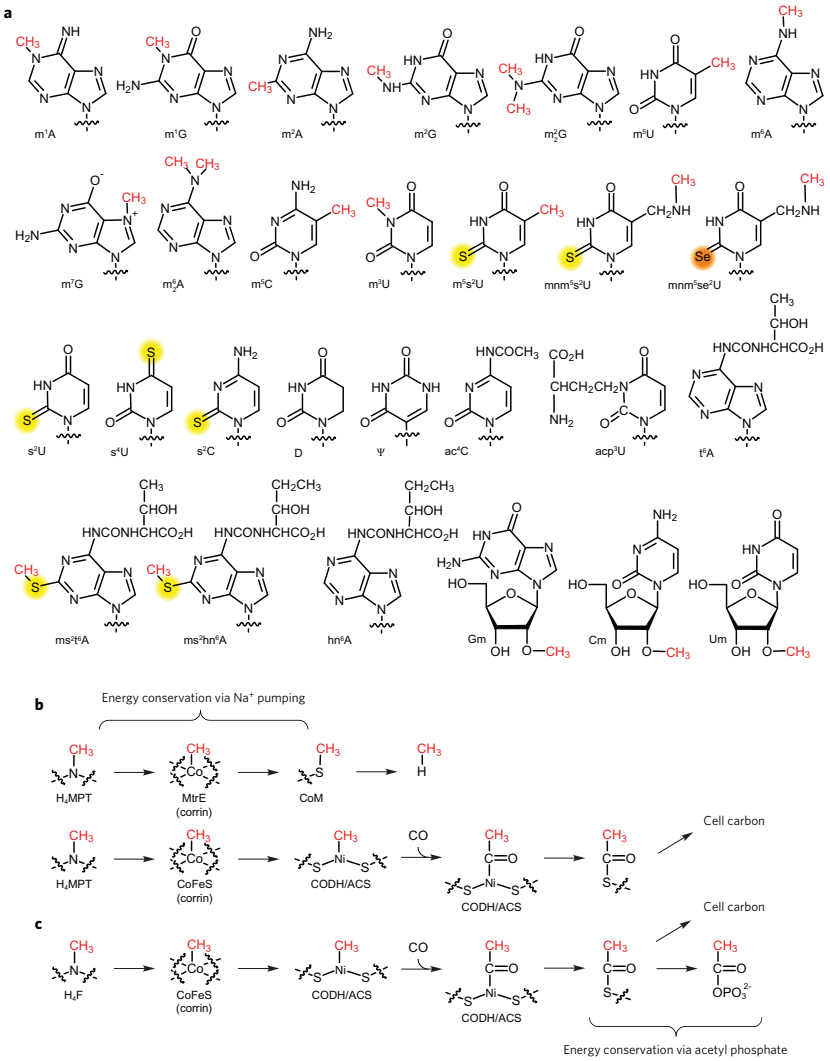


Figure 4 | Methyl groups in conserved modified nucleosides and in anaerobic autotroph metabolism. a, Structures of modified nucleosides found in tRNA and rRNA of archaea and bacteria[36]. Sulfur and selenium are highlighted in yellow and orange, respectively, while methyl groups are indicated in red. Many of these methylations are performed by SAM-dependent enzymes, which are mainly involved in nucleoside modifications and cofactor biosyntheses (Supplementary Table 7), suggesting a very central role of SAM in early metabolism and at the origin of tRNA–mRNA–rRNA interactions. **b**, In hydrogenotrophic methanogens, energy conservation occurs via Na^+ pumping during the N-to-Co(I)-to-S methyl transfer at the MtrA-H methyltransferase complex[14,44], while carbon assimilation[6] involves methyl transfer from H_4MPT to the corrinoid iron-sulfur protein (CoFeS)[45] and CODH/ACS[21], which reduces CO_2 to CO and catalyses the synthesis of acetyl-CoA. **c**, In hydrogenotrophic acetogens[15], the methyl group is transferred to CoFeS and CODH/ACS, where carbonyl insertion[21] and thiolysis to generate acetyl-CoA for carbon assimilation or for energy conservation[11] occur. Abbreviations: m^1A, 1-methyladenosine; m^2A, 2-methyladenosine; m^6A, N^6-methyladenosine; t^6A, N^6-threonylcarbamoyladenosine; ms^2t^6A, 2-methylthio-N^6-threonylcarbamoyladenosine; m^5C, 5-methylcytidine; Cm, 2′-O-methylcytidine; ac^4C, N^4-acetylcytidine; m^1G, 1-methylguanosine; m^2G, N^2-methylguanosine; m^7G, 7-methylguanosine; Gm, 2′-O-methylguanosine; m^2_2G, N^2,N^2-dimethylguanosine; **ψ**, pseudouridine; D, dihydrouridine; m^5U, ribosylthymine; Um, 2′-O-methyluridine; s^2U, 2-thiouridine; s^4U, 4-thiouridine; m^5s^2U, 5-methyl-2-thiouridine; acp^3U, 3-(3-amino-3-carboxypropyl)uridine; mnm^5s^2U, 5-methylaminomethyl-2-thiouridine; mnm^5se^2U, 5-methylaminomethyl-2-selenouridine; s^2C, 2-thiocytidine; hn^6A, N^6-hydroxynorvalylcarbamoyladenosine; ms^2hn^6A, 2-methylthio-N^6-hydroxynorvalyl carbamoyladenosine; m^3U, 3-methyluridine; m^6_2A, N^6,N^6-dimethyladenosine; H_4MPT, tetrahydromethanopterin; CoM, coenzyme M; CoFeS, corrinoid iron-sulfur protein; CODH/ACS, carbon monoxide dehydrogenase/acetyl-CoA synthase; H_4F, tetrahydrofolate; MtrE, a component of the MtrA-H methytransferase complex[44].

massive loss of anaerobe-specific genes in many lineages, leading to an underestimation of LUCA's gene content.

In addition, LUCA's gene list reveals only nine nucleotide biosynthesis and five amino acid biosynthesis proteins (Supplementary Table 2). The paucity of enzymes for essential amino acid, nucleoside and cofactor biosyntheses is most easily attributed to three factors: (1) the missing genes in question have been subject to interdomain LGT; (2) the genes are not well conserved at the sequence level, such that the bacterial and archaeal homologues do not fall into the same cluster; (3) LUCA had not yet evolved the genes in question prior to the bacterial–archaeal split, the pathway products for LUCA being provided by primordial geochemistry instead.

Low sequence conservation for proteins that were present in LUCA can also yield underestimates of LUCA's gene content. Because our criteria for presence in LUCA (domain monophyly) involve trees, the sequences need to be sufficiently well conserved to permit multiple sequence alignments and ML phylogenies, so a clustering threshold of 25% global identity was used. Yet genes that were present in LUCA that were not subject to transdomain LGT, but are not well conserved, can still fall into separate domain specific clusters. Relative to archaea, bacteria are overrepresented in the 355 families by a ratio of 134:1,847 in terms of genome sequences, for example in the far right of Supplementary Fig. 1. Despite our phylogenetic criteria, some LGTs might be among them. Enzymes for lipid metabolism in LUCA are scarce. The presence of a few enzymes involved in acyl-CoA metabolism might reflect multiple LGTs, as several archaea have acquired bacterial genes for fatty acid and aliphatic degradations[47].

Finally, one might ask what happens if we allow for a little bit of transdomain LGT? The minimum amount of transdomain LGT for which to allow would be reflected by a tree that fulfils our criteria of being present in two members each in two archaeal and two bacterial phyla (see Methods), but in addition, one bacterial sequence is misplaced within the archaea (or vice versa). If we allow for such cases representing one single transdomain LGT, then 124 new trees would be included (Supplementary Table 8), expanding our list to 479 members. If we allow for the next increment of LGT, namely that not one sequence but sequences from one archaeal phylum are misplaced within the bacteria (or vice versa), then 97 additional trees would be included (Supplementary Table 9), bringing the list to 576 proteins. The functional annotations in those expanded lists are very much in line with those reflected in the list of 355 summarized in Supplementary Tables 1 and 2.

Discussion

Our findings clearly support the views that FeS and transition metals are relics of ancient metabolism[23,24], that life arose at hydrothermal vents[12,13], that spontaneous chemistry in the Earth's crust driven by rock–water interactions at disequilibrium thermodynamically underpinned life's origin[41,48] and that the founding lineages of the archaea and bacteria were H_2-dependent autotrophs that used CO_2 as their terminal acceptor in energy metabolism[17,19]. Spontaneous reactions involving C1 compounds and intermediate methyl groups in modern submarine hydrothermal systems[30–33] link observable geochemical processes with the earliest forms of carbon and energy metabolism in bacteria and archaea. In the same way that biochemists have long viewed FeS clusters as relics of ancient catalysis[23,24], methyl groups appear here as relics of primordial carbon and energy metabolism.

Although the paucity in LUCA of genes for amino acid and nucleoside biosyntheses could, in principle, be attributable to post-LUCA LGT, we note that there is no viable alternative to the view that LUCA, regardless of how envisaged, ultimately arose from components that were synthesized abiotically via spontaneous, exergonic syntheses somewhere during the history of early Earth[48]. Prior to the origin of genes, proteins and the code, LUCA's origin was hence dependent on spontaneous organic syntheses, which are thermodynamically favourable under the high H_2 activities of submarine hydrothermal vents[40,41], and which still occur today in some geochemical environments[30–33]. The notion that early replicating systems tapped environmental supplies of biologically relevant compounds provided by spontaneous (exergonic) chemical reactions might seem to be a very radical proposition in the present Article, but it is inherent, often implicitly, to all theories for the prebiotic origin of replicating systems and life.

Genome data depict LUCA as a strictly anaerobic, H_2-dependent thermophilic, diazotrophic autotroph with a WL pathway and that lived in a hydrothermal vent setting. These are attributes of acetogenic clostridia and methanogens, lineages that branch deeply in trees of LUCA's genes and that occupy the Earth's crust today. Methyl groups were central to carbon and energy metabolism in LUCA, and they also persist to the present as chemical relicts in tRNA–mRNA interactions at the ribosome, suggesting that LUCA not only lived in a hydrothermal vent setting rich in H_2, CO_2, transition metals, sulfur and reactive C1 species of geochemical origin, but that LUCA's genetic code arose there as well. The data provide evidence in favour of autotrophic origins[11] over heterotrophic origins[49] and converge with independent geochemical evidence[30–33], favouring theories that posit a single hydrothermal environment rich in H_2 and transition metals for LUCA's origin[11–13,16,17,19,48] over theories that entail many different kinds of chemical environments[50] catalysing one reaction each.

Methods

Sequence clustering and gene phyletic pattern reconstruction. Protein sequences of 1,981 complete prokaryotic genomes were downloaded from the NCBI RefSeq[51] database (version June 2012). These genomes were grouped into 13 archaeal and 23 bacterial groups corresponding to NCBI taxonomic orders, phylum or class, respectively, as in ref. 9. Markov chain clustering[52] (MCL) of sequences was performed as previously described[9], with the reciprocal best BLAST[53] (v. 2.2.28) hit (rBBH) procedure and an *E*-value threshold of $\leq 10^{-10}$. rBBH-pairs with global amino acid identity not smaller than 25%, calculated with needle[54] from EMBOSS 6.6.0.0, were clustered using MCL. Proteins with no significant homologues were classified as singletons and discarded from further analysis. Protein families with archaeal and bacterial sequences were considered further.

Multiple sequence alignment and reconstruction of phylogenetic trees. Sequences in each of the MCL clusters were aligned using MAFFT[55] version 7.130 with the options *–localpair*, *–maxiterate = 1000* and *–anysymbol*. The heads-or-tails method[56,57] was used to compare alignment reliability for different sets of clusters with an inbuilt program. ML trees were reconstructed using RAxML[58] v7.8.6 under the PROTCATWAG model, with special amino acid characters U, O and J converted into C, K and X. These trees were rerooted between archaea and bacteria and parsed for monophyly of sets of sequences using the programs *nw_reroot* and *nw_clade* with the *-m* option from Newick Utilities[59] version 1.6. Single group paraphyly was identified as those trees where a single archaeal or bacterial group was placed within the other domain.

LUCA gene identification. Phyletic patterns for 1,981 genomes were analysed and a gene family was only considered as present in LUCA if there were at least two archaeal groups and two bacterial groups, each of which had a minimum of two members, respectively, present in the given family and if archaea and bacteria were monophyletic in the corresponding phylogenetic tree. Exceptions were made for the four under-represented groups ('Methanocellales', 'Thermoplasmatales', 'Archaeoglobales' and 'other archaea'), where the presence of only one sequence instead of a minimum of two was counted as present.

Functional annotation and cofactor determination. The protein families were annotated using COG[60] and KEGG[61] functional categories. COG identifiers (COG IDs) of all sequences from 355 candidate LUCA clusters were extracted from the NCBI RefSeq database (June 2012). A particular COG ID was assigned to a cluster if it was assigned to more sequences in that cluster than any other COG ID. Each COG ID was then mapped to the COG functional categories. If a COG ID mapped to more than one category, the category R (general function prediction only) was assigned. The same procedure was repeated to annotate families using the KEGG database. In addition, for each protein family, the most frequent PFAM-A domain annotation (version 28.0, June 2015) was obtained by using the HMM approach as available at PFAM[62]. The identification of protein cofactors and catalytic centres present in the families as well as their organization into categories was performed through case by

case inspection of each of the 355 families. The functional categories listed in Supplementary Tables 1, 2, 4 and 5 are designed to reflect microbial physiology (for example, the category redox), so they do not correspond 1:1 to COG or KEGG functional categories; however, the COG and KEGG categories for each protein are given in Supplementary Table 2. For a test of independence (see section 'Test of independence'), COG categories were used. Cofactors occurring only in one protein family are not shown in Supplementary Table 1. These include thiamine pyrophosphate, pyridoxal phosphate, biotin, methyltetrahydrofolate, menaquinone, iron and rubredoxin. Copper (two occurrences) is also not shown because it is only present as a metal centre in oxygen-related protein families. For protein families corresponding to subunits of a protein complex, a subunit was scored as presence of the complex, such that cofactors of the complex were counted even if not all subunits were present in our list. For complexes scored as present, cofactors were counted only once. Ferredoxin, flavodoxin and methanophenazine were counted as ferredoxin because proteins annotated as coenzyme F_{420}-reducing hydrogenase can correspond to protein families that bind either ferredoxin or methanophenazine, while the electron carrier for NifH can be either ferredoxin or flavodoxin. Thioredoxin was not counted as a cofactor.

Presence of selenoproteins. All amino acid sequences present in the 355 clusters were searched for the presence of the one letter code amino acid 'U' that represents selenocysteine. Selenium was considered present in a protein family when selenocysteine was found in at least one protein sequence of the cluster.

Deeply branching archaeal and bacterial lineages. Deeply branching groups in each domain were identified based on two different factors: (1) the smallest split that contains both archaea and bacteria, or (2) sequences connecting both archaea and bacteria with shortest evolutionary distance in the tree. The smallest split that contains both archaea and bacteria was obtained by parsing the bipartitions of the tree, and two different cases were considered: (1) in the $Tree_{Pure}$ method, intradomain basal branches containing sequences from only one phylum/group were counted; (2) in the $Tree_{Mixed}$ method, the intradomain basal branches containing sequences belonging to organisms from more than one phylum/group were counted. To identify sequences connecting archaea and bacteria at the shortest evolutionary distance, trees were translated into a distance matrix of branch lengths using the *nw_distance* program from Newick Utilities[59] (version 1.6). The sequence pair with shortest distance connecting archaea and bacteria was identified from this distance matrix ($Dist_{Root}$ method).

Ribosome, tRNA and nucleoside modifications. Sequences of prokaryotic tRNA containing modifications were downloaded from the Modomics[63] and tRNAdb[64] databases (December 2014). Bacterial and archaeal tRNA modifications were mapped into the tRNA structure and common modifications occurring at the same tRNA position in both prokaryotic domains were identified. Supplementary Fig. 3a was prepared using VMD[65] version 1.9.2. All chemical structures were drawn in ACD/ChemSketch (2015 release, Advanced Chemistry Development; www.acdlabs.com).

Test of independence. To determine whether functional annotations and taxonomical groups were distributed non-randomly in the 355 candidate LUCA clusters, a χ^2 test of independence was performed. All 11,093 clusters that contained homologues from both archaea and bacteria were used to calculate the expected distribution. COG categories were separated into five groups: information, metabolism, cellular, poorly characterized and not declared (including clusters that were not annotated by the COG database). There were 36 taxonomical groups used for the lineage distribution: 13 archaeal groups (Archaeoglobales, Desulfurococcales, Halobacteria, Methanobacteriales, Methanocellales, Methanococcales, Methanomicrobiales, Methanosarcinales, Sulfolobales, Thermococcales/Pyrococcales, Thermoplasmatales, Thermoproteales, other archaea) and 23 bacterial groups (Acidobacteria, Actinobacteria, Alphaproteobacteria, Aquificae, Bacilli, Bacteroidetes, Betaproteobacteria, Chlamydiae, Chlorobi, Chloroflexi, Clostridia, Cyanobacteria, Deinococcus-Thermus, Deltaproteobacteria, Epsilonproteobacteria, Fusobacteria, Gammaproteobacteria, Negativicutes, Planctomycetes, Spirochaetes, Tenericutes, Thermotogae, other bacteria). The distribution of functional annotations and taxonomical groups of the 355 candidate LUCA clusters were compared with those expected (degrees of freedom = 4 and degrees of freedom = 35, respectively). *P* values were calculated using MATLAB R2015a and its function chi2cdf.

Received 19 April 2016; accepted 21 June 2016; published 25 July 2016

References

1. Fox, G. E. *et al.* The phylogeny of prokaryotes. *Science* **209,** 457–463 (1980).
2. Arndt, N. & Nisbet, E. Processes on the young Earth and the habitats of early life. *Annu. Rev. Earth Planet Sci.* **40,** 521–549 (2012).
3. Woese, C. The universal ancestor. *Proc. Natl Acad. Sci. USA* **95,** 6854–6859 (1998).
4. Koonin, E. V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Rev. Microbiol.* **1,** 127–136 (2003).
5. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504,** 231–236 (2013).
6. Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl Acad. Sci. USA* **112,** 6670–6675 (2015).
7. Ouzounis, C. A., Kunin, V., Darzentas, N. & Goldovsky, L. A minimal estimate for the gene content of the last universal common ancestor—exobiology from a terrestrial perspective. *Res. Microbiol.* **157,** 57–68 (2006).
8. Kannan, L., Li, H., Rubinstein, B. & Mushegian, A. Models of gene gain and gene loss for probabilistic reconstruction of gene content in the last universal common ancestor of life. *Biol. Direct.* **8,** 32 (2013).
9. Nelson-Sathi, S. *et al.* Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517,** 77–80 (2015).
10. Say, R. F. & Fuchs, G. Fructose 1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. *Nature* **464,** 1077–1081 (2010).
11. Fuchs, G. Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu. Rev. Microbiol.* **65,** 631–658 (2011).
12. Baross, J. A. & Hoffman, S. E. Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Origins Life Evol. B* **15,** 327–345 (1985).
13. Russell, M. J. & Hall, A. J. The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. *J. Geol. Soc. Lond.* **154,** 377–402 (1997).
14. Buckel, W. & Thauer, R. K. Energy conservation via electron bifurcating ferredoxin reduction and proton/Na^+ translocating ferredoxin oxidation. *Biochim. Biophys. Acta* **1827,** 94–113 (2013).
15. Schuchmann, K. & Müller, V. Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria. *Nature Rev. Microbiol.* **12,** 809–821 (2014).
16. Ferry, J. G. & House, C. H. The step-wise evolution of early life driven by energy conservation. *Mol. Biol. Evol.* **23,** 1286–1292 (2006).
17. Martin, W. & Russell, M. J. On the origin of biochemistry at an alkaline hydrothermal vent. *Phil. Trans. R. Soc. Lond. B* **362,** 1887–1925 (2007).
18. Mulkidjanian, A. Y., Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Evolutionary primacy of sodium bioenergetics. *Biol. Direct.* **3,** 13 (2008).
19. Lane, N. & Martin, W. F. The origin of membrane bioenergetics. *Cell* **151,** 1406–1416 (2012).
20. Déclais, A. C., Marsault, J., Confalonieri, F., La Tour de, C. B. & Duguet, M. Reverse gyrase, the two domains intimately cooperate to promote positive supercoiling. *J. Biol. Chem.* **275,** 19498–19504 (2000).
21. Ragsdale, S. W. Nickel-based enzyme systems. *J. Biol. Chem.* **284,** 18571–18575 (2009).
22. Broderick, J. B., Duffus, B. R., Duschene, K. S. & Shepard, E. M. Radical *S*-adenosylmethionine enzymes. *Chem. Rev.* **114,** 4229–4317 (2014).
23. Eck, R. V. & Dayhoff, M. O. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science* **152,** 363–366 (1966).
24. Hall, D. O., Cammack, R. & Rao, K. K. Role of ferredoxins in the origin of life and biological evolution. *Nature* **233,** 136–138 (1971).
25. Böck, A., Forchhammer, K., Heider, J. & Baron, C. Selenoprotein synthesis: an expansion of the genetic code. *Trends Biochem. Sci.* **16,** 463–467 (1991).
26. Liu, Y. C., Beer, L. L. & Whitman, W. B. Methanogens: a window into ancient sulfur metabolism. *Trends Microbiol.* **20,** 251–258 (2012).
27. Evans, P. N. *et al.* Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350,** 434–438 (2015).
28. Lever, M. A. Acetogenesis in the energy-starved deep biosphere—a paradox? *Front. Microbiol.* **2,** 284 (2012).
29. Schönheit, P., Buckel, W. & Martin, W. F. On the origin of heterotrophy. *Trends Microbiol.* **24,** 12–25 (2016).
30. Schrenk, M. O., Brazelton, W. J. & Lang, S. Q. Serpentinization, carbon, and deep life. *Rev. Mineral. Geochem.* **75,** 575–606 (2013).
31. Etiope, G. & Schoell, M. Abiotic gas: atypical, but not rare. *Elements* **10,** 291–296 (2014).
32. Proskurowski, G. *et al.* Abiogenic hydrocarbon production at Lost City hydrothermal field. *Science* **319,** 604–607 (2008).
33. McDermott, J. M., Seewald, J. S., German, C. R. & Sylva, S. P. Pathways for abiotic organic synthesis at submarine hydrothermal fields. *Proc. Natl Acad. Sci. USA* **112,** 7668–7672 (2015).
34. Chow, C. S., Lamichhane, T. N. & Mahto, S. K. Expanding the nucleotide repertoire of the ribosome with post-transcriptional modifications. *ACS Chem. Biol.* **2,** 610–619 (2007).
35. Agris, P. F., Vendeix, F. A. P. & Graham, W. D. tRNA's wobble decoding of the genome: 40 years of modification. *J. Mol. Biol.* **366,** 1–13 (2007).
36. Grosjean, H., Gupta, R., & Maxwell, E. S. in *Archaea: New Models for Prokaryotic Biology* (ed. Blum, P.) 171–196 (Caister Academic Press, 2008).

37. Seewald, J. S., Tolotov, M. Y. & McCollom, T. Experimental investigation of single carbon compounds under hydrothermal conditions. *Geochim. Cosmochim. Acta* **70,** 446–460 (2006).
38. He, C., Tian, G., Liu, Z. & Feng, S. A mild hydrothermal route to fix carbon dioxide to simple carboxylic acids. *Org. Lett.* **12,** 649–651 (2010).
39. Horita, J. & Berndt, M. Abiogenic methane formation and isotopic fractionation under hydrothermal conditions. *Science* **285,** 1055–1057 (1999).
40. Amend, J. P. & Shock, E. L. Energetics of amino acid synthesis in hydrothermal ecosystems. *Science* **281,** 1659–1662 (1998).
41. Amend, J. P., LaRowe, D. E., McCollom, T. M. & Shock, E. L. The energetics of organic synthesis inside and outside the cell. *Phil. Trans. R. Soc. Lond. B* **368,** 20120255 (2013).
42. Yokoyama, S., Watanabe, K. & Miyazawa, T. Dynamic structures and functions of transfer ribonucleic acids from extreme thermophiles. *Adv. Biophys.* **23,** 115–147 (1987).
43. Helm, M. Post-transcriptional nucleotide modification and alternative folding of RNA. *Nucleic Acids Res.* **34,** 721–733 (2006).
44. Gottschalk, G. & Thauer, R. K. The Na^+-translocating methyltransferase complex from methanogenic archaea. *Biochim. Biophys. Acta* **1505,** 28–36 (2001).
45. Svetlitchnaia, T., Svetlitchnyi, V., Meyer, O. & Dobbek, H. Structural insights into methyltransfer reactions of a corrinoid iron–sulfur protein involved in acetyl-CoA synthesis. *Proc. Natl Acad. Sci. USA* **103,** 14331–14336 (2006).
46. Raymond, J. & Segre, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311,** 1764–1767 (2006).
47. Dibrova, D. V., Galperin, M. Y. & Mulkidjanian, A. Y. Phylogenomic reconstruction of archaeal fatty acid metabolism. *Environ. Microbiol.* **16,** 907–918 (2014).
48. Shock, E. L. & Boyd, E. S. Geomicrobiology and microbial geochemistry: principles of geobiochemistry. *Elements* **11,** 389–394 (2015).
49. Mansy, S. S. *et al.* Template-directed synthesis of a genetic polymer in a model protocell. *Nature* **454,** 122–125 (2008).
50. Patel, B. H., Percivalle, C., Ritson, D. J., Duffy, C. D. & Sutherland, J. D. Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nature Chem.* **7,** 301–307 (2015).
51. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40,** D130–D135 (2011).
52. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An ancient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
53. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
54. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology open software suite. *Trends Genet.* **16,** 276–277 (2000).
55. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
56. Landan, G. & Graur, D. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.* **24,** 1380–1383 (2007).
57. Landan, G. & Graur, D. Local reliability measures from sets of co-optimal multiple sequence alignments. *Pac. Symp. Biocomput.* **13,** 15–24 (2008).
58. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
59. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26,** 1669–1670 (2010).
60. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28,** 33–36 (2000).
61. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27,** 29–34 (1999).
62. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–D285 (2016).
63. Machnicka, M. A. *et al.* MODOMICS: a database of RNA modification pathways—2013 update. *Nucleic Acids Res.* **41,** D262–D267 (2013).
64. Jühling, F. *et al.* tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.* **37,** D159–D162 (2009).
65. Humphrey, W., Dalke, A. & Schulten, K. VMD—visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).

Acknowledgements
The authors thank J. Baross and N. Lane for discussions. The authors acknowledge the Zentrum für Informations- und Medientechnologie (ZIM) of the Heinrich-Heine University for computational support and the European Research Council for funding (ERC AdG 666053 to W.F.M.).

Author contributions
M.C.W., F.L.S., S.N., M.R. and S.N.-S. performed the bioinformatics analysis. F.L.S. and N.M. carried out the functional classification of the protein families. All authors analysed and discussed the results. W.F.M., F.L.S. and S.N.-S. designed the research. M.C.W., F.L.S., S.N., N.M., S.N.-S. and W.F.M. wrote the paper.

Additional information
Supplementary information is available online. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to W.F.M.

Competing interests
The authors declare no competing financial interests.

5.2 Reply to ‘Is LUCA a thermophilic progenote‘

Madeline C. Weiss, Sinje Neukirchen, Mayo Roettger, Natalia Mrnjavac, Shijulal Nelson-Sathi, William F. Martin und Filipa L. Sousa

Affiliations

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

Dieser Artikel wurde am 25 November 2016 in *Nature Microbiology* Ausgabe 1 veröffentlicht.

Beitrag von Madeline C. Weiß:

Die Analyse, auf die eingegangen wird, habe ich für die Publikation Weiss *et al.* (2016b) durchgeführt. Des Weiteren war ich an der Erstellung des Textes durch konstruktive Kritik und fachbezogene Diskussionen beteiligt. Weiterhin habe ich die Literaturrecherche durchgeführt und den Text formatiert.

PUBLISHED: 25 NOVEMBER 2016 | ARTICLE NUMBER: 16230 | DOI: 10.1038/NMICROBIOL.2016.230

correspondence

Reply to 'Is LUCA a thermophilic progenote?'

Weiss *et al.* reply — In response to our recent paper[1], Gogarten and Deamer[2] write in with five paragraphs. They focus on traditional views concerning the nature of the last universal common ancestor (LUCA). We find the current exchange worthwhile in that it highlights several important differences in older and newer concepts concerning both LUCA and approaches to inference of its properties.

Their first paragraph, which summarizes some, but by no means all, virtues of submarine hydrothermal vents in the context of life's origin, requires no response, although John Baross[3], Mike Russell[4] and Everett Shock[5] can explain far better than we can how warmly the idea that life arose at hydrothermal vents was "welcomed"[2].

Gogarten and Deamer's second paragraph revisits older inferences about LUCA and the progenote that are based on the three-domain tree[6]. However, improved phylogenetic methods and better archaeal lineage sampling now deliver a different picture of domain relationships, called the two-domain tree, in which the archaeal partner at eukaryote origin arises from within the archaea, not as the sister of the archaea[7,8]. We can hardly be faulted that newer methods and data obtain the two-domain tree. Why is the issue of the three-domain verus two-domain tree important? It is this. Eukaryotes are derived from a single common ancestor that had both mitochondria and a very narrow sample of bacterial carbon and energy metabolism, demonstrating facultatively anaerobic chemoorganoheterotrophy[9]. Inferences about LUCA that trace eukaryotic properties to the first cells (the three-domain tree) always exclude most forms of microbial physiology — for example, nitrification, mineral oxidations, the knallgas reaction, sulfate reduction, methanogenesis, and so forth[5] — and can never recover traits such as anaerobic chemolithoautotrophy or the Wood–Ljungdahl pathway in LUCA for lack of such physiology in eukaryotes. Investigations of LUCA based on a three-domain approach to the problem can therefore not recover sets of genes similar to the ones we found using phylogenetic criteria that embrace the newer two-domain tree, which has been germane to our formulations of hydrothermal origins right from the beginning[10]. Gogarten and Deamer also lament in this paragraph that our results lead to an inference of LUCA that is not as complex as a free-living cell[2], or "half-alive", as we put it. We will return to this point in closing.

Their third paragraph deals with potential caveats of our method to address LUCA's properties. In our paper[1], under the subheading "Spelling out caveats and allowing for some LGT", we explained the issues concerning possible false positives and possible false negatives, but in greater depth and detail than Gogarten and Deamer[2]. We also mentioned examples[1]. Their comment just restates points that we made first.

The fourth paragraph complains that we do not obtain the same results as earlier studies that addressed the temperature at which LUCA lived, studies that were based upon the three-domain tree. Yes, that is correct. Our results really do differ from those in previous studies. In addition, the molecular thermometer papers they cite, which estimate growth temperatures from inferred GC content in selected regions of some genes, as modelled along the branches of the three-domain tree, would need to be repeated on the basis of more current archaeal lineage sampling along the branches of the two-domain tree[7,8].

In their fifth paragraph, Gogarten and Deamer criticize various aspects of the theory that life arose at hydrothermal vents, in particular the idea that the naturally chemiosmotic nature of alkaline hydrothermal vents could have been important for life's origin. The aspects of the hydrothermal vent theory that they criticize have been developed and discussed in many earlier papers by us and others[3,4,9–12]. In our recent genomic investigation study, we do not modify the theory, we merely find independent, genome-based evidence compatible with it. Hence, their critique applies to older papers and not ours[1], the findings of which are compatible with numerous aspects of some versions of the hydrothermal vent theory (the roles of metals, acetogenesis, methanogenesis, methyl groups, clostridia, methanogens, the acetyl-CoA pathway[11]), including the view that gradients[3] and chemiosmosis[4,9–12] were important at the origin of life. Can we be faulted that genomes deliver that result? Gogarten and Deamer[2] are particularly concerned that our study did not recover complete lipid biosynthesis, but we reported a number of membrane proteins within our dataset[1], indicating the presence of lipids in LUCA. Not all of the membrane proteins are drawn in Fig. 3, which summarizes physiology, but they are reported in Fig. 2 and elsewhere in the paper[1]. Yet, not only is lipid synthesis poorly represented, the majority of amino acid and nucleotide biosyntheses are missing too. As we wrote[1], lack of such essential functions among LUCA's gene set could indicate (1) that the missing genes unspectacularly underwent transdomain lateral gene transfer (LGT) post-LUCA, and hence were filtered out by our method, (2) that missing chemical components were provided by spontaneous abiotic syntheses during early Earth history, or (3) a combination thereof. LGT between the prokaryotic domains is both normal and natural[13] and all theories for the origin of cells, without exception, require abiotic syntheses; hence, we do not see any fundamental problems here. Their examples of universally present genes that are lacking from LUCA's list are not criticisms of our findings, rather they merely underscore what we wrote, namely that genes present in all genomes "can be affected by LGTs, such that only subsets of even universally present genes will also meet the domain monophyly criterion"[1].

Finally, Gogarten and Deamer[2] protest that our inference of LUCA contained "ribosomes, translation, genes, and a genetic code", offering that such a level of organization goes "far beyond what most would imagine as the first form of life"[2]. This is a very important comment. If the first form that Gogarten and Deamer consider to be alive lacks ribosomes, translation, genes and the genetic code, how can they possibly be worried that our version of LUCA, which has genes, ribosomes, the code, translation, exergonic carbon and energy metabolism, nitrogen fixation and many basics of cofactor biosyntheses, is only half-alive? Clearly, our version of early life[1] comes much closer to something that one might find in microbial culture collections (in

freezers with acetogens and methanogens), than their version[2], which would not be deposited among archaea, bacteria, eukaryotes or viruses, but amidst organisms that lack ribosomes, the code, translation, and genes, and for lack of genes would have no heritable metabolism for either carbon, energy or nitrogen.

As the genetic code and some ribosomal proteins are universal to all cells, it seems inescapable that LUCA and the first forms of anything that we would call alive possessed genes. Our study is based on genes. It is clearly not possible to test the concept of gene-lacking life forms[2] using genomic data. In closing, they ask for suggestions for how to test the idea that life arose at hydrothermal vents. Laboratory[12,14] and field[15] investigations have already been reported and, looking at the matter openly, our contribution is a test of the hypothesis as it relates to genomes[1]. In summary, Gogarten and Deamer[2] summarize their views concerning LUCA and we have welcomed this opportunity to summarize ours. ❐

References

1. Weiss, M. C. *et al. Nat. Microbiol.* **1,** 16116 (2016).
2. Gogarten, J. P. & Deamer, D. *Nat. Microbiol.* **1,** 16229 (2016).
3. Baross, J. A. & Hoffman, S. E. *Orig. Life Evol. Biosph.* **15,** 327–345 (1985).
4. Russell, M. J. & Hall, A. J. *J. Geol. Soc.* **154,** 377–402 (1997).
5. Amend, J. P. & Shock, E. L. *FEMS Microbiol. Rev.* **25,** 175–243 (2001).
6. Woese, C. R., Kandler, O. & Wheelis, M. L. *Proc. Natl Acad. Sci. USA* **87,** 4576–4579 (1990).
7. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. *Nature* **504,** 231–236 (2013).
8. Müller, M. *et al. Microbiol. Mol. Biol. Rev.* **76,** 444–495 (2012).
9. Martin, W. & Russell, M. J. *Philos. Trans. R. Soc. Lond. B* **362,** 1887–1925 (2007).
10. Lane, N. & Martin, W. F. *Cell* **151,** 1406–1416 (2012).
11. Sousa, F. L. *et al. Philos. Trans. R. Soc. Lond. B* **368,** 20130088 (2013).
12. Sojo, V., Pomiankowski, A. & Lane, N. *PLoS Biol.* **12,** e1001926 (2014).
13. Nelson-Sathi, S. *et al. Nature* **517,** 77–80 (2015).
14. Roldan, A. *et al. Chem. Commun.* **51,** 7501–7504 (2015).
15. McDermott, J. M., Seewald, J. S., German, C. R. & Sylva, S. P. *Proc. Natl Acad. Sci. USA* **112,** 7668–7672 (2015).

Madeline C. Weiss, Sinje Neukirchen†, Mayo Roettger, Natalia Mrnjavac, Shijulal Nelson-Sathi†, William F. Martin and Filipa L. Sousa†

Institute for Molecular Evolution, Heinrich-Heine University Düsseldorf, Universitätsstraβe 1, 40225 Düsseldorf, Germany. †Present address: Department of Ecogenomics and Systems Biology, University of Vienna, 1090 Vienna, Austria (S.N., F.L.S.); Computational Biology & Bioinformatics Group, Rajiv Gandhi Centre for Biotechnology (RGCB), Trivandrum, Kerala 695014, India (S.N.-S.).
e-mail: bill@hhu.de

5.3 Physiology, phylogeny and LUCA

William F. Martin[1,2], Madeline C. Weiss[1], Sinje Neukirchen[3], Shijulal Nelson-Sathi[4], Filipa L. Sousa[3]

Affiliations

1 Institut für Molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Universitätsstrasse 1, 40225 Düsseldorf, Deutschland.
2 Instituto de Tecnologia QuõÂmica e Biológica, Universidade Nova de Lisboa, 2780-157 Oeiras, Portugal.
3 Department of Ecogenomics and Sytems Biology, University of Vienna, Althanstrasse 14, 1090 Wien, Österreich.
4 Computational Biology and Bioinformatics Group, Rajiv Gandhi Centre of Biotechnology (RGCB), Trivandrum, Kerala 695014, Indien.

Dieser Artikel wurde am 25 November 2016 in *microbial cell* Ausgabe 3 veröffentlicht.

Beitrag von Madeline C. Weiß:

Die in der Publikation erwähnte Analyse stammt aus der Publikation Weiss *et al.* (2016b) und wurde von mir durchgeführt. Weiterhin habe ich die Litertaurrecherche durchgeführt und den Text in das richtige Format für das Journal gebracht. Des Weiteren habe ich den Text mit überarbeitet.

microbial cell

Viewpoint

www.microbialcell.com

Physiology, phylogeny, and LUCA

William F. Martin*[,1,2], Madeline C. Weiss[1], Sinje Neukirchen[3], Shijulal Nelson-Sathi[4], Filipa L. Sousa[3]

[1] Institute for Molecular Evolution, Heinrich-Heine Universität Düsseldorf, Universitätstrasse 1, 40225 Düsseldorf, Germany.
[2] Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, 2780-157 Oeiras, Portugal.
[3] Department of Ecogenomics and Systems Biology, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria.
[4] Computational Biology & Bioinformatics Group, Rajiv Gandhi Centre for Biotechnology (RGCB), Trivandrum, Kerala 695014, India.
* Corresponding Author:
William F. Martin, Molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Building: 26.13, Floor/Room: 01.34, Universitätsstraße 1; 40225 Duesseldorf, Germany; Tel: +49 211 81 13011; Fax: +49 211 81 13554; E-mail: bill@hhu.de

ABSTRACT **Genomes record their own history. But if we want to look all the way back to life's beginnings some 4 billion years ago, the record of microbial evolution that is preserved in prokaryotic genomes is not easy to read. Microbiology has a lot in common with geology in that regard. Geologists know that plate tectonics and erosion have erased much of the geological record, with ancient rocks being truly rare. The same is true of microbes. Lateral gene transfer (LGT) and sequence divergence have erased much of the evolutionary record that was once written in genomes, and it is not obvious which genes among sequenced genomes are genuinely ancient. Which genes trace to the last universal ancestor, LUCA? The classical approach has been to look for genes that are universally distributed. Another approach is to make all trees for all genes, and sift out the trees where signals have been overwritten by LGT. What is left ought to be ancient. If we do that, what do we find?**

doi: *10.15698/mic2016.12.545*
Received *originally: 21.09.2016; in revised form: 27.09.2016,*
Accepted *28.09.2016.*
Published *25.11.2016.*

Keywords*: early evolution, autotrophy, geochemistry, acetogens, methanogens.*

Early evolution and the nature of the very first kinds of life are interesting topics. They concern the phase of Earth history where our most distant ancestors emerged from the elements on an otherwise lifeless planet. The questions of how the initial evolutionary transition — from inanimate to animate matter — might have happened and what the first kinds of life were like in terms of habitat and lifestyle are just plain interesting. People generally want to know about how things were in the past, including the most distant past. It is apparently part of human nature to wonder where we came from.

An important concept in very early evolution is the last universal common ancestor, LUCA for short, because it represents the organism, cell, thing, or chemical reaction, depending on one's concept of LUCA, from which all life forms we know are descended. Thoughts about the nature of LUCA abound in the literature and are immensely diverse; the search term 'last universal common ancestor' alone returns 188 articles since 1997 in standard literature databases. Diversity of thoughts on LUCA is partly due to the circumstance that when we, as scientists, conceptually delve as deep as LUCA in evolutionary history, we are not far removed from the topic of life's origin. Thoughts on the origin of life are even more diverse than on LUCA, with over 2200 articles in literature databases appearing with 'origin of life' as the query. How can one learn more about the biology of LUCA, the starting point of early evolution?

If we look around, there are presently only two ways to empirically approach early evolution: geology and genomes. A prominent geologist, Andy Knoll, likes to say "Earth records its own history" [1], which is spot-on. Geology can indeed tell us when life arose. The oldest sedimentary rocks, which are ca. 3.8 billion years of age, harbour traces for life in the form of light carbon isotopes, evidence for biological CO_2 fixation at that time [2,3]. But the presence of CO_2 fixation, possibly even as far back as 4.1 Ga [4] does not tell us everything that we might want to know about early life. Indeed, plate tectonics and erosion have erased much of the Earth's recorded history, with truly ancient rocks being rare and their evidence for early life often being difficult to interpret. Nonetheless, the geochemical record does harbor evidence for physiological processes.

A problem arises, though, in that physiological processes among prokaryotes are not generally restricted to any particular phylogenetic group. A glaring exception to that rule are the cyanobacteria, who also infringe upon the rule that Earth records its own history, because since cyanobacteria have been around, they have been editing a lot of Earth's recorded text with their waste product, oxygen [5]. Outside of the cyanobacteria, phylogeny and physiology are decoupled by the reality of lateral gene transfer (LGT)

among prokaryotes: sulfate reduction [6], anoxygenic photosynthesis [5], fermentations [7], and respirations [8] are distributed among many different prokaryotic lineages, but because of LGT, not because of differential loss: LUCA could not do everything, it can hardly have possessed a genome of Eden. One might interject that methanogenesis *is* restricted to a particular phylogenetic group, the methanogens, but new phylogenetic depictions of the 'tree of life' have methanogens basal among the archaea, with loss of methanogenesis in many independent groups [9,10], those losses corresponding to gene acquisitions from bacteria in some cases [11], thereby decoupling phylogeny from physiology in the methanogens, too, which no longer appear as a monophyletic group.

Curiously, genomes also record their own history. But lateral gene transfer (much like plate tectonics) and sequence divergence (much like erosion) have erased much of the evolutionary signal that the very first genomes on our planet contained. Nonetheless we can be sure that there was a time and a place and an environment where those very first genomes did exist. How can one harness genomes to find out more about what the first life forms were like, and how to get a better picture of LUCA?

In the modern era (since the discovery of archaea), the ribosomal RNA tree of life, or the three domain tree [12], has been the main starting point for inferences about the nature of LUCA. But as progress has accrued with genomes, three issues have come to the fore that bear on inferences of LUCA's gene set: i) the effects of lateral gene transfer on our picture of LUCA, ii) the question of whether the three domain tree is correct, and iii) the issue of how universally distributed genes need to be in order to trace to LUCA.

The LGT issue is fairly straightforward. One avenue of investigation into LUCA has been to see which, what kind of and how many genes are common to archaea, bacteria and eukaryotes (all three domains). All things being equal, and barring LGT, such genes would trace to LUCA. So by simply looking for gene presence, Ouzounis *et al.* [13] could attribute about 1000 genes to LUCA, if LUCA was taken as the common ancestor of prokaryotes, or up to 1400 genes, if eukaryotes were included and if one allowed for widespread gene loss and excluded LGT. But like earlier investigations [14] and later investigations [15], Ouzounins *et al*. [13] attributed all absences of genes among lineages descended from LUCA to differential loss. If genes were distributed across domains by LGT, rather than differential loss, then presence of a gene in all three domains (or in both prokaryotic domains) would not reflect presence in LUCA, it would just reflect transdomain LGT. If not identified and removed, LGT generates overestimates of LUCA's gene content. Kannan *et al.* [16] very clearly spelled out the problem that transdomain LGT introduces into the study of LUCA's genes, and they also explained why it is not trivial to circumvent the LGT problem. The real problem with transdomain LGT is not that it has been known for many years to be an issue in early evolution [17], rather the real issue is its prevalence in nature today and in the past. Phylogenetic studies spanning all genes from many hundreds of genomes uncover thousands of cases of transdomain LGT, mainly from bacteria to archaea [11,18]. If such LGT cases are identified and filtered out, maybe a picture of LUCA will come into focus.

The influence of the three domain tree on the issue of LUCA is somewhat more complicated. Many investigators on the issue of LUCA have adhered strictly to the three domain tree, meaning that if one wants to address LUCA, one must first place a root somewhere on the three domain tree. Investigations of anciently duplicated genes [19,20] led to placement of the root on the bacterial branch [12]. But even among proponents of the three domain tree, the bacterial root was not universally accepted. For example, there have been strong proponents of the view that, the three domain tree is correct, but its root should be on the eukaryotic branch, coupled with the view that LUCA was more similar to eukaryotes than it was to prokaroytes [21-24] — a line of inference that has led its proponents to argue that the term 'prokaryote' be banned from the literature altogether. Di Guilio [25] also argues that we should ban the use of the term prokaryotes, albeit on grounds that do not hinge upon arguments that the first cells were eukaryote-like. Such discussions result in suggestions for terms like acaryotes, akaryotes, arkarya, and syncaryote [26] to replace the very useful concepts of prokaryotes and eukaryotes, terms which the more physiologically minded among us [27] are (wisely, we think) unwilling to surrender.

While debates about LUCA and higher order microbial nomenclature have been brewing, something else far more threatening for the three domain tree has been gnawing on its trunk: the three domain tree apparently has the domain relationships wrong. Recently, a small revolution in deep phylogenetic views has occurred, with newer methods of phylogenetic inference and investigations based on broader sampling of archaeal lineages having brought forth a new view of domain relationships, in which the archaeal component of eukaryotes branches *within* the archaea, not *as a sister* to them [9, 28-32]. Jim Lake will be quick to point out that some people had been saying that for 30 years [33]. Defenders of the three domain tree counter that there is no need to worry, the three domain tree will persist [34]. But people keep on finding the new tree of domain relationships, which is currently being called the two domain tree [29]. Lake [33] (1988) called it the eocyte tree but the name did not stick well. In the two domain tree — which incidentally fits very well with what some of us have been saying about eukaryote origin for a long time [35] — genes that trace to LUCA need not be present in eukaryotes at all. That is because in the two domain tree, eukaryote genomes arose from a very small sample of prokaryotic gene diversity, in the simplest case from the symbiotic association of two prokaryotic genomes in the form of an archaeal host with a bacterial symbiont, the ancestor of mitochondria and hydrogenosomes [36, 37].

Related to the issue of the three domain tree is the issue of how universal gene distributions need to be to trace a gene to LUCA. Regardless of where the root is, one can still look for genes that trace to LUCA by virtue of the density of their distribution. If one is strict, requiring that

genes be universally distributed across genomes, about 30-36 genes trace to LUCA [38-40]; if one allows for a bit of loss, about 100 genes trace to LUCA [41]; if one allows for a bit more loss, then about 500-600 genes trace to LUCA [42]; and if we allow for a lot of loss, then we are redirected to the issue above, namely that presence/absence patterns might be due to transdomain LGT rather than to differential loss, such that simple presence of a gene in bacteria and one archaeon or vice versa [15] is not solid ground for saying that said gene was present in LUCA.

In addition, if LUCA's gene set is defined in such a way that has to include genes that are present in eukaryotes (by the criterium of being present in *three* domains), then we quickly end up with an inference of LUCA that had a glycolytic pathway [42] and that used oxygen as a terminal acceptor [23], because that is how most eukaryotes obtain their energy [43]. But we know from physiology that the first free-living cells cannot have been chemoorganotrophs (satisfying their energy needs by the oxidation or disproportionation of reduced carbon compounds) because organics from space are nonfermentable substrates [44]. We also know from physiology that the producers of oxygen, cyanobacteria, represent a bioenergetically very advanced stage in physiological evolution [45,46], and thus cannot have preceded LUCA to generate oxygen for it to breathe. We also know from physiology that the mitochondria of many eukaryotes do not require oxygen for ATP synthesis [36].

Aware of the foregoing, we recently undertook a phylogenetic investigation based upon the two domain tree in search of insights into LUCA that might illuminate its microbial lifestyle [47]. Rather than looking for genes that are universally distributed (or nearly universally distributed), we looked for genes that trace to LUCA by virtue of being ancient. As our criterion for ancient, we looked for genes that are present in bacteria and archaea, but not because of LGT. This approach embraces the two domain tree, in which eukaryotes have nothing to do with life's origin, thereby excluding eukaryotes from the analysis. But how to exclude LGT? We looked for genes that fulfill two very simple criteria: i) the gene is present in two members each of two major groups of archaea and bacteria and ii) the domains are monophyletic. Genes that fulfill those criteria are unlikely to have a distribution that results from LGT.

In order to identify such genes, there is presently no obvious alternative to making all trees for all genes in all sequenced genomes and separating the wheat (the trees that show domain monophyly in the two domain tree) from the chaff (the trees that show archaea and bacteria interleaving). We have been making trees for large numbers of genes for some time [11,18, 48-50]. Trees for all genes are important because it has become evident that in prokaryotes, each gene has its own independent evolutionary history and that "trees of life", whether based on rRNA or the currently popular collection of ribosomal proteins [29,30,38] are not good proxies for what genes will be present in the rest of the genome and how those genes will be related to homologues from other genomes, because LGT is very prevalent among prokaryotes.

When we were done sorting the trees, what we found in our analysis were 355 genes that depict LUCA as an anaerobic autotroph that lived in a hot, gas-rich, metal-rich environment [47]. Its inferred energy metabolism was dependent upon H_2 and CO_2, it could fix N_2, it had a heavy dependence upon transition metals, its metabolism revealed an extremely prominent role for methyl groups, one electron transfers, radical reactions, and redox chemistry. Its carbon metabolism was based on the acetyl-CoA pathway, the oldest of the six known CO_2 fixation pathways. It was capable of substrate level phosphorylation using the acetyl-CoA pathway and it could harness chemiosmotic potential. It had modified bases, mostly involving methylations, suggesting that not only LUCA, but also the genetic code arose in an environment where reactive methyl groups were abundant. Previous studies had uncovered little information about LUCA's physiology and habitat. That is probably because earlier studies had focused on genes that are universally distributed (or nearly so). We also found that the trees of genes that trace to LUCA implicate clostridia (which harbour many acetogens) and methanogens as the earliest-branching forms of bacteria and archaea respectively. That fits with the functions of the genes we found, because acetogens and methanogens have carbon and energy metabolism that depends upon H_2 and CO_2, they can fix N_2, they have a heavy dependence upon transition metals, and their core physiology reveals an extremely prominent role for methyl groups, one electron transfers, radical reactions, and redox chemistry.

The results that we obtained fit very well with the idea that life arose in submarine hydrothermal vents and that the first cells were autotrophs that satisfy both their carbon and their energy needs from the reduction of CO_2 with electrons from H_2 [51-53]. Notably, H_2 is still continuously generated in modern hydrothermal vents today by the process of serpentinization [54], a spontaneous and exergonic geochemical reaction in which Fe^{2+} in oceanic crust reduces H_2O to generate H_2, which can reach many concentrations in vent effluent of many millimols per liter [55]. We found no evidence for a role of photosynthesis in LUCA's physiology, in particular there was no evidence for ZnS-based photosynthesis in LUCA (a physiology that is unknown among modern life forms anyway), in contrast to the predictions of some other recent theories [56]. Rather we found evidence linking LUCA to known forms of microbial physiology — acetogenesis and methanogenesis without cytochromes [57] — that are manifest among the strictest anaerobes [58, 59], with evidence for a role of sulfur metabolism [60], and with a very important role for Fe, Ni, Mo, and Co, transition metals that play a central role in the metabolism of anaerobic autotrophs today.

Our recent findings depart from phylogeny-based views of LUCA germane to the three domain tree and uncover connections between modern microbial physiology and geochemical environments on the early Earth. Some will surely complain that 355 genes is not enough and that essential functions like lipid synthesis, amino acid and nucleotide biosyntheses are very poorly represented in LUCA's gene set. How can anything live without that? As we

wrote, lack of such essential functions among LUCA's gene set could indicate i) that the missing genes unspectacularly underwent transdomain lateral gene transfer (LGT) post-LUCA and hence were filtered out by our method, ii) that some missing chemical components were provided by spontaneous abiotic syntheses during early Earth history, or iii) a combination thereof. Transdomain LGT is both normal and natural, and all theories for the origin of cells, without exception, require abiotic syntheses, hence we do not see any fundamental problems in that regard. There was a time on the early Earth when there was no life and there was a time when there was life. If we filter out the effects of 4 billion years of LGT — which is, in essence, what we did — a picture of LUCA emerges that represents something that was half-alive, an intermediate in the transition from rocks and water on a young, barren planet to something that could scratch a living out of gasses and mineral salts. For some reason, that sounds quite reasonable to us, others will surely disagree.

It is very interesting that acetogens and methanogens inhabit the crust today [10,61]. Geochemists say that the convective currents of water that permeate the Earth's crust to drive serpentinization have been going on since there was water on Earth [62]. Let us presume, just for a moment, that the first bacteria and archaea were acetogens and methanogens respectively. On an uninhabited planet, they have no competitors, and life multiplies quickly given ample growth substrates. The founders of their respective domains would have bubbled off into the ocean bottom waters to be spread around by currents and eventually to be introduced back into hydrothermal systems in the crust, where they would have found the diet that they were raised on. It is possible that some anaerobic autotrophs that live from the reduction of CO_2 with H_2 still inhabit the same *niche* in which life arose, albeit not the same *rocks* because during Earth history oceanic crust is constantly recycled into the mantle via subduction. In that sense, acetogens and methanogens really might provide a glimpse into the biology of the very first microbes on Earth, as some microbiologists familiar with the physiology of these organisms have been saying for some time [45, 60, 63].

Over four decades ago, biochemists thought that FeS clusters are ancient [64] and that acetogens and methanogens are ancient [45], based on good intuition, common sense, and some straightforward principles of physiology. With the discovery of archaea, the three domain tree led to avenues of thought about early evolution that were guided by phylogeny rather than physiology. LGT conflates phylogeny. But LGT does not conflate physiology, it just decouples it from phylogeny. When we filter out the LGT from all of the gene trees that we can make from genomes, we end up with a picture of LUCA that looks very much like what experts familiar with the physiology of anaerobes had in mind in the late 1960's [45], and still have in mind today [65, 66]. If we return to the geochemical record, the first evidence for life we see is evidence for autotrophs [3,4], which is also what genomes recently uncovered about LUCA [47]. Thus, on the issue of autotrophs being ancient, geology and physiology converge. The version of LUCA that is obtained by taking all the data and simply removing the obvious LGT interfaces well with Earth history, with microbial physiology, and even with the new two domain tree. It also bears out the predictions of some specific formulations the theory that life arose at submarine hydrothermal vents.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

COPYRIGHT

Please cite this article as: William F. Martin, Madeline C. Weiss, Sinje Neukirchen, Shijulal Nelson-Sathi, Filipa L. Sousa (2016). Physiology, phylogeny, and LUCA. **Microbial Cell** 3(12): 451-456. doi: 10.15698/mic2016.12.545

REFERENCES

1. Gaidos E, and Knoll AH (**2012**). Our evolving planet: From dark ages to evolutionary renaissance. In: Impey C, Lunine J, Funes J, editors. Frontiers of Astrobiology. **Cambridge University Press, Cambridge**; pp 132-153.

2. Mojzsis SJ, Arrhenius G, McKeegan KD, Harrison TM, Nutman AP and Friend CR (**1996**). Evidence for life on Earth before 3,800 million years ago. **Nature** 384: 55-59.

3. Ueno Y, Yurimoto H, Yoshioka H, Komiya T, and Maruyama S (**2002**). Ion microprobe analysis of graphite from ca. 3.8 Ga metasediments, Isua crustal belt, West Greenland: Relationship between metamorphism and carbon isotopic composition. **Geochimia Et Cosmochimia Acta** 66(7):1257–1268.

4. Bell EA, Boehnke P, Harrison TM, and Mao WL (**2015**). Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. **Proc Nat Acad Sci U S A** 112(47):14518–14521.

5. Fischer WW, Hemp J, and Johnson JE (**2016**). Evolution of oxygenic photosynthesis, **Ann Rev Earth Planet Sci** 44:647-683.

6. Rabus R, Venceslau SS, Wöhlbrand L, Voordouw G, Wall JD, and Pereira IAC (**2015**). A post-genomic view of the ecophysiology, catabolism and biotechnological relevance of sulphate-reducing prokaryotes. **Adv Microb Physiol** 66:55–321.

7. Barker HA (**1961**). Fermentation of nitrogenous compounds. In: Gunsalus IC and Stanier RY, editors. The Bacteria. A Treatise on Structure and Function. **Academic Press, New York** pp. 151-207.

8. Marreiros BC, Calisto F, Castro PJ, Duarte AM, Sena FV, Silva AF, Sousa FM, Teixeira M, Refojo PN, and Pereira MM (**2016**). Exploring membrane respiratory chains. **Biochim Biophys Acta** 1857(8):1039-1067.

9. Raymann K, Brochier-Armanet C, and Gribaldo S (**2015**). The two-domain tree of life is linked to a new root for the Archaea. **Proc Nat Acad Sci U S A** 112:6670–6675.

10. Evans PN, Parks DH, Chadwick GL, Robbins SJ, Orphan VJ, Golding SD, and Tyson GW (**2015**). Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. **Science** 350:434–438.

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

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

13. Ouzounis CA, Kunin V, Darzentas N, and Goldovsky L (**2006**). A minimal estimate for the gene content of the last universal common ancestor–exobiology from a terrestrial perspective. **Res Microbiol** 157:57-68.

14. Castresana J (**2001**). Comparative genomics and bioenergetics. **Biochim Biophys Act** - **Bioenerg** 1506:147–162.

15. Nitschke W, and Russell MJ (**2013**). Beating the acetyl coenzyme A-pathway to the origin of life. **Phil Trans Roy Soc Lond B**. 368:20120258.

16. Kannan L, Li H, Rubinstein B, and Mushegian A (**2013**). Models of gene gain and gene loss for probabilistic reconstruction of gene content in the last universal common ancestor of life. **Biol Direct** 8:32.

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

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

19. Iwabe N, Kuma K, Hasegawa M, Osawa S, and Miyata T (**1989**). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. **Proc Natl Acad Sci U S A** 86:9355-9359.

20. Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, Konishi J, Denda K, and Yoshida M (**1989**). Evolution of the vacuolar H^+-ATPase: Implications for the origin of eukaryotes. **Proc Natl Acad Sci U S A**, 86:6661–6665.

21. Forterre P (**1995**). Thermoreduction, a hypothesis for the origin of prokaryotes. **C R Acad Sci III** 318:415-422.

22. Poole A, Jeffares D, and Penny D (**1999**). Early evolution: prokaryotes, the new kids on the block. **BioEssays** 21: 880-889.

23. Glansdorff N, Xu Y, and Labedan B (**2008**). The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner. **Biol Direct** 3:29.

24. Harish A, Tunlid A, and Kurland CG (**2013**). Rooted phylogeny of the three superkingdoms. **Biochimie**. 95:1593–1604.

25. Di Giulio M (**2015**). The non-biological meaning of the term "Prokaryote" and its implications. **J Mol Evol** 80:98–101.

26. Forterre P (**2015**). The universal tree of life: an update. **Front Microbiol** 6:717.

27. Whitman EB (**2009**). The modern concept of the procaryote. **J Bact** 191:2000-2005.

28. Cox CJ, Foster PG, Hirt RP, Harris SR and Embley TM (**2008**). The archaebacterial origin of eukaryotes. **Proc Nat Acad Sci U S A** 105:20356–20361.

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

30. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, and Ettema TJ (**2015**). Complex archaea that bridge the gap between prokaryotes and eukaryotes. **Nature** 521:173–179.

31. McInerney J, Pisani D, and O'Connell MJ (**2015**). The ring of life hypothesis for eukaryote origins is supported by multiple kinds of data. **Phil Trans R Soc B** 370:20140323.

32. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, and Banfield JF (**2016**). A new view of the tree of life. **Nature Microbiol** 1:16048.

33. Lake JA (**1988**). Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. **Nature** 331:184–-186.

34. Forterre P (**2013**). The common ancestor of archaea and eukarya was not an archaeon. **Archaea**. 2013: 372396.

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

36 . 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 .

37. Sousa FL, Neukirchen S, Allen JF, Lane N, and Martin WF (**2016**). Lokiarchaeon is hydrogen dependent. **Nature Microbiol**. 1:16034.

38. Hansmann S, and Martin W (**2000**). Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. **Int J Syst Evol Microbiol** 50(4):1655-1663.

39. Charlebois RL, and Doolittle WF (**2004**). Computing prokaryotic gene ubiquity: Rescuing the core from extinction. **Genome Res** 14: 2469-2477.

40. Ciccarelli FD, Doerks T, Mering von C, Creevey CJ, Snel B, and Bork P (**2006**). Toward automatic reconstruction of a highly resolved Tree of Life. **Science** 311, 1283–1287.

41. Puigbò P, Wolf YI, and Koonin EV (**2009**). Search for a 'Tree of Life' in the thicket of the phylogenetic forest. **J Biol** 8:59.

42. Koonin EV (**2003**). Comparative genomics, minimal gene-sets and the last universal common ancestor. **Nat Rev Microbiol** 1: 127–136.

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

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

45. Decker K, Jungerman K, and Thauer RK (**1970**). Energy production in anaerobic organisms. **Angew Chem Int Ed**. 9:138–158.

46. Martin WF, and Sousa FL (**2016**). Early microbial evolution: the age of anaerobes. **Cold Spring Harbor Persp Biol** 8:a018127.

47. 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. **Nature Microbiol** 1(9):16116.

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

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

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

51. Russell MJ, and Martin W (**2004**). The rocky roots of the acetyl-CoA pathway. **Trends Biochem Sci** 29:358–363.

52. Martin W, and Russell MJ (**2007**). On the origin of biochemistry at an alkaline hydrothermal vent. **Phil Trans Roy Soc Lond B** 367:1887–1925.

53. Lane N, and Martin WF (**2012**). The origin of membrane bioenergetics. **Cell** 151:1406–1416.

54. Russell MJ, Hall AJ, and Martin W (**2010**). Serpentinization as a source of energy at the origin of life. **Geobiol** 8:355–371.

55. Schrenk MO, Brazelton WJ, and Lang SQ (**2013**). Serpentinization, carbon, and deep life. **Rev Mineral Geochem** 75:575–606.

56. Mulkidjanian AY, and Galperin MY (**2009**) On the origin of life in the zinc world. 2. Validation of the hypothesis on the photosynthesizing zinc sulfide edifices as cradles of life on Earth. **Biol Direct** 4:27.

57. Buckel W, and Thauer RK (**2013**). Energy conservation via electron bifurcating ferredoxin reduction and proton/Na^+ translocating ferredoxin oxidation. **Biochim Biophys Acta** 1827: 94–113.

58. Thauer RK, Kaster AK, Seedorf H, Buckel W, and Hedderich R (**2008**). Methanogenic archaea: ecologically relevant differences in energy conservation. **Nat Rev Microbiol**. 6:579-59.

59. Schuchmann K, and Müller V (**2014**). Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria. **Nat Rev Microbiol**. 12:809–821.

60. Liu Y, Beer LL, and Whitman WB. (**2012**). Methanogens: a window into ancient sulfur metabolism. **Trends Microbiol**. 20:251 – 258.

61. Lever MA (**2012**). Acetogenesis in the energy-starved deep biosphere—a paradox? **Front Microbiol** 2:284.

62. Sleep NH, Meibom A, Fridriksson T, Coleman RG, and Bird DK (**2004**) H_2-rich fluids from serpentinization: geochemical and biotic implications. **Proc Natl Acad Sci U S A** 101:12818-12823.

63. Ferry JG, House CH (**2006**). The step-wise evolution of early life driven by energy conservation. **Mol Biol Evol** 23:1286–1292.

64. Eck RV, and Dayhoff MO (**1966**). Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. **Science** 152:363–366.

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

66. Basen M, and Müller V (**2016**). "Hot" acetogenesis. **Extremophiles** 1-12.

5.4 Wo lebten die ersten Zellen - und wovon?

William F. Martin, Verena Zimorski, Madeline C. Weiss

Affiliations

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

Dieser Artikel wurde am 14. Juni 2017 in *Biologie in unserer Zeit* Ausgabe 47 veröffentlicht.

Beitrag von Madeline C. Weiß:

Für die Publikation wurde von mir die Literaturrecherche durchgeführt sowie der Text korrektur gelesen. Des Weiteren habe ich die Abbildung 2 nach Vorgaben von Prof. Dr. Martin erstellt und die Abbildung 3 übersetzt und formatiert. Die Abbildung 3 stammt original aus der Publikation „The physiology and habitat of the last universal common ancestor“, Weiss *et al.* (2016b).

DOI:10.1002/biuz.201710622

Frühe Evolution

Wo lebten die ersten Zellen – und wovon?

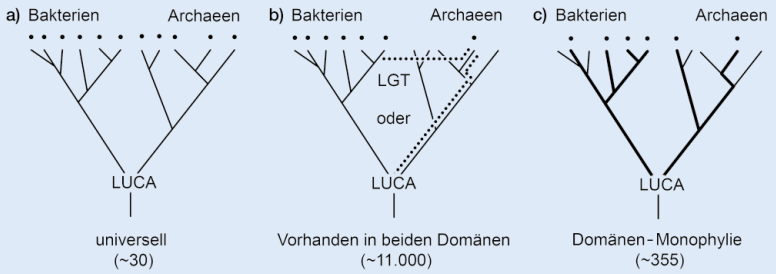
WILLIAM F. MARTIN | VERENA ZIMORSKI | MADELINE C. WEISS

Der letzte gemeinsame Vorfahre allen Lebens (Luca – Last Universal Common Ancestor) nimmt eine zentrale Rolle bei Studien zur frühen Evolution und zum Ursprung des Lebens ein. Luca verbindet die abiotische Phase der Erdgeschichte mit den ersten geochemischen Spuren mikrobiellen Lebens. Bisher gab es keine Informationen, wo und wovon Luca gelebt hat. Eine neue Studie identifizierte 355 Gene, die über Lucas Lebensweise und seinen Lebensraum Auskunft geben.

ABB. 1 Tiefsee-Hydrothermalquellen waren vermutlich das Habitat von Luca: Weiße Raucher im Westpazifischen Ozean in ca. 1500 m Tiefe. Die „Schornsteine" sind 20 cm breit und 50 cm hoch und geben 103 °C heißes Wasser ab. Bild: NOAA Photo Library.

Online-Ausgabe unter: wileyonlinelibrary.com

ABB. 2 | WEGE, UM RÜCKSCHLÜSSE AUF GENE VON LUCA ZU ZIEHEN



a) Gene, die universell in allen Genomen vorhanden sind, führen zu Luca. Ungefähr 30 Gene erfüllen dieses Kriterium [2], und circa 100, wenn ein paar Verluste erlaubt sind [3]. Die Anwesenheit der Gene wird jeweils mit schwarzen Punkten angezeigt.
b) Ein weiterer Weg, um Gene zu Luca zurückzuverfolgen: Gene, die sowohl in Archaeen als auch in Bakterien vorhanden sind, werden als Lucas Gene betrachtet [5]. Ein moderner Genomdatensatz von circa 11.000 Genen erfüllte dieses Kriterium [1]. Ein Genom mit 11.000 Genen ist allerdings viel zu groß für Luca. Viele Gene, die in den heutigen Bakterien und Archaeen vorhanden sind, gab es nicht in Luca, sie wurden zwischen Bakterien und Archaeen über die Domänengrenzen hinweg transferiert. In der Tat wurden tausende von Genen identifiziert, welche zwischen Bakterien und Archaeen transferiert wurden [6].
c) Durch das Betrachten der phylogenetischen Bäume können lateral transferierte Gene herausgefiltert werden. Was übrigbleibt sind Gene, die die Domänen-Monophylie erhalten. Gene, die nur in einem Bakterien- oder Archaeen-Phylum vorhanden sind, können das Ergebnis lateralen Gentransfers sein. Das Vorhandensein in zwei Phyla pro Domäne ist bei gleichzeitiger Domänen-Monophylie schwieriger durch lateralen Gentransfer zu erhalten. Solche Gene sind gute Kandidaten für Luca [1].

Das Leben ist eine chemische Reaktion, eine zwar sehr komplizierte, aber eine chemische Reaktion. In allen Formen des Lebens gibt es exergone (Energie liefernde) Reaktionen im zentralen Kern des Energiestoffwechsels, die jene chemische Energie liefern, die notwendig ist, um alle individuellen Reaktionen in der Zelle in die richtige Richtung zu steuern: weg vom Zustand des Gleichgewichts. Alle Lebewesen gehen auf einen gemeinsamen Vorfahren zurück, weil alle Zellen denselben universellen genetischen Code verwenden. Allen Lebewesen gemeinsam ist auch das Prinzip der Energiekonservierung. Aber die chemischen Reaktionen, die das Leben nutzt, damit Zellen Energie aus ihrer Umgebung umwandeln können, zeigen eine fast endlose Vielfalt.

Diese Vielfalt im Energiestoffwechsel war nicht vom Anfang an da, sondern hat sich entwickelt. Die ersten Zellen, die den genetischen Code nutzten, hatten ebenfalls einen Energiestoffwechsel. Wie haben die ersten Zellen gelebt? Wo haben sie gelebt und vor allem wovon? Diese Fragen haben wir mithilfe mikrobieller Genomdaten untersucht [1]. Die Daten sprechen dafür, dass der letzte universelle gemeinsame Vorfahr allen Lebens wahrscheinlich von Gasen lebte – Wasserstoff (H_2), Kohlenstoffdioxid (CO_2), Kohlenmonoxid (CO) und Stickstoff (N_2) – und zwar in einer Umgebung, die heutigen Tiefsee-Hydrothermalquellen (Abbildung 1) sehr ähnlich war. Diese Ergebnisse lieferten erste Einblicke in die Physiologie und das Habitat von Luca.

Rekonstruktion Lucas durch bioinformatische Analysen

Es gibt viele genombasierte Untersuchungen zu Luca. Diese lassen sich in zwei generelle Kategorien einteilen. Die klassische Herangehensweise ist, eine Stichprobe an Genomen aus allen Domänen des Lebens zu nehmen und herauszufinden, welche Gene allen Genomen gemeinsam sind. Gene, die in allen modernen Lebensformen vorkommen, waren im Rückschluss auch in Luca vorhanden (Abbildung 2a). Unter Anwendung strikter Kriterien für die Definition des universellen Auftretens von Genen (d. h. wirklich in allen Genomen der Stichprobe), konnten etwa 35 Gene auf Luca zurückgeführt werden [2]. Gene können aber in der Evolution auch verloren gehen, weshalb „universell vorhanden“ als Kriterium vielleicht zu strikt ist. Wenn die Kriterien ein wenig gelockert werden, um auch Genverluste erfassen zu können, wächst die Liste an universellen Genen auf ungefähr 100 Gene. Dabei werden Fälle einbezogen, bei denen ein Gen nur in einigen Linien verloren ging. Diese Gene sind dann immerhin annähernd universell [3]. Universell und annähernd universell konservierte Gene codieren typischerweise für Proteine, die mit Ribosomen, also der Translation, oder anderen Aspekten der Informations-Verarbeitung assoziiert sind. Dies zeigt, dass die ersten Zellen Energie nutzten, weil sie Ribosomen synthetisieren und die Translation in Gang bringen mussten. Die Proteinsynthese ist die kostspieligste Reak-

IN KÜRZE

- *Alle Lebewesen gehen auf **einen gemeinsamen Vorfahren** zurück: Luca (Last universal common ancestor).*
- *Wo haben die ersten Zellen gelebt? Wovon haben die ersten Zellen gelebt? Diesen Fragen ging eine neue Studie nach. Dabei wurden mittels bioinformatischer Analysen mikrobieller Genomdaten **355 Gene** identifiziert, die über Lucas Lebensweise und seinen Lebensraum Auskunft geben.*
- *Die ersten Zellen lebten wahrscheinlich in einer Umgebung, die heutigen **Tiefsee-Hydrothermalquellen** sehr ähnlich war.*
- *Luca war wahrscheinlich **anaerob, thermophil und lebte von Gasen**: Wasserstoff, Kohlendioxid, Kohlenmonoxid und Stickstoff. Sein Energiestoffwechsel ähnelt stark demjenigen heutiger acetogener Bakterien und methanogener Archaeen und zeigte zudem Ähnlichkeiten zu den geochemischen Reaktionen an Hydrothermalquellen.*
- *Kurze **Animationsfilme zum Ursprung des Lebens** und dem symbiotischen **Ursprung der Eukaryoten** finden Sie im Internet unter www.molevol.hhu.de/movies.html.*

tion einer Zelle und kostet sehr viel ATP. Ungefähr 75 % der Zellenergie (ATP-Verbrauch) fließt in die Proteinsynthese [4]. Wer Gene hatte und Proteine synthetisierte, hatte daher Energie, aber woher?

Ein anderer Ansatz, um Lucas Eigenschaften aus Genomen abzuleiten, ist es, nicht einige Genverluste, sondern beliebig viele Genverluste zuzulassen. Dieses Vorgehen bei der Untersuchung der Genomdaten führt dazu, dass jedes Gen, das sowohl in Bakterien als auch in Archaeen vorkommt, auf Luca zurückzuführen ist [5]. Ein Problem dieses Ansatzes ist, dass Gene, die in Vertretern beider Domänen des Lebens vorhanden sind, auf zwei Wegen zu ihrer heutigen Verteilung gekommen sein könnten (Abbildung 2b).

Sie könnten entweder in Luca vorhanden gewesen und dann unterschiedlich verloren gegangen sein oder sie könnten in einer Linie entstanden sein, die lange nach Luca gelebt hat, und anschließend von Bakterien zu Archaeen (oder umgekehrt) über lateralen Gentransfer weitergegeben worden sein. Eine unserer aktuellen Arbeiten zeigt, dass tausende solcher Gentransfers zwischen Bakterien und Archaeen während der Evolution stattgefunden haben [6]. Evolutionär spät entstandene Gene, zum Beispiel diejenigen für Proteine der Sauerstoffatmung, wurden von den Bakterien über die Domänengrenze hinweg zu den Archaeen transferiert. Solche interdomän-transferierten Gene zeigen dieselbe *Verteilung* wie Gene, die in Luca vorhanden waren, d.h. sie sind in Bakterien und Archaeen vorhanden. Deshalb gilt: Gene, die in Bakterien und Archaeen vorkommen, müssen nicht bereits in Luca vorhanden gewesen sein. Sie könnten genau so gut spät in der Evolution entstanden und mittels lateralen Gentransfers zwischen den Domänen verteilt worden sein. Genstammbäume können hier weiterhelfen, denn sie zeigen, ob ein Gen über lateralen Gentransfer verteilt wurde oder nicht.

Zur Erforschung Lucas haben wir daher eine neue Herangehensweise mit Hilfe von Genomdaten gewählt (Abbildung 2c). Bei ca. 6.000.000 Proteinen, die in etwa 2.000 Genomen codiert sind, haben wir nicht nach den universellen Genen gesucht, und auch nicht nach den Genen, die nur in Archaeen und Bakterien vorkommen. Stattdessen haben wir gefragt: Welche Gene kommen zwar in Bakterien und Archaeen vor, aber nicht aufgrund von lateralem Gentransfer zwischen den Domänen? Diese Gene sollten in Luca vorhanden gewesen und seit Lucas Zeit vertikal innerhalb der Domänen vererbt worden sein. Wie aber identifiziert man diese Gene? Dazu muss man Stammbäume erstellen, viele Stammbäume und zwar von jedem Gen, das einen Stammbaum abbilden kann. Diese Stammbäume wurden anschließend nach zwei einfachen Kriterien gefiltert, die Gene mit lateralem Gentransfer (die Spreu) von solchen mit vertikaler Vererbung (dem Weizen) trennen sollten:

1. Das Gen (somit auch das korrespondierende Protein) sollte in mindestens zwei höheren Taxa der Bakterien und Archaeen vorhanden sein und
2. der korrespondierende Stammbaum des Proteins sollte eine Monophylie der Bakterien und der Archaeen zeigen.

Bei Genen (Proteinen), die diese beiden Kriterien erfüllen, ist es unwahrscheinlich, dass sie einem Gentransfer zwischen den Domänen unterlagen. Sie waren daher wahrscheinlich im Genom von Luca vorhanden. Zum Glück ist es nicht erforderlich, alle Stammbäume per Hand zu analysieren. Hierzu gibt es Hochleistungscomputer, die allerdings für unsere Berechnungen immer noch mehrere Monate benötigten. Als alle 11.093 Stammbäume, die Sequenzen aus Bakterien und Archaeen enthielten, nach Gentransfer zwischen den Domänen durchforstet waren, blieb nur eine Liste mit 355 Genen übrig. Diese 355 Gene definieren jeweils eine Proteinfamilie. Diese 355 Gene waren nicht ursprünglich allein anhand von Verteilungskriterien, sondern sie waren ursprünglich anhand von phylogenetischen Kriterien. Das hat es vorher nicht gegeben.

Was Lucas Gene und die ursprünglichen Mikroben gemeinsam haben

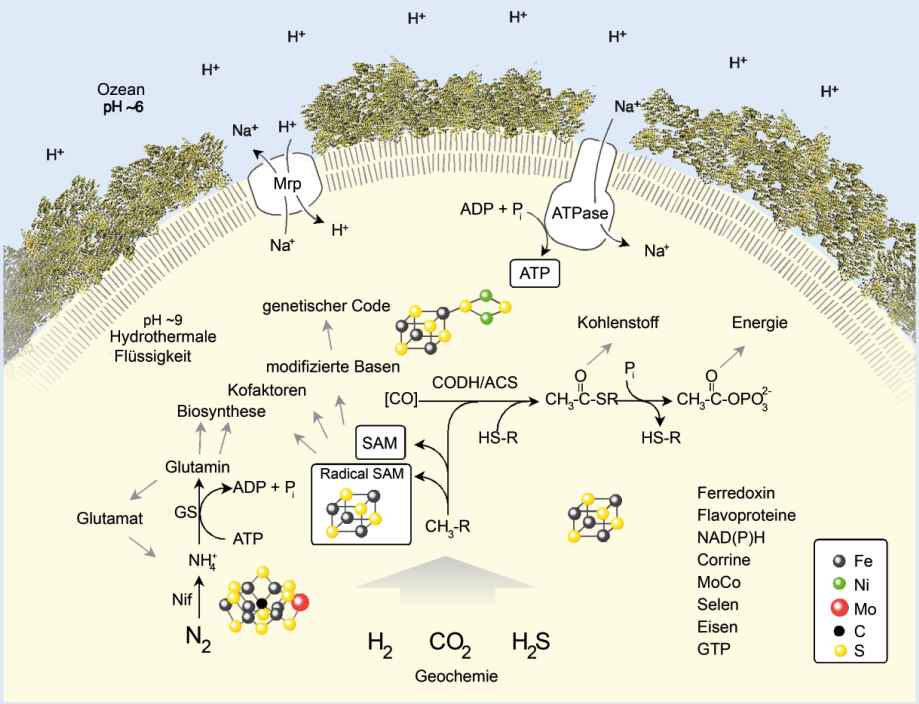
Die Überraschung war nicht die Anzahl an Genen, die wir gefunden haben (355), sondern wofür die Gene codieren. Für mehr als die Hälfte der Proteinfamilien gab es Informationen über die biochemische Funktion des codierten Proteins in den Datenbanken. Lucas Gene codierten für Enzyme, die typisch für heutige Zellen sind, die in strikter Abwesenheit von Sauerstoff von Substanzen leben, die nachweislich auf der frühen Erde vorkamen – H_2, CO_2, CO, N_2. Der Kern des Energiestoffwechsels von Luca ähnelt damit stark demjenigen in heutigen anaeroben, wasserstoffabhängigen Chemolithoautotrophen. Kurz gesagt, Luca hatte einen Lebensstil, der sehr ähnlich zu dem einiger heutiger Prokaryoten ist: den acetogenen Bakterien und den methanogenen Archaeen. Acetogene Bakterien bilden aus den Substraten Kohlendioxid und Wasserstoff (Elektronendonor) das Endprodukt Essigsäure, während die methanogenen Archaeen aus denselben Substraten Methan als Endprodukt bilden. Der Energiestoffwechsel in beiden Gruppen dieser modernen anaeroben autotrophen Mikroben zeigt somit Ähnlichkeiten zu den geochemischen Reaktionen an Hydrothermalquellen.

Acetogene Bakterien und methanogene Archaeen nutzen den an geochemische Reaktionen erinnernden reduktiven Acetyl-CoA-Weg (auch Wood-Ljungdahl-Weg genannt) als Hauptweg der CO_2-Fixierung. Der reduktive Acetyl-CoA-Weg ist der ursprünglichste von sechs bekannten Wegen der CO_2-Fixierung und der einzige, der sowohl in Archaeen als auch in Bakterien vorkommt. Der Acetyl-CoA-Weg gliedert sich in zwei Segmente: die Methyl-Synthese und die Acetyl-Synthese. Während die Methyl-Synthese durch Wasserstoff und CO_2 bei Archaeen und Bakterien von jeweils anderen, strukturell nicht verwandten Enzymen und verschiedenen Kofaktoren katalysiert wird, ist die Acetyl-Synthese durch Methylgruppen und Kohlenmonoxid gleichartig und konserviert unter den Archaeen und Bakterien und es werden ausschließlich Übergangsmetalle zur Katalyse genutzt. Als Konsequenz nutzen die acetogenen Bakterien Kohlendioxid und Wasserstoff

(Elektronendonor) als Substrate, um daraus das Endprodukt Essigsäure zu bilden. Die Methanogenen bilden aus denselben Substraten Methan als Endprodukt. Diese Unterschiede deuten daraufhin, dass die Acetyl-Synthese im letzten gemeinsamen Vorfahren aller Mikroben vorhanden war, die Methyl-Synthese sich jedoch später und unabhängig voneinander in den Abstammungslinien, die zu den Acetogenen und Methanogenen führten, entwickelte. In Luca wurden nur Gene des reduktiven Acetyl-CoA-Wegs gefunden. Somit sprechen die Daten dafür, dass Luca diesen Weg der CO_2-Fixierung nutzte, wie die heutigen acetogenen Bakterien und methanogenen Archaeen.

Aber Lucas Energiestoffwechsel war vermutlich noch einfacher als der heutiger Acetogener und Methanogener. Wir fanden zum Beispiel keine Hinweise auf Proteine, die an der sogenannten Elektronenbifurkation [7] beteiligt sind: Wenn moderne Methanogene und Acetogene CO_2 mit Elektronen aus H_2 reduzieren, müssen sie zunächst reduziertes

ABB. 3 | HAUPTINTERAKTIONEN VON LUCA

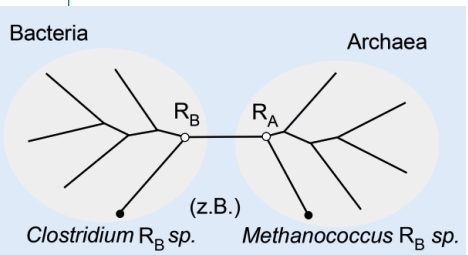


Gezeigt werden die Hauptinteraktionen von Luca mit seiner geochemischen Umgebung, wie sie von den Genomdaten abgeleitet werden können. Diese Umgebung war vermutlich heutigen Hydrothermalquellen ähnlich [1]. Luca lebte von Gasen (H_2, CO_2, H_2S, CO, N_2), die Ausgangspunkt der Energiegewinnung und der Biosynthese von Kofaktoren, modifizierten Basen und des genetischen Codes waren.
In Luca waren Ferredoxin (mit 4Fe-4S Cluster; schematisch links neben Ferredoxin dargestellt), Flavoproteine, Reduktionsäquivalente (NAD(P)H), Corrine, Molybdän-Cofaktoren (MoCo), Selen, Eisen und GTP vorhanden. Ein Na^+/H^+-Antiporter (Mrp; links oben) könnte einen geochemischen pH-Gradienten wie er in alkalinen Hydrothermalquellen vorkommt, in einen stabileren Na^+-Gradienten umgewandelt haben, um eine erste Na-abhängige ATP-Synthase anzutreiben.
Luca besaß unbestreitbar Gene, weil er den genetischen Code besaß. Die Frage, welche Gene hier vorhanden waren, war bisher schwierig zu beantworten. Übergangsmetalle (Fe, Ni, Mo) sind im katalytischen Zentrum (schematisch dargestellt) der Nitrogenase (Nif) und der CODH/ACS vorhanden und spielen somit eine wichtige Rolle beim Energiestoffwechsel und bei Biosynthesen. Die Abbildung macht keine Aussagen über die Quelle des Kohlenmonoxids in ursprünglichen Metaboliten (entweder unkatalysiert z. B. über die Wassergas-Shift-Reaktion oder katalysiert von Übergangsmetallen). Daher wurde Kohlenmonoxid durch [CO] dargestellt. Für weitere Einzelheiten siehe [1].
Abkürzungen: CODH/ACS, Kohlenstoffmonoxid Dehydrogenase/Acetyl CoA-Synthase; Nif, Nitrogenase; GS, Glutaminsynthase; Mrp, MrP Typ Na^+/H^+ Antiporter; CH3-R, Methylgruppen; HS-R organisches Thiol.
Bild: mit freundlicher Genehmigung der Nature Publishing Group entnommen aus [1].

Ferredoxin erzeugen, was nur über einen bioenergetischen Trick geht: Um ein Elektron im Elektronenpaar von H_2 energetisch „bergauf" auf Ferredoxin zu befördern, wird das andere energetisch „bergab" geschickt, zum Beispiel auf NAD^+ bei den Acetogenen, oder auf das Heterodisulfid CoM-S-S-CoB bei den Methanogenen. Elektronenbifurkation ist eine neue Art der energetischen Kopplung im Stoffwechsel, sie wurde von Rolf Thauer und Wolfgang Buckel in Marburg entdeckt [7] und ist erst seit etwa zehn Jahren bekannt. Und woher bezog Luca reduziertes Ferredoxin? Möglicherweise direkt von der FeS-haltigen geochemischen Umwelt, in der Luca selbst gelebt hat (Abbildung 3).

Die Phylogenien der 355 Proteinfamilien in unserer Analyse haben Bakterien auf der einen Seite der Wurzel und Archaeen auf der anderen Seite. „Reziprok gewurzelt" ist der technische Ausdruck für solche Stammbäume, ein schematisches Beispiel für einen reziprok gewurzelten Baum wird in Abbildung 4 gezeigt. Für Bäume, in denen die Archaeen und Bakterien monophyletisch waren, definiert der Ast, der die Domänen verbindet, zugleich die Wurzel beider Teilbäume. Setzten die Bakterien im Teilbaum für die Archaeen an einem Ast an, der zu einer *Methanococcus*-Art führt, so erscheint diese als das frühestabzweigende Archaeon. Analog verhält es sich bei den Bakterien, sollte der Archaeen-Ast bei einem *Clostridium*-Vertreter ansetzen. In einigen Fällen kann der frühest abzweigende Ast Vertreter mehrerer Taxa enthalten. In solchen Fällen zählt man die einzelnen Vertreter anteilig. Experten werden merken, dass an der Wurzel entweder der Ast mit vielen Arten („Blätter" ist hierfür der Fachausdruck) oder mit wenigen Arten (Blättern) als die ursprüngliche Gruppe angesehen werden kann. In solchen Fällen kann man sowohl die Anzahl der Blätter als auch die Astlängen (Wurzel bis Spitze) hinzuziehen, um den frühesten Abzweig zu ermitteln (für Details, siehe [1]). Die reziprok gewurzelten Bäume können genutzt werden, um zu fragen, welche modernen Gruppen der Mikroben am tiefsten in den phylogenetischen Stammbäumen abzweigen. Erneut war die Antwort: Acetogene (Clostridien) und Methanogene. Es gab zudem Hinweise darauf, dass Luca thermophil war, da Gene für das Enzym Reverse Gyrase vorhanden waren. Dieses katalysiert ein ATP-abhängiges positives Supercoiling (Überspiralisierung der DNA), wodurch DNA repariert und stabilisiert wird. Reverse Gyrase ist typisch für hyperthermophile Organismen. Auch waren Lucas Enzyme übersät mit Übergangsmetallen als Elektronenträger und Katalysatoren, insbesondere mit FeS- und FeNiS-Zentren. Zusammenfassend deutet unsere Analyse darauf hin, dass der letzte gemeinsame Vorfahre aller Zellen in einer heißen Umgebung, in der Metalle und metallische Sulfide reichlich vorkommen, mit Gasen als Kohlenstoff- und Energiequelle heranwuchs (Abbildung 3).

ABB. 4 | EIN REZIPROK GEWURZELTER BAUM FÜR BAKTERIEN UND ARCHAEEN



R_B, Wurzel (engl. root) für die Bakterien. R_A, Wurzel für die Archaeen.

Lucas Nachfahren

Wo leben Acetogene und Methanogene heutzutage? Sie bewohnen viele strikt anaerobe Habitate, in denen H_2 als wichtigster chemischer Treibstoff für die CO_2-Reduktion reichlich vorkommt. Das kann der Verdauungstrakt von Tieren sein, welcher bekanntlich keine ursprüngliche Umgebung ist. Oder es kann organisches Sediment am Seegrund oder in Ozeanen sein, was ebenfalls keine ursprüngliche Umgebung ist, weil es vor vier Milliarden Jahren zu Lucas Zeit kein fermentierbares organisches Sediment gab [8]. Oder es kann die Erdkruste sein, die eine ursprüngliche Umgebung ist. Sowohl Methanogene, als auch Acetogene leben in der Erdkruste [9, 10] und gewinnen ihre Energie durch die Produktion von Methan oder Acetat aus H_2, CO_2 und CO. Auf der frühen Erde war alles strikt anaerob, weil Sauerstoff ein Produkt der biologischen Evolution ist. Es gab reichlich CO_2, vielleicht 1000-mal mehr als heute in den Ozeanen vorkommt, während das Angebot an Wasserstoff beschränkter war. Woher kam dann H_2 als Treibstoff für Lucas Stoffwechsel?

Heute gibt es zwei Hauptquellen für Wasserstoff in der Umwelt. Einerseits wird H_2 von Mikroben während der Fermentation von faulender Biomasse produziert, andererseits wird H_2 geochemisch in der Erdkruste durch einen Prozess gebildet, der Serpentinisierung heißt – benannt nach dem Mineral Serpentinit ($Mg_{2,85}Fe_{0,15}Si_2O_5(OH)_4$), das dabei gebildet wird. Dieser Prozess tritt auf, wenn Wasser durch die Kruste hydrothermaler Systeme zirkuliert. Zu Lucas Zeit gab es keine Biomasse, die hätte fermentiert werden können. Luca war ein Pionier auf einem zuvor unbesiedelten Planeten voller Gestein, Wasser und CO_2. Biologischer Wasserstoff war somit nicht verfügbar. Geochemischer Wasserstoff war dagegen im Überfluss vorhanden, da Hydrothermalquellen seit dem Vorkommen von Wasser auf der Erde reichlich Wasserstoff produzieren [11]. Wenn Methanogene und Acetogene die ersten Zellen auf der Erde gewesen sind, wie es Lucas Daten nahelegen, und sie in der Erdkruste lebten, könnte es sein, dass sie noch heute am selben Ort leben, an dem Leben entstand? Ja – sie leben noch im selben chemischen Habitat, aber nicht am gleichen Ort im engeren Sinne, weil die Erdkruste im Lauf der Jahrmillionen ständig durch den Erdmantel über Subduk-

tion (ein nach unten gerichteter Fluss) und Wiederauftreten als Magma an Spreizungszonen erneuert wird. Die ökologische Nische der Tiefseekruste ist jedoch uralt.

Die Serpentinisierung als Quelle des Lebens

Wasserstoff stammt von Interaktionen zwischen Gestein und Wasser in der Erdkruste. Der Prozess der Serpentinisierung wurde von Geochemikern gut charakterisiert [11]. Dabei wird Wasser (durch die Erdanziehungskraft) bis kilometertief in Rissen der Erdkruste nach unten gezogen. Die Tiefseekruste besteht hauptsächlich aus Eisen-Magnesium-Silikaten. Vor vier Milliarden Jahren (und zum größten Teil auch noch heute) lag das Eisen vorwiegend in der Oxidationsstufe Fe^{2+} vor, wie im Mineral Olivin ($Mg_{1,8}Fe_{0,2}SiO_4$), das mit Wasser während der Serpentinisierung reagiert (siehe unten stehende Reaktionsgleichungen). Dabei wird Fe^{2+} zu Fe^{3+}, wie es im Magnetit vorliegt. Fe^{3+} hat ein Elektron weniger als Fe^{2+}, und Fe^{2+} wird durch Wasser oxidiert. Während der Serpentinisierung werden die Elektronen auf Wasser übertragen und dabei Wasserstoff erzeugt. H_2 ist als Gas im Ausstrom hydrothermaler Systeme in Konzentrationen der Größenordnung von 10 mM oder mehr gelöst [12] und wird in den Ozean abgegeben.

$2\ Mg_{1,8}Fe_{0,2}SiO_4$ (Olivin) + $3\ H_2O \rightarrow Mg_{2,85}Fe_{0,15}Si_2O_5(OH)_4$ (Serpentinit) + $Mg_{0,75}Fe_{0,25}(OH)_2$ (eisenhaltiges Bucit)

$57\ Mg_{0,75}Fe_{0,25}(OH)_2$ (eisenhaltiges Brucit) + $30\ SiO_2$ (aq) $\rightarrow 15\ Mg_{2,85}Fe_{0,15}Si_2O_5(OH)_4$ (Serpentinit) + $23\ H_2O + 4\ Fe_3O_4$ (Magnetit) + $4\ H_2$

Während der Serpentinisierung können Elektronen auch auf CO_2 übertragen und organische Verbindungen sowie Methan generiert werden [13]. Dieser Prozess der organischen Synthese innerhalb der Erde ist sehr interessant, sowohl im Kontext der frühen Erde als auch im Kontext der frühen Evolution [14]. Die Gesamtreaktion, die Methan in hydrothermalen Systemen erzeugt (geochemische Methanogenese), ist dieselbe, die methanogene Archaeen heute in ihrer Energiegewinnungsreaktion (biologische Methanogenese) nutzen, lediglich die chemischen Intermediate sind verschieden: $4H_2 + CO_2 \rightarrow CH_4 + 2H_2O$. Die Reaktion läuft ab, weil sie Energie freisetzt, sowohl für die Erde als auch für das Leben. Die Ursprünglichkeit der Methanbildung passt gut zum Konzept der methanogenen Archaeen als eine Form der Ur-Mikroben. Das macht die Serpentinisierung – die spontane Reduktion von CO_2 zu Methan – von allen bislang bekannten, natürlich auftretenden geochemischen Redoxreaktionen zur einzigen, die Ähnlichkeiten mit den bioenergetischen Kernreaktionen moderner methanogener Archaeen hat. Lucas Gene weisen sehr deutlich auf eine frühe Abstammung der Methanogenen unter den Archaeen hin und dies passt zur biochemischen Ursprünglichkeit der Methanbildung, die Biologen eigentlich schon immer vermutet haben [15].

Es ist bemerkenswert, dass keine Gene für die Nutzung von Licht als Energiequelle in Lucas Genom auftauchten.

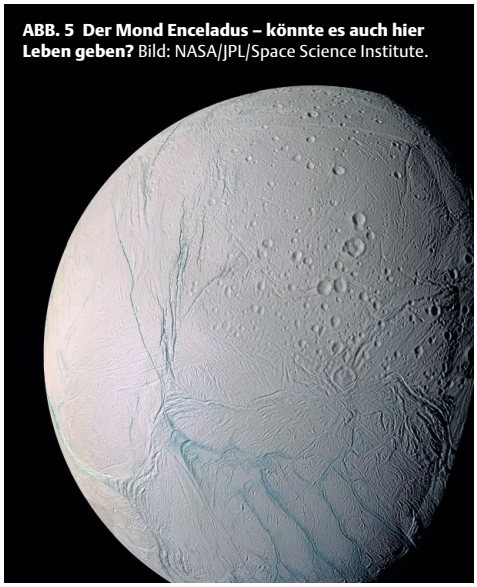


ABB. 5 Der Mond Enceladus – könnte es auch hier Leben geben? Bild: NASA/JPL/Space Science Institute.

Luca lebte von chemischer Energie. Alles was für Lucas Gedeihen erforderlich war, war in Gestein, Metallen, CO_2, Wasser und H_2 aus hydrothermaler Aktivität vorhanden. Weder Sonnenlicht noch ultraviolettes Licht waren erforderlich, um Luca zum Leben zu erwecken oder am Leben zu halten. Daraus folgt, dass auf der Suche nach weiterem Leben in unserem Sonnensystem Licht kein limitierender Faktor sein muss. Wenn man nach außerirdischem Leben sucht, geraten daher weit entfernte Monde wie Enceladus (Abbildung 5) in den Fokus des Interesses: Er umkreist Saturn und hat einen flüssigen Ozean aus Wasser, einen steinernen, metallreichen Kern und eine dramatische hydrothermale Aktivität (Serpentinisierung?) an seinem Südpol [16]. So eine chemische Umgebung könnte prinzipiell das Entstehen eines Organismus wie Luca, unserem eigenen gemeinsamen Vorfahren, unterstützen. Ob zukünftige Missionen zu Enceladus Hinweise für die Existenz von komplexen chemischen Reaktionen erbringen werden, oder gar molekulare Bausteine des Lebens ergeben könnten, bleibt abzuwarten. Wir sollten aber investieren, um Raumfahrzeuge zu bauen, die zum Saturn fliegen, einen genaueren Blick auf Enceladus werfen und diesen an die Erde übermitteln. Es ist durchaus möglich, dass zumindest eine interessante Gestein-Wasser-Kohlenstoff-Chemie dort in völliger Dunkelheit, unter dem Eis eines weit entfernten Mondes stattfindet.

Zusammenfassung

Der letzte universelle gemeinsame Vorfahre allen Lebens (Luca) wurde lange als der gemeinsame Vorfahre von Bakterien, Archaeen und Eukaryoten betrachtet. Neue Stammbäume des Lebens zeigen jedoch den Ursprung der Eukary-

oten innerhalb der Archaeen, somit wird Luca zum gemeinsamen Vorfahren der Bakterien und Archaeen. Informationen über Lucas Lebensraum und Lebensweise (Physiologie) gab es bisher nicht. Eine neue Arbeit konnte die mikrobielle Ökologie Lucas nun rekonstruieren. Nach phylogenetischen Kriterien gehen 355 Gene (Proteinfamilien) auf Luca zurück. Danach war Luca anaerob, CO_2- und N_2-fixierend, H_2-abhängig und thermophil. Lucas Proteine waren reichlich mit FeS-Zentren und radikalen Reaktionsmechanismen ausgestattet und erforderten Kofaktoren, bei denen Übergangsmetalle eine tragende Rolle spielen. Luca bewohnte eine geochemisch aktive Umgebung, die reich an H_2, CO_2 und Eisen war. Diese mikrobielle Ökologie ähnelt der heutiger acetogener Bakterien und methanogener Archaeen, den physiologisch ursprünglichsten Mikroben.

Summary

How, and where, did the first cells on Earth grow?

The last universal common ancestor of all cells (Luca) was long considered as the common ancestor of bacteria, archaea and eukaryotes. New trees of life have a host for the origin of mitochondria (of eukaryotes) branching within the archaea, making Luca the common ancestor of bacteria and archaea. New comparative genomic investigations have reconstructed Luca's microbial ecology. The 355 protein families that trace back to Luca by phylogenetic criteria describe Luca as anaerobic, CO_2- and N_2-fixing, H_2-dependent and thermophilic. Luca's biochemistry was replete with FeS clusters and radical reaction mechanisms, its cofactors reveal an essential role for transition metals in its metabolism. Luca lived in an anaerobic geochemical active environment rich in H_2, CO_2 and iron. This lifestyle is similar to modern acetogens (bacteria) and methanogens (archaea), the physiologically most ancient microbes.

Schlagworte

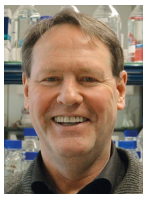
Letzter universeller gemeinsamer Vorfahre allen Lebens (Luca). Serpentinisierung, Hydrothermalquellen, Methanogene, Acetogene, Ursprung des Lebens

Literatur

[1] M. C. Weiss, F. L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, W. F. Martin, *The physiology and habitat of the last universal common ancestor*, Nat. Microbiol. 2016, 1, 16116.
[2] R. L. Charlebois, W. F. Doolittle, *Computing prokaryotic gene ubiquity: Rescuing the core from extinction*, Genome Res. 2004, 14, 2469–2477.
[3] P. Puigbò, Y. I. Wolf, E. V. Koonin, *Search for a ‚Tree of Life' in the thicket of the phylogenetic forest*, J. Biol. 2009, 8, 59.
[4] H. Stouthamer, *Energy yielding pathways in The bacteria* (Hrsg. I. C. Gunsalus, R. Y. Stanier), Vol. 4, Academic Press, New York, 1978, 389–462.
[5] F. Baymann, E. Lebrin, M. Brugna, B. Schoepp-Cothenet, M.-T. Giudici-Orticoni, W. Nitschke, *The redox protein construction kit: Pre-last universal common ancestor evolution of energy-conserving enzymes*, Phil. Trans. R. Soc. Lond. 2003, 358, 267–274.
[6] S. Nelson-Sathi, F. L. Sousa, M. Roettger, N. Lozada-Chávez, T. Thiergart, A. Janssen, D. Bryant, G. Landan, P. Schönheit, B. Siebers, J. O. McInerney, W. F. Martin, *Origins of major archaeal clades correspond to gene acquisitions from bacteria*, Nature 2015, 517, 77–80.
[7] W. Buckel, R. K. Thauer, *Energy conservation via electron bifurcating ferredoxin reduction and proton/Na⁺ translocating ferredoxin oxidation*, BBA Bioenergetics. 2013, 1827, 94–113.
[8] P. Schönheit, W. Buckel, W. F. Martin, *On the origin of heterotrophy*, Trends Microbiol. 2015, 24, 12–25.
[9] F. H. Chapelle, K. O'Neill, P. M. Bradley, B. A. Methé, S. A. Ciufo, L. L. Knobel, D. R. Lovley, *A hydrogen-based subsurface microbial community dominated by methanogens*, Nature 2002, 415, 312–315.
[10] M. A. Lever, V. B. Heuer, Y. Morono, N. Masui, F. Schmidt, M. J. Alperin, F. Inagaki, K.-U. Hinrichs, A. Teske, *Acetogenesis in deep subseafloor sediments of the Juan de Fuca Ridge Flank: A synthesis of geochemical, thermodynamic, and gene-based evidence*, Geomicrobiol. J. 2010, 27, 183–211.
[11] W. Bach, H. Paulick, C. J. Garrido, B. Ildefonse, W. P. Meurer, S. E. Humphris, *Unraveling the sequence of serpentinization reactions: Petrography, mineral chemistry, and petrophysics of serpentinites from MAR 15°N (ODP Leg 209, Site 1274)*, Geophys. Res. Lett. 2006, 33.
[12] M. J. Russell, A. J. Hall, W. Martin, *Serpentinization as a source of energy at the origin of life*, Geobiology 2010, 8, 355–371.
[13] G. Proskurowski, M. D. Lilley, J. S. Seewald, G. L. Früh-Green, E. J. Olson, J. E. Lupton, S. P. Sylva, D. S. Kelley, *Abiogenic hydrocarbon production at Lost City hydrothermal field*, Science 2008, 319, 604–607.
[14] T. M. McCollom, *Abiotic methane formation during experimental serpentinization of olivine*, Proc. Natl. Acad. Sci. USA 2016, 113, 13965–13970.
[15] K. Decker, K. Jungermann, R. K. Thauer, *Energy production in anaerobic organisms*, Angew. Chem. Int. Ed. 1970, 9, 138–158.
[16] H.-W. Hsu, F. Postberg, Y. Sekine, T. Shibuya, S. Kempf, M. Horányi, A. Juhász, N. Altobelli, K. Suzuki, Y. Masaki, T. Kuwatani, S. Tachibana, S. Sirono, G. Moragas-Klostermeyer, R. Srama, *Ongoing hydrothermal activities within Enceladus*, Nature 2015, 519, 207–210.

Korrespondenz:
Prof. Dr.
William F. Martin
Institute of Molecular Evolution
Heinrich-Heine-Universität Düsseldorf
Universitätsstraße 1
40225 Düsseldorf
E-Mail: bill@hhu.de

Die Autoren



William F. (Bill) Martin wurde 1957 in Bethesda, Maryland (USA) geboren. Er kam 1981 nach Deutschland und begann sein Studium der Biologie an der Technischen Universität Hannover, das er 1985 mit dem Diplom abschloss. Er promovierte 1988 am Max-Planck-Institut für Züchtungsforschung in Köln unter Prof. Heinz Saedler in Genetik. Anschließend ging er als Post-Doc an die Technische Universität Braunschweig und habilitierte dort im Jahr 1992 am Institut für Genetik, Prof. Rüdiger Cerff. 1999 wurde er zum Professor an die Heinrich-Heine-Universität Düsseldorf berufen. Seine Hauptinteressen sind die Gebiete der frühen Evolution und der Endosymbiose.

Verena Zimorski wurde 1978 in Oberhausen (Deutschland) geboren. Sie studierte Biologie (Diplom) an der Heinrich-Heine-Universität Düsseldorf, wo sie 2010 in Biologie promovierte. Seitdem führt sie ihre Forschung in Bill Martins Gruppe fort.

Madeline C. Weiß wurde 1989 in Geseke (Deutschland) geboren. Sie begann 2009 ihr Studium der Biologie an der Universität Kassel, das sie 2013 mit dem Bachelor abschloss. Sie setzte ihr Studium der Biologie an der Heinrich-Heine-Universität Düsseldorf fort und schloss es mit dem Master mit Schwerpunkt Bioinformatik im Jahre 2015 ab. Seitdem promoviert sie in Bill Martins Gruppe.

5.5 The last univesal common ancestor between ancient Earth chemistry and the onset of genetics

Madeline C. Weiss[1], Martina Preiner[1], Joana C. Xavier[1], Verena Zimorski[1] und William F. Martin[1,2]

Affiliations

1 Institut für Molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Deutschland.
2 Instituto de Tecnologia QuõÂmica e BioloÂgica, Universidade Nova de Lisboa, Oeiras, Portugal

Dieser Artikel wurde am 16. August 2018 in *PLoS Genetics* Ausgabe 14 veröffentlicht.

Beitrag von Madeline C. Weiß:

Ich habe für die Publikation die Abbildungen 2 und 4 erstellt und die Abbildung 3 formatiert. Des Weiteren habe ich die in Abschnitt „The physiology of LUCA“ erwähnte Analyse, aus Weiss *et al.* (2016b) durchgeführt und auch an den anderen Abschnitten habe ich mitgewirkt.

PLOS GENETICS

REVIEW

The last universal common ancestor between ancient Earth chemistry and the onset of genetics

Madeline C. Weiss[1], Martina Preiner[1], Joana C. Xavier[1], Verena Zimorski[1], William F. Martin[1,2]*

1 Institute of Molecular Evolution, Heinrich Heine University, Düsseldorf, Germany, **2** Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal

* bill@hhu.de

Abstract

All known life forms trace back to a last universal common ancestor (LUCA) that witnessed the onset of Darwinian evolution. One can ask questions about LUCA in various ways, the most common way being to look for traits that are common to all cells, like ribosomes or the genetic code. With the availability of genomes, we can, however, also ask what genes are ancient by virtue of their phylogeny rather than by virtue of being universal. That approach, undertaken recently, leads to a different view of LUCA than we have had in the past, one that fits well with the harsh geochemical setting of early Earth and resembles the biology of prokaryotes that today inhabit the Earth's crust.

Citation: Weiss MC, Preiner M, Xavier JC, Zimorski V, Martin WF (2018) The last universal common ancestor between ancient Earth chemistry and the onset of genetics. PLoS Genet 14(8): e1007518. https://doi.org/10.1371/journal.pgen.1007518

Editor: Mark Achtman, Warwick Medical School, UNITED KINGDOM

Published: August 16, 2018

Funding: This work was supported by grants from the ERC (666053 https://erc.europa.eu/), Volkswagen Foundation (93 046 https://www.volkswagenstiftung.de/en/funding.html), and GIF ([I-1321-203.13/2015] http://www.gif.org.il/Pages/default.aspx) to WFM. The funders had no role in the preparation of the article.

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

Introduction

The very earliest phases of life on Earth witnessed the origin of life and genetics from the elements. There was a time when there was no life on Earth, and there was a time when there were DNA-inheriting cells. The transitions are hard to imagine. Some dates and constraints on the order of events helps us to better grasp the problem. The Earth is 4.5 billion years (Ga) old [1]. By about 4.4 Ga, the moon-forming impact turned the Earth into a ball of boiling lava [1]. Magma oceans with temperatures over 2,000˚K forced all water from early accretion into the gas phase and converted all early accreted carbon to atmospheric carbon dioxide (CO_2) [1,2]. By 4.2 to 4.3 Ga, the Earth had cooled sufficiently enough that there was liquid water [3] —those first oceans were about twice as deep as today's [1,2]. Only later, hydrothermal convection currents started sequestering water to the primordial crust and mantle, which today bind one extra ocean volume [4,5]. The first signs of life appear as carbon isotope signatures in rocks 3.95 billion years of age [6]. Thus, somewhere on the ocean-covered early Earth and in a narrow window of time of only about 200 million years, the first cells came into existence. Because the genetic code [7] and amino acid chirality [8] are universal, all modern life forms ultimately trace back to that phase of evolution. That was the time during which the last universal common ancestor (LUCA) of all cells lived.

PLOS GENETICS

LUCA, the tree of life, and its roots

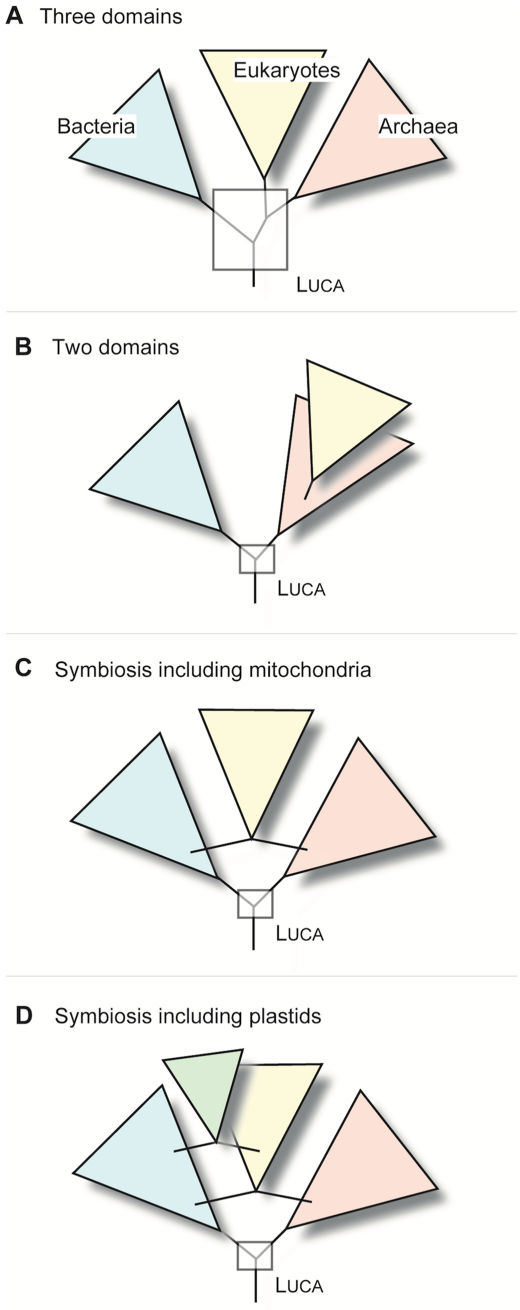
LUCA is a theoretical construct—it might or might not have been something we today would call an organism. It helps to bridge the conceptual gap between rocks and water on the early Earth and ideas about the nature of the first cells. Thoughts about LUCA span decades. Various ideas exist in the literature about how LUCA was physically organized and what properties it possessed. These ideas are traditionally linked to our ideas about the overall tree of life and where its root might lie [9–18]. Phylogenetic trees are, however, ephemeral. It is their inescapable fate to undergo change as new data and new methods of phylogenetic inference emerge. Accordingly, the tree of life has been undergoing a great deal of change of late.

The familiar three-domain tree of life presented by ribosomal RNA [19] depicted LUCA as the last common ancestor of archaea, bacteria, and eukaryotes (Fig 1A). In that framework, efforts to infer the gene content, hence the properties of LUCA, boiled down to identifying genes that were present in eukaryotes, archaea, and bacteria. When the first genomes came out, there were a great many such investigations [20–22], all of which were confronted with the same two recurrent and fundamental problems: 1) How are the three domains related to one another so that gene presence patterns would really trace genes to LUCA as opposed to another evolutionarily more derived branch? 2) Does presence of a gene in two domains (or three) indicate that it was present in the common ancestor of those domains, or could it have reached its current distribution via late invention in one domain and lateral gene transfer (LGT) from one domain to another?

The first problem (the root of the domains) has been the subject of much recent work. Phylogenetic advances and new metagenomic data are changing the three-domain tree [19] into a two-domain tree [24,25]. This is partially a development around phylogenetic methods [24,26–28] but also entails new archaeal lineages that are now being assembled from metagenomic data and that appear to be more closely related to the host that acquired the mitochondrion than any other archaea known so far [29,30]. The two-domain tree showing an "archaeal origin of eukaryotes" [24,28] (Fig 1B) only tells part of the story, though, because eukaryote genomes harbor more bacterial genes than they do archaeal genes by a factor of about 3:1 [31–33], and those bacterial genes furthermore trace to the eukaryote common ancestor [23]. Eukaryotes are not just big, complex archaea; genomically and at the cellular level, they are true chimeras in that they possess archaeal ribosomes in the cytosol and bacterial ribosomes in mitochondria (Fig 1C) [34]. That polarizes cellular evolution in the right direction (there were once debates about eukaryotes being ancestral [10,13,14,22], as discussed elsewhere [35–37]) and identifies eukaryotes as latecomers in evolution, descendants of prokaryotes [38].

Current versions of the two-domain tree focus on the phylogeny of a handful of about 30 genes, mostly for ribosomal proteins (Box 1) but also on sequences from metagenomic samples. The metagenomic studies [29,30] have generated debate. Metagenomic data can bring forth alignments of genes that were sequenced accurately but have the wrong taxonomic label. For example, Da Cunha and colleagues [39] reported that published trees [29] hinge upon a strong signal stemming from one gene out of 30 and that the gene in question (an elongation factor [EF2]) might not be archaeal but eukaryotic instead. Spang and colleagues [40] defended their tree, eliciting more debate [41]. Errors can also occur in the assembly pipeline [42] en route to alignments [43], independent of contamination. Notwithstanding current debate about metagenomics-based trees of life [24,39,40,42,43], we should recall that rRNA itself produces the two-domain tree when various tree construction parameters are employed [24,26,27]. Both data and methods bear upon efforts to construct trees of life. It remains possible that some aspects of domain relationships might never be resolved to everyone's satisfaction—even the endosymbiotic origin of mitochondria is still debated [37]. But the bacterial

PLOS GENETICS



PLOS GENETICS

Fig 1. Different views on domain relationships in the tree of life. (A) The three-domain tree: based on rRNA phylogeny, the three domains were of equal rank. (B) The two-domain tree: modern trees show eukaryote cytosolic ribosomes branching within the diversity of archaeal ribosomes. (C) As eukaryotes are not just grownup archaea, the eukaryote ancestor possessed mitochondria. If mitochondrial-derived genes are taken into account, the tree is no longer a bifurcating graph. (D) If plastids are included, the tree becomes even less tree-like because the photosynthetic lineages of eukaryotes also acquired many genes from the plastid ancestor [23].

https://doi.org/10.1371/journal.pgen.1007518.g001

origin of mitochondria and their presence in the eukaryote common ancestor [44–47], together with the tendency of eukaryotes to branch within archaeal lineages as archaeal lineage sampling [29,30,48] and phylogenetic methods [24,26,27,32] improve, indicates that eukaryotes arose from prokaryotes and that genes that trace to the common ancestor of archaea and bacteria trace to LUCA.

The second problem (how much LGT has there been between domains) that has impaired progress on LUCA has arguably been more difficult to resolve than the rooting issue. If a given gene is present in bacteria and archaea, was it present in LUCA, or could it have been transferred between domains via LGT? As one important example, early studies pondered the presence of bacterial type oxygen (O_2)-consuming respiratory chains in archaea [21]. Does that

Box 1. The tree of 1% and the tree of everything else

A traditional approach to LUCA has been to simply look for the genes that are present in all genomes. That is easy enough, but the results are sobering. What one finds is a collection of about 30 genes, mostly for ribosomal proteins, telling us that LUCA had a ribosome and had the genetic code, which we already knew [63–65]. That collection of about 30 genes has been in use for about 20 years as concatenated alignments to make trees of lineages based on larger amounts of data than rRNA sequences have to offer [66]. The genes that are present in all lineages (or nearly all) inform us about how LUCA translated mRNA into protein, but they do not tell us about how or where LUCA lived. That information concerns ecophysiology, and physiological traits are not universally conserved—they are what makes microbes different from one another. One can relax the criteria of universal presence a bit and allow for some gene loss in some lineages, in which case, one finds about 100 proteins that are nearly universal [67]. If one puts no size constraints on LUCA's genome and allows loss freely, then all genes present in at least one archaeon and one bacterium trace to LUCA, making it the most versatile organism that ever lived [51]. New insights about microbial phylogeny are emerging from concatenated alignments [24,29,30,42,48,68]. But one has to take care not to get genes from different lineages mixed up, which can be difficult when metagenomes are involved [39,43]. Furthermore, data concatenation has its own pitfalls [66,69,70]. Most modern concatenation studies [29,30,48] employ site-filtering methods in an attempt to remove "noise," but even sites that look "noise free" can still contain bias and conflicting data [63]. Another problem is that popular methods of phylogenetic inference produce inflated confidence intervals on phylogenies and branches [71]. Trees of ca. 30 concatenated proteins are no more immune to phylogenetic error than rRNA is and are prone to additional kinds of error [72]. As it relates to LUCA, regardless of the backbone tree, we still need to know what all proteins say individually about their own phylogenies.

mean that archaea are ancestrally O_2 consumers? As O_2 is the product of cyanobacterial photosynthesis [49] if we presume archaeal O_2 respiration to be an ancestral trait of archaea, it means that archaea arose after cyanobacteria, which are only about 2.5 billion years old and gave rise to plastids (Fig 1D) only about 1.5 billion years ago [50]. If ancestral archaea were oxygen respirers, and ancestral bacteria were too, suddenly neither the two-domain tree nor the three-domain tree (Fig 1) make sense because everything is upside down and rooted in cyanobacteria. Similar issues are encountered for many genes and traits [51]. Lateral gene transfer among prokaryotic domains helps to resolve such problems because it decouples physiology (ecological trait evolution) from phylogeny (ribosomal lineage evolution) [52], but it also makes genes more difficult to trace to LUCA.

Has lateral gene transfer obscured all records?

That takes us to the other extreme. If all genes have been subjected to LGT, as some early claims had it [53], then LUCA would be altogether unknowable from the standpoint of genomes. Early archaeal genomes did indeed uncover abundant transdomain LGT [54], and many bacteria to archaea transfers can be correlated to changes in physiology [55], including the transfer of O_2-consuming respiratory chains [55–58]. For reconstructing LUCA, the issue boils down to determining i) which genes are present in both archaea and bacteria, ii) which of those are present in both prokaryotic domains because of LGT between archaea and bacteria, and iii) which are present because of vertical inheritance from LUCA. For that, there are currently two methodological approaches. One involves making a backbone reference tree from universally conserved genes that are present in each genome—the tree of 1% [59] (see Box 1)—plotting all gene distributions on the tips of that tree, and then estimating which genes trace to LUCA on the basis of various assumed gain and loss parameters [60–62]. If we permit loss freely, many genes will trace back to LUCA; if we assume many gains, LUCA will have few genes [61]. Constraining ancestral genome sizes helps constrain estimates of which genes trace to LUCA [61] but only if we assume that the tree of each gene is compatible with the reference tree, which is a very severe assumption and unlikely to be true. Each gene has its own individual history (Box 1).

Each gene records its own evolutionary history

If any protein-coding genes have been vertically inherited from LUCA, their trees should reflect that. To find such trees, one has to make all trees for all proteins, meaning one has to make clusters for all protein-coding genes from large numbers (thousands) of sequenced genomes. Clusters correspond to "natural" protein families of shared amino acid sequence similarity. Given modern computers, making alignments for all such clusters and making maximum likelihood trees for all such alignments is a tractable undertaking. Because LGT among prokaryotes is a real and pervasive process shaping prokaryote genome evolution [55,58,73–77], one has to treat each gene as a marker of its own evolution, not as a proxy for other genes or as a function that is subordinate to ribosomal phylogeny.

Genes that are present in several bacterial lineages and one archaeal lineage (or vice versa) might have been present in LUCA, but they might also have been the result of LGT [55,56,58]. An example illustrates how each gene tree can discriminate between vertical inheritance from LUCA and interdomain LGT. A recent study investigated the 6.1 million proteins encoded in 1,981 prokaryotic genomes (1,847 bacteria and 134 archaea) [78]. The proteins were clustered using the standard Markov Cluster (MCL) method [79]. The first step in that procedure is a matrix containing 18.5 trillion elements ($(n^2-n)/2$), each element corresponding to a pairwise amino acid sequence comparison. The clustering of such a matrix requires substantial

PLOS GENETICS

computational power and is aided by the availability of several terabytes of memory in a single machine. The MCL algorithm samples the distribution of values in the matrix and then starts removing the weak edges, with the value of "weak" being specified by the user. Two kinds of thresholds are typically used in MCL clustering: BLAST e-values and amino acid identity in pairwise alignments.

When the goal of clustering is to make alignments and trees, our group has found that a clustering threshold of 25% amino acid identity is a good rule of thumb. At lower thresholds, amino acid identity starts to approach random values and generates random errors in alignments [80], carrying over as erroneous topologies in trees [81]. That is why Russell F. Doolittle coined the term "twilight zone" for amino acid identity at or below the 20% range [82,83]. Of course, many proteins or domains that clearly share a common ancestry by the measure of related crystal structures do not share more than a random amino acid sequence identity [84]. Such ancient folds will fall into separate clusters at the 25% identity threshold and might thus generate false negatives when it comes to presence in LUCA (but see next section).

From thousands of clusters and trees, a handful remain

Using the 25% identity threshold, the 6.1 million prokaryotic proteins sampled fall into 286,514 clusters of at least two sequences, and 11,093 of those clusters include sequences found in both archaea and bacteria [78]. Many of those clusters involve oxygen-dependent respiratory chains. Did LUCA have 11,000 genes in its genome and breathe oxygen? That is, was LUCA (and hence archaea) descended from cyanobacteria? Neither prospect seems likely enough to warrant further discussion [85]. Knowing that transdomain LGT is prevalent [54–56] and that thousands of typically bacterial genes are shared with only one archaeal group [58], Weiss and colleagues [78] reasoned that a simple way to exclude some LGTs would be to set the minimal phylogenetic criteria that 1) a gene needs to be present in bacteria and archaea, 2) it needs to be present in at least two phylum-level clades, and 3) the tree needs to preserve domain monophyly (Fig 2). Genes that do not fulfil criterion 1 are not candidates for LUCA anyway. The two-phylum-plus-monophyly criteria 2 and 3 make it less likely but not impossible that such a gene attained that distribution via LGT. How so? Criteria 2 and 3 would require one transdomain transfer followed by intradomain transfers to different phyla, while allowing no subsequent, independent transdomain transfers. The last condition is the restrictive one.

Of the 11,093 clusters that harbored sequences in bacteria and archaea, only 355 (3%) passed the simple LGT filter [78]. Put another way, 97% of the sequences present in bacteria and archaea apparently underwent some transdomain LGT, underscoring the degree to which transdomain LGT has influenced gene history since LUCA and underscoring the need to employ phylogenetic filters in search of genes that trace to LUCA [21,51]. The 97% LGT value is important with regard to the 25% clustering threshold and possible false negatives; 97% of all false negatives founded in low-sequence conservation would still not trace to LUCA because of transdomain LGTs. But transdomain LGT has apparently not erased all signals, as 355 genes passed the LGT test, and those genes tell us things about LUCA that we did not know before.

The physiology of LUCA

Most earlier depictions of LUCA focused on what it was like [16]; for example, whether it was like RNA [86], like a virus [87], whether it was like prokaryotes in terms of its genetic code [88], or like eukaryotes in terms of its cellular organization [22]. But traditional approaches lacked information about how and from what LUCA lived [16]. Our phylogenetic approach to LUCA [78] uncovered information about what LUCA was doing: its physiology, its ecology, and its environment. The genes for those physiological traits are not necessarily widespread

PLOS GENETICS

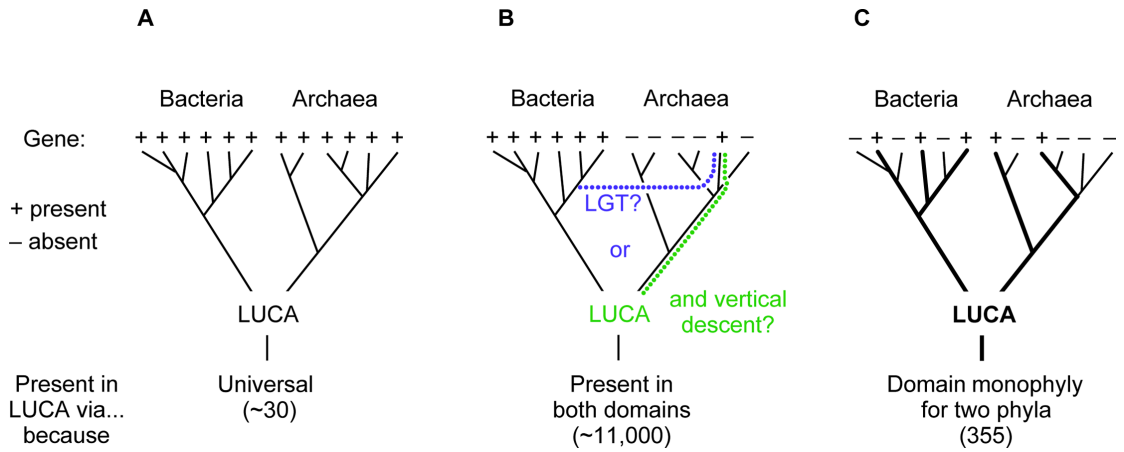


Fig 2. Three ways to infer genes present in LUCA. The gene presence is indicated with a plus sign, absence with a minus sign. a) Genes found universally in both domains, regardless of their tree, trace to LUCA. About 30 fulfil this criterion. b) Another way to trace genes to LUCA is to say that any gene found in both archaea and bacteria was present in LUCA. However, thousands of these genes will have been transferred between bacteria and archaea by LGT so were not necessarily present in LUCA. c) Genes present in only one bacterial or archaeal phylum could easily be the result of LGT and are removed. But presence in two phyla per domain while preserving domain monophyly yields good candidates to have been present in LUCA. Such phylogenies would only result from LGT under very specific and restrictive conditions. They require exactly one transdomain transfer followed by either i) one additional transdomain LGT from the same donor lineage to a different recipient phylum or ii) retention during phylum divergence in the recipient domain, plus—in addition to either criteria i) or ii)—an additional, more subtle but highly restrictive criterion: No further transdomain LGTs occurred during all of evolution. Subsequent transdomain LGT would violate domain monophyly for the gene. Indeed, transdomain LGT is common, and 97% of the trees examined by Weiss and colleagues [78] did not exclude transdomain LGT (remaining 3%, 355 trees, provided in S1 Appendix). LGT, lateral gene transfer; LUCA, last universal common ancestor.

https://doi.org/10.1371/journal.pgen.1007518.g002

among modern genomes, but the filtering criteria by Weiss and colleagues [78] only require that these genes are ancient. What Weiss and colleagues [78] found is schematically summarized in Fig 3.

LUCA was an anaerobe, as long predicted by microbiologists [89]. Its metabolism was replete with O_2-sensitive enzymes. These include proteins rich in O_2-sensitive iron–sulfur (FeS) clusters and enzymes that entail the generation of radicals (unpaired electrons) via S-adenosyl methionine (SAM) in their reaction mechanisms. That fits well with the 50-year-old [90] but still modern view that FeS clusters represent very ancient cofactors in metabolism [91–93]. It also fits with newer insights about the ancient and spontaneous (nonenzymatic) chemistry underlying SAM synthesis [94].

LUCA lived from gasses. For carbon assimilation, LUCA used the simplest and most ancient of the six known pathways of CO_2 fixation, called the acetyl–CoA (or Wood–Ljungdahl) pathway [95–97], which is increasingly central for our concepts on early evolution because of its chemical simplicity [97,98] and exergonic nature [99–101]. In the acetyl–CoA pathway, CO_2 is reduced with hydrogen (H_2) to a methyl group and CO. The methyl group is synthesized by the methyl branch of the pathway, which employs different one-carbon (C1) carriers in bacteria (tetrahydrofolate) and archaea (tetrahydromethanopterin), cofactors that are synthesized by unrelated biosynthetic pathways [96]. Carbon monoxide (CO) is synthesized by carbon monoxide dehydrogenase (CODH), the archaeal and bacterial versions of which are distinct but related [96]. The methyl and carbonyl moieties are condensed to an enzyme-bound acetyl group that is removed from a metal cluster in acetyl–CoA synthase (ACS) as an energy rich thioester. Thioesters harbor chemically reactive bonds [102] that play

PLOS GENETICS

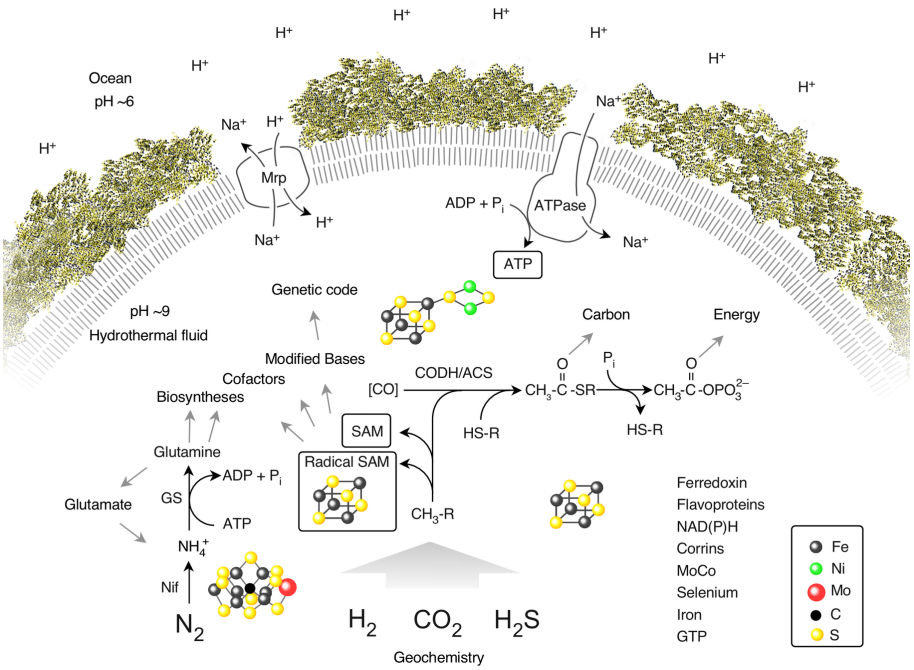


Fig 3. The physiology of LUCA. Summary of the main interactions of LUCA with its environment, reprinted with permission from [78] (supporting trees in S1 Appendix). Components listed at the lower right are present in LUCA. The figure does not make a statement regarding the source of CO in primordial metabolism, symbolized by [CO]. LUCA indisputably possessed genes because it had a genetic code. Transition metal clusters are symbolized. CH_3-R, methyl groups; CODH/ACS, carbon monoxide dehydrogenase/acetyl–CoA synthase; GS, glutamine synthetase; HS-R, organic thiols; LUCA, last universal common ancestor; Mrp, MrP type Na^+/H^+ antiporter; Nif, nitrogenase; SAM, S-adenosyl methionine.

https://doi.org/10.1371/journal.pgen.1007518.g003

a crucial role in energy metabolism [101] and in metabolism in general, both modern and ancient [101,103,104]. Although CODH/ACS clearly does trace to LUCA [78,96], this is not true for the methyl synthesis branch, which consists of unrelated enzymes in bacteria and archaea [78,96].

A recent report [105] argued that the presence of CODH in LUCA did not exclude a heterotrophic lifestyle for LUCA. This argument is problematic because no single enzyme defines a trophic lifestyle. Even Rubisco (D-ribulose-1, 5-bisphosphate carboxylase/oxygenase), the classical Calvin cycle enzyme, is not a marker for autotrophy because Rubisco also functions in a simpler heterotrophic pathway of RNA fermentation [106–108] that is common among archaea and bacteria in marine sediment environments [109]. Moreover, all heterotrophs are derived from autotrophs due to the former requiring the latter as a source of chemically defined growth substrates. The reason is that CO_2 constituted the main carbon source on Earth after the moon-forming impact [1,110], while carbon delivered from space was either too reduced to be fermented (polyaromatic hydrocarbons), too heterogeneous in structure to support microbial growth, or both [108]. Autotrophs with CODH can obtain ATP from CO_2 reduction with H_2 [98,101,110]. Autotrophs without CODH cannot. If we base inferences

about LUCA's lifestyle on broad criteria rather than single genes [105], LUCA was an autotroph [78,108].

Life is about harnessing energy [44]. Thioesters are chemically reactive—they forge direct links between carbon metabolism and energy metabolism (ATP synthesis) as they give rise to acetyl phosphate, the possible precursor of ATP in evolution as a currency of high-energy bonds [111]. Relics of ATP synthesis via acetyl phosphate were found in LUCA's genes [78], as were subunits of the rotor–stator ATP synthase itself. The ATP synthase might appear to present a paradox because no proteins of the proton-pumping machinery that cells use to generate the ion gradient that drives the ATP synthase traced to LUCA [78]. Yet some theories have it that the first cells arose at alkaline hydrothermal vents [91,96,111], meaning that the inside of the vent is more alkaline than the ocean outside. Such naturally existing pH gradients could have been harnessed by LUCA to synthesize ATP (Fig 3). Ancestral ATPases might have harnessed either proton gradients or sodium gradients generated by proton/sodium (H^+/Na^+) dependent antiporters [112], or they might have even been promiscuous for both kinds of ions, similar to the ATPase of modern microbes that live near the thermodynamic limits of life [113].

LUCA's environment was rich in sulfur; thioesters, SAM, proteins rich in FeS and iron–nickel–sulfur (FeNiS) clusters, sulfurtransferases, and thioredoxins were part of its repertoire, as were hydrogenases that could channel electrons from environmental H_2 to reduced ferredoxin, which is the main currency of reducing power (electrons) in anaerobes [114]. A recent report provided phylogenetic evidence that archaea are ancestrally H_2-dependent methanogens [62], compatible with an autotrophic, H_2-dependent lifestyle of LUCA.

LUCA had a reverse gyrase, an enzyme typical of thermophiles, suggesting that LUCA liked it hot. But independent of the reverse gyrase, simple chemical kinetics provide strong evidence in favor of a thermophilic origin for the first cells [115,116]. The reason is that only uncatalysed or inorganically catalysed reactions existed before there were enzymes. Their rates of reaction were lower than the enzymatically catalyzed reactions. Between 0˚C and 120˚C (the biologically relevant temperature range), organic chemical reaction rates generally increase with temperature [115,116]. Before there were enzymes, high-temperature environments were more conducive to organic chemical reactions than low-temperature environments [115,116]. Taken together, LUCA's requirement for gasses (CO_2, H_2, CO, nitrogen [N_2]), the prevalence of sulfide, its affinity to high temperature and metals, plus an ability to use but not generate ion gradients all point to the same environment: alkaline hydrothermal vents.

In addition to shedding light on physiology, the 355 trees that showed domain monophyly (S1 Appendix) [78] also have another interesting property: they are reciprocally rooted. That is, the bacteria are rooted in an archaeal outgroup and vice versa. Genes present in LUCA contain information about their lineages and about the groups of bacteria and archaea that branched most deeply in each domain. In both cases, the answer was clostridia (bacteria) and methanogens (archaea). Those are strictly anaerobic prokaryotes that use the acetyl–CoA pathway; live from CO_2, H_2, and CO; fix N_2, and today inhabit hydrothermal environments in the Earth's crust [117–119].

The onset of genetics

Though the organization of inanimate matter into living cells with genetics can be charted in mathematical terms [120,121], the biochemical details remain elusive. For example, it is controversial whether LUCA had DNA or not [87]. Several DNA-binding proteins trace to LUCA [78], so it would appear that LUCA possessed DNA, but it is unresolved whether LUCA could

PLOS GENETICS

actually replicate DNA. For LUCA, DNA might just have been a chemically stable repository for RNA-based replication [122].

A novel and interesting aspect of LUCA's biology concerns modified bases and the genetic code. Transfer RNA requires modified bases for proper interaction with mRNA (codon–anticodon wobble base pairing) and with rRNA in the ribosome during translation. That is, modified bases are part of the universal genetic code (Fig 4), which was present in LUCA. Many RNA-modifying enzymes trace to LUCA, particularly the enzymes that modify tRNA. Several of those enzymes are methyltransferases (many SAM dependent), and they remind us that, before the genetic code arose, the four main RNA bases could hardly have been in great supply in pure form because there were no genes or enzymes, only chemical reactions [123]. Spontaneous synthesis of bases in a real early Earth environment like a hydrothermal vent, an environment that lacks the control of a modern laboratory [124], is not likely to generate the four main bases in pure form. Many side products will accumulate, including chemically modified bases [111]. Chemically modified bases from living cells have been reported since the 1970s by pioneering RNA chemists such as Mathias Sprinzl [125] and Henri Grosjean [126]. There are 28 modified bases, mainly occurring in tRNA, that are shared by bacteria and archaea [127]. The modifications are chemically simple, such as the introduction of methyl groups or sulfur and occasionally of acetyl groups and the like (Fig 4).

Chemical modifications in the tRNA anticodon are essential for codon–anticodon interactions to work [128,129]. Modifications of the rRNA are concentrated around the peptidyl transferase site and are also essential for tRNA ribosome interactions [130]. It is possible that the genetic code itself arose in the same chemically reactive environment where LUCA arose and that modified bases in tRNA carry the chemical imprint of that environment [78]. That would forge a link between the early Earth and genetics as we know it. New laboratory syntheses of RNA molecules in the origin of life context now also include investigations of modified bases [131], as it is becoming increasingly clear that these are crucial components at the very earliest phases of molecular and biological evolution.

Moving forward

Investigations of LUCA based on phylogenies of all genes pose new opportunities and new challenges. As environmental sequencing and metagenomics progresses, the number of microbial sequences and new lineages is exploding [48,109]. How will that aspect of metagenomics affect investigations of LUCA? If the criteria for gene age are phylogenetic (prokaryote domain monophyly, presence in at least two bacterial and archaeal "phyla"), then the correct taxonomic assignment of each sequence is very important. A problematic aspect of metagenomic data is that some data handling steps can assign incorrect higher taxon labels to genes [39,41,43], which in turn can falsify phylogenetic relationships. Analyses of cultured microbes or complete genome sequences limit the available sample size but deliver reliable taxon labels, at least at the level of archaea versus bacteria. Clearly, there are trade-offs.

At first sight, LUCA's genome appears doomed to shrinkage. As the sample of complete genomes grows, the list of 355 genes that trace to LUCA by domain monophyly criteria [78] will shrink because each new genome offers new opportunities to uncover recent LGT events for the 355 genes. Recalling that only 3% of the 11,093 clusters investigated [78] appeared free of transdomain LGT, it is evident that the inclusion of new genomes will eventually cause the number 355 to asymptotically approach zero, unless some genes never undergo transdomain LGT, which seems unlikely. What to do? Filtering out recent LGT events would help save LUCA's genome from shrinking to zero. For example, the tree for gene X might violate domain monophyly by one LGT event. If the LGT was recent, affecting members of only one

PLOS GENETICS

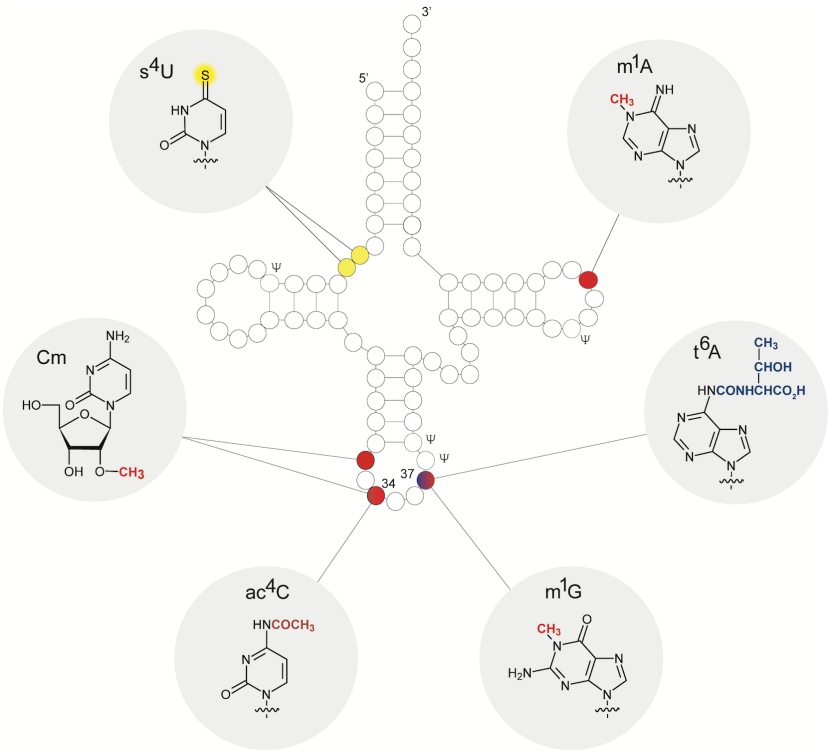


Fig 4. Modified tRNA and nucleoside structures (adapted from [78]). Cloverleaf secondary structure representation of tRNA showing post-transcriptional nucleoside modifications that are conserved among bacteria and archaea in both identity and position. The structures of respective conserved modified nucleosides are highlighted in grey. Methyl and acetyl groups are shown in red and dark red, respectively; sulfur in yellow; and the threonylcarbamoyl group in blue.

https://doi.org/10.1371/journal.pgen.1007518.g004

recipient genus or family, it would hardly affect inferences about LUCA, adding gene X to LUCA's list. To identify recent LGTs in prokaryote phylogeny, standard criteria like incomplete amelioration [132], anomalously high-sequence identity [133], or presence in the auxiliary genome [134] will be useful, as will new methods that root unrooted trees [135]. Identifying recent LGTs should allow us to trace more genes to LUCA.

There is also the issue of clustering thresholds to consider, as discussed above. Stringent thresholds produce many small clusters and more relaxed thresholds produce a smaller number of very large clusters [136]. One can argue that large clusters (low stringency) allow one to look further back into time, but they also can generate clusters whose origins trace to duplications in LUCA, in which domain monophyly is violated but not because of LGT. Another factor concerns gene fusions. Genes tend to undergo fusion and fission during evolution [137,138]. In clustering procedures, gene fusions tend to slightly reduce the number of clusters because when they occur, they can bring two fused genes into one alignment, and the weaker phylogenetic signal in the fusion is obscured [23]. Methods to detect fusions exist [139,140]. By detecting gene fusions and dissecting them into their component parts, it might be possible to increase the number of trees that trace to LUCA by phylogenetic criteria.

PLOS GENETICS

Investigations into early evolution always elicit protest. For example, there were criticisms [141] of the term "progenote," which Woese and Fox [142] introduced to designate a state of organization below that of a free-living cell [143,144], as shown in Fig 3. In addition, multiple LGTs can, in principle, generate false positives by mimicking vertical inheritance from LUCA [78], but very specific conditions have to be fulfilled (Fig 1C). The challenge is to distill a chronicle of microbial evolution that takes all genes and LGT [145] into account and that conveys information about physiology [146], the energy-releasing reactions that power microbial evolution.

Conclusions

More clues about LUCA's lifestyle are emerging. Investigations of modern biochemical pathways hone in on the same kinds of reactions as the phylogenetic approach [103]. Similarly, laboratory experiments also demonstrate the spontaneous synthesis of end products and intermediates of the acetyl–CoA pathway, the mainstay of LUCA's physiology; new findings show that formate, methanol, acetyl moieties, and even pyruvate arise spontaneously at high yields and at temperatures conducive to life (30˚C–100˚C) from CO_2, native metals, and water [98,147]. Those conditions are virtually impossible to underbid in terms of chemical simplicity [98], yet they bring forth the core of LUCA's carbon and energy metabolism [78,96,97,101, 103] overnight. Did the origin of genetics hinge upon hydrothermal chemical conditions that gave rise to the first biochemical pathways that in turn gave rise to the first cells? Genes that trace to LUCA [78], ancient biochemical pathways [103], and aqueous reactions of CO_2 with iron and water [98,110] all seem to converge on similar sets of simple, exergonic chemical reactions as those that occur spontaneously at hydrothermal vents [148]. From the standpoint of genes, physiology, laboratory chemistry, and geochemistry, it is beginning to look like LUCA was rooted in rocks.

Supporting information

S1 Appendix. ML trees for the 355 protein families that trace to LUCA by phylogenetic criteria. The trees are for the 355 clusters that, after alignment and tree construction, generated ML trees that preserve domain monophyly while also having homologues in $\geq$2 archaeal and $\geq$2 bacterial lineages. These 355 proteins trace to LUCA by those phylogenetic criteria [78]. LUCA, last universal common ancestor; ML, maximum likelihood.
(ZIP)

Acknowledgments

We thank Filipa Sousa (Universität Wien), Harun Tüysüz (Max-Planck-Institut für Kohlenforschung, Mülheim an der Ruhr), and Joseph Moran (University of Strasbourg) for discussions.

References

1. Zahnle K, Arndt N, Cockell C, Halliday A, Nisbet E, Selsis F, et al. Emergence of a habitable planet. Space Sci Rev. 2007; 129: 35–78.
2. Arndt N, Nisbet E. Processes on the young Earth and the habitats of early life. Annu Rev Earth Planet Sci. 2012; 40: 521–549.
3. Mojzsis SJ, Harrison TM, Pidgeon RT. Oxygen-isotope evidence from ancient zircons for liquid water at the Earth's surface 4,300 Myr ago. Nature. 2001; 409: 178–181. https://doi.org/10.1038/35051557 PMID: 11196638

4. Hirschmann MM. Water, melting, and the deep earth H_2O cycle. Annu Rev Earth Planet Sci. 2006; 34: 629–653.
5. Fei H, Yamazaki D, Sakurai M, Miyajima N, Ohfuji H, Katsura T, et al. A nearly water-saturated mantle transition zone inferred from mineral viscosity. Sci Adv. 2017; 3: e1603024. https://doi.org/10.1126/sciadv.1603024 PMID: 28630912
6. Tashiro T, Ishida A, Hori M, Igisu M, Koike M, Méjean P, et al. Early trace of life from 3.95 Ga sedimentary rocks in Labrador, Canada. Nature. 2017; 549: 516–518. https://doi.org/10.1038/nature24019 PMID: 28959955
7. Kubyshkin V, Budisa N. Synthetic alienation of microbial organisms by using genetic code engineering: Why and how? Biotechnol J. 2017; 12: 1600097–1600110.
8. Haldane JBS. Origin of Life. The Rationalist Annual. 1929; 148: 3–10.
9. Crick FHC. The recent excitement in the coding problem. Prog Nucleic Acid Res Mol Biol. 1963; 1: 163–217.
10. Doolittle WF, Brown JR. Tempo, mode, the progenote, and the universal root. Proc Natl Acad Sci USA. 1994; 91: 6721–6728. PMID: 8041689
11. Pace NR. Origin of life-facing up to the physical setting. Cell. 1991; 65: 531–533. PMID: 1709590
12. Woese C. The universal ancestor. Proc Natl Acad Sci USA. 1998; 95: 6854–6859. PMID: 9618502
13. Forterre P, Philippe H. Where is the root of the universal tree of life? Bioessays. 1999; 21: 871–879. https://doi.org/10.1002/(SICI)1521-1878(199910)21:10<871::AID-BIES10>3.0.CO;2-Q PMID: 10497338
14. Penny D, Poole A. The nature of the last universal common ancestor. Curr Opin Genet Dev. 1999; 9: 672–677. PMID: 10607605
15. Koonin EV. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol. 2003; 1: 127–136. https://doi.org/10.1038/nrmicro751 PMID: 15035042
16. Becerra A, Delaye L, Islas S, Lazcano A. The very early stages of biological evolution and the nature of the last common ancestor of the three major cell domains. Annu Rev Ecol Evol Syst. 2007; 38: 361–379.
17. Di Giulio M. The last universal common ancestor (LUCA) and the ancestors of archaea and bacteria were progenotes. J Mol Evol. 2011; 72: 119–126. https://doi.org/10.1007/s00239-010-9407-2 PMID: 21079939
18. Fox GE. Origin and evolution of the ribosome. Cold Spring Harb Perspect Biol. 2010; 2: a003483. https://doi.org/10.1101/cshperspect.a003483 PMID: 20534711
19. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA. 1990; 87: 4576–4579. PMID: 2112744
20. Kyprides N, Overbeek R, Ouzounis C. Universal protein families and the functional content of the Last Universal Common Ancestor. J Mol Evol. 1999; 49: 413–423. PMID: 10485999
21. Castresana J, Moreira D. Respiratory chain in the last common ancestor of living organisms. J Mol Evol. 1999; 49: 453–460. PMID: 10486003
22. Doolittle WF. The nature of the universal ancestor and the evolution of the proteome. Curr Opin Struct Biol. 2000; 10: 355–358. PMID: 10851188
23. Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, et al. Endosymbiotic origin and differential loss of eukaryotic genes. Nature. 2015; 524: 427–432. https://doi.org/10.1038/nature14963 PMID: 26287458
24. Williams TA, Foster PG, Cox CJ, Embley TM. An archaeal origin of eukaryotes supports only two primary domains of life. Nature. 2013; 504: 231–236. https://doi.org/10.1038/nature12779 PMID: 24336283
25. McInerney J, Pisani D, O'Connell MJ. The ring of life hypothesis for eukaryote origin is supported by multiple kinds of data. Philos Trans R Soc Lond B Biol Sci. 2015; 370: 20140323. https://doi.org/10.1098/rstb.2014.0323 PMID: 26323755
26. Lake JA. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature.1988; 331: 184–186. https://doi.org/10.1038/331184a0 PMID: 3340165
27. Galtier N, Tourasse N, Gouy M. A nonhyperthermophilic common ancestor to extant life forms. Science. 1999; 283: 220–221. PMID: 9880254
28. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci USA. 2008; 105: 20356–20361. https://doi.org/10.1073/pnas.0810647105 PMID: 19073919

29. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015; 521: 173–179. https://doi.org/10.1038/nature14447 PMID: 25945739

30. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature. 2017; 541: 353–358. https://doi.org/10.1038/nature21031 PMID: 28077874

31. Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, et al. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol. 2004; 21: 1643–1660. https://doi.org/10.1093/molbev/msh160 PMID: 15155797

32. Pisani D, Cotton JA, McInerney JO. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol. 2007; 24: 1752–1760. https://doi.org/10.1093/molbev/msm095 PMID: 17504772

33. Cotton JA, McInerney JO. Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. Proc Natl Acad Sci USA. 2010; 107: 17252–17255. https://doi.org/10.1073/pnas.1000265107 PMID: 20852068

34. Martin WF, Tielens AGM, Mentel M, Garg SG, Gould SB. The physiology of phagocytosis in the context of mitochondrial origin. Microbiol Mol Biol Rev. 2017; 81: e00008–17. https://doi.org/10.1128/MMBR.00008-17 PMID: 28615286

35. Forterre P.The common ancestor of Archaea and Eukarya was not an archaeon. Archaea 2013;372396. https://doi.org/10.1155/2013/372396 PMID: 24348094

36. Mariscal C, Doolittle WF. Eukaryotes first: How could that be? Philos Trans R Soc Lond B Biol Sci. 2015; 370: 20140322. https://doi.org/10.1098/rstb.2014.0322 PMID: 26323754

37. Harish A, Kurland CG. Mitochondria are not captive bacteria. J Theor Biol. 2017; 434: 88–98. https://doi.org/10.1016/j.jtbi.2017.07.011 PMID: 28754286

38. Dagan T, Roettger M, Bryant D, Martin W. Genome networks root the tree of life between prokaryotic domains. Genome Biol Evol. 2010; 2: 379–392. https://doi.org/10.1093/gbe/evq025 PMID: 20624742

39. Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLoS Genet. 2017; 13: e1006810. https://doi.org/10.1371/journal.pgen.1006810 PMID: 28604769

40. Spang A, Eme L, Saw JH, Caceres EF, Zaremba-Niedzwiedzka K, Lombard J, Guy L, Ettema TJG. Asgard archaea are the closest prokaryotic relatives of eukaryotes. PLoS Genet. 2018; 14: e1007080. https://doi.org/10.1371/journal.pgen.1007080 PMID: 29596421

41. Da Cunha V, Gaia M, Nasir A, Forterre P. Asgard archaea do not close the debate about the universal tree of life topology. PLoS Genet. 2018; 14: e1007215. https://doi.org/10.1371/journal.pgen.1007215 PMID: 29596428

42. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J–F, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013; 499: 431–437. https://doi.org/10.1038/nature12352 PMID: 23851394

43. Williams TA, Embley TM. Archaeal "dark matter" and the origin of eukaryotes. Genome Biol Evol. 2014; 6: 474–481. https://doi.org/10.1093/gbe/evu031 PMID: 24532674

44. Judson OP. The energy expansions of evolution. Nat Ecol Evol. 2017; 1: 138. https://doi.org/10.1038/s41559-017-0138 PMID: 28812646

45. Zachar I, Szathmáry E. Breath-giving cooperation: critical review of origin of mitochondria hypotheses. Biology Direct. 2017; 12. https://doi.org/10.1186/s13062-017-0190-5 PMID: 28806979

46. Lane N. Serial endosymbiosis or singular event at the origin of eukaryotes? J Theor Biol. 2017; 434: 58–67. https://doi.org/10.1016/j.jtbi.2017.04.031 PMID: 28501637

47. Gould SB. Membranes and evolution. Curr Biol. 2018; 28: R381–R385. https://doi.org/10.1016/j.cub.2018.01.086 PMID: 29689219

48. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016; 1: 16048. https://doi.org/10.1038/nmicrobiol.2016.48 PMID: 27572647

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

50. Sánchez-Baracaldo P, Raven JA, Pisani D, Knoll AH. Early photosynthetic eukaryotes inhabited low-salinity habitats. Proc Natl Acad Sci USA. 2017; 114: E7737–E7745. https://doi.org/10.1073/pnas.1620089114 PMID: 28808007

51. Baymann F, Lebrun E, Brugna M, Schoepp-Cothenet B, Giudici-Orticoni MT, Nitschke W. The redox protein construction kit: Pre-last universal common ancestor evolution of energy-conserving enzymes.

Philos Trans R Soc Lond B Biol Sci. 2003; 358: 267–274. https://doi.org/10.1098/rstb.2002.1184 PMID: 12594934

52. Martin WF, Bryant DA, Beatty JT. A physiological perspective on the origin and evolution of photosynthesis. FEMS Microbiol Rev. 2017; 42: 205–231.

53. Doolittle WF. Phylogenetic classification and the universal tree. Science. 1999; 284: 2124–2129. PMID: 10381871

54. Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, et al. The genome of *Methanosarcina mazei*: Evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotechnol. 2002; 4: 453–461. PMID: 12125824

55. Wagner A, Whitaker RJ, Krause DJ, Heilers JH, van Wolferen M, van der Does C, et al. Mechanisms of gene flow in archaea. Nat Rev Microbiol. 2017; 15: 492–501. https://doi.org/10.1038/nrmicro.2017.41 PMID: 28502981

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

57. López-García P, Zivanovic Y, Deschamps P, Moreira D. Bacterial gene import and mesophilic adaptation in archaea. Nat Rev Microbiol. 2015; 13: 447–456. https://doi.org/10.1038/nrmicro3485 PMID: 26075362

58. Nelson-Sathi S, Sousa FL, Roettger M, Lozada-Chávez N, Thiergart T, Janssen A, et al. Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature. 2015; 517: 77–80. https://doi.org/10.1038/nature13805 PMID: 25317564

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

60. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. The net of life: Reconstructing the microbial phylogenetic network. Genome Res. 2005; 15: 954–959. https://doi.org/10.1101/gr.3666505 PMID: 15965028

61. Dagan T, Martin W. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc Natl Acad Sci USA. 2007; 104: 870–875. https://doi.org/10.1073/pnas.0606318104 PMID: 17213324

62. Williams TA, Szöllősi GJ, Spang A, Foster PG, Heaps SE, Boussau B, et al. Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc Natl Acad Sci USA. 2017; 114: E4602–E4611. https://doi.org/10.1073/pnas.1618463114 PMID: 28533395

63. 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: 1655–1663. https://doi.org/10.1099/00207713-50-4-1655 PMID: 10939673

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

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

66. Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. Gene transfer to the nucleus and the evolution of chloroplasts. Nature. 1998; 393: 162–165. https://doi.org/10.1038/30234 PMID: 11560168

67. Puigbò P, Wolf YI, Koonin EV. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. J Biol. 2009; 8: 59. https://doi.org/10.1186/jbiol159 PMID: 19594957

68. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015; 523: 208–211. https://doi.org/10.1038/nature14486 PMID: 26083755

69. Gadagkar SR, Rosenberg MS, Kumar S. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J Exp Zool B Mol Dev Evol. 2005; 304: 64–74. https://doi.org/10.1002/jez.b.21026 PMID: 15593277

70. Thiergart T, Landan G, Martin WF. Concatenated alignments and the case of the disappearing tree. BMC Evol Biol. 2014; 14: 266. https://doi.org/10.1186/s12862-014-0266-0 PMID: 25547755

71. Yang Z, Zhu T. Bayesian selection of misspecified models is over confident and may cause spurious posterior probabilities for phylogenetic trees. Proc Natl Acad Sci USA. 2018; 115: 1854–1859. https://doi.org/10.1073/pnas.1712673115 PMID: 29432193

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

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

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

75. Treangen TJ, Rocha EP. 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 PMID: 21298028

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

77. Bapteste E, Boucher Y, Leigh J, Doolittle WF. Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol. 2004; 12: 406–411. https://doi.org/10.1016/j.tim.2004.07.002 PMID: 15337161

78. Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, 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 PMID: 27562259

79. Enright AJ, Van Dongen S, Ouzounis CA. An ancient algorithm for largescale detection of protein families. Nucleic Acids Res. 2002; 30: 1575–1584. PMID: 11917018

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

81. Lockhart PJ, Steel M, Hendy MD, Penny D. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol. 1994; 11: 605. https://doi.org/10.1093/oxfordjournals.molbev.a040136 PMID: 19391266

82. Doolittle RF. Of URFs and ORFs: A primer on how to analyze derived amino acid sequences. 1st ed. Mill Valley, CA: University Science Books; 1986.

83. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999; 12: 85–94. PMID: 10195279

84. Rossmann MG, Moras D, Olsen KW. Chemical and biological evolution of nucleotide-binding protein. Nature. 1974; 250: 194–199. PMID: 4368490

85. Martin WF, Sousa FL. Early microbial evolution: The age of anaerobes. Cold Spring Harb Perspect Biol. 2015; 8: a018127. https://doi.org/10.1101/cshperspect.a018127 PMID: 26684184

86. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res. 2002; 30: 1427–1464. PMID: 11917006

87. Forterre P. Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: A hypothesis for the origin of cellular domain. Proc Natl Acad Sci USA. 2006; 103: 3669–3674. https://doi.org/10.1073/pnas.0510333103 PMID: 16505372

88. Di Giulio M. An autotrophic origin for the coded amino acids is concordant with the coevolution theory of the genetic code. J Mol Evol. 2016; 83: 93–96. https://doi.org/10.1007/s00239-016-9760-x PMID: 27743002

89. Decker K, Jungerman K, Thauer RK. Energy production in anaerobic organisms. Angew. Chem. Int. Ed. 1970; 9: 138–158.

90. Eck RV, Dayhoff MO. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science. 1966; 152: 363–366. https://doi.org/10.1126/science.152.3720.363 PMID: 17775169

91. Russell MJ, Hall AJ. The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. J Geol Soc Lond. 1997; 154: 377–402.

92. Camprubi E, Jordan SF, Vasiliadou R, Lane N. Iron catalysis at the origin of life. IUBMB Life. 2017; 69: 373–381. https://doi.org/10.1002/iub.1632 PMID: 28470848

93. Shock EL, McCollom T, Schulte MD. The emergence of metabolism from within hydrothermal systems. In: Wiegel J, Adams MWW, editors. Thermophiles: The Keys to Molecular Evolution and the Origin of Life. Taylor and Francis; 1998. pp. 59–76.

94. Laurino P, Tawfik DS. Spontaneous emergence of S-adenosylmethionine and the evolution of methylation. Angew Chem Int Ed Engl. 2017; 56: 343–345. https://doi.org/10.1002/anie.201609615 PMID: 27901309

95. Ragsdale SW. Enzymology of the Wood-Ljungdahl pathway of acetogenesis. Ann NY Acad Sci. 2008; 1125: 129–136. https://doi.org/10.1196/annals.1419.015 PMID: 18378591

96. Sousa FL, Martin WF. Biochemical fossils of the ancient transition from geoenergetics to bioenergetics in prokaryotic one carbon compound metabolism. Biochim Biophys Acta. 2014; 1837: 964–981. https://doi.org/10.1016/j.bbabio.2014.02.001 PMID: 24513196

97. Fuchs G. Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? Annu Rev Microbiol. 2011; 65: 631–658. https://doi.org/10.1146/annurev-micro-090110-102801 PMID: 21740227

98. Varma SJ, Muchowska KB, Chatelain P, Moran J. Native iron reduces CO_2 to intermediates and end-products of the acetyl-CoA pathway. Nat Ecol Evol. 2018; 2: 1019–1024. https://doi.org/10.1038/s41559-018-0542-2 PMID: 29686234

99. Fuchs G. Variations of the acetyl-CoA pathway in diversely related microorganisms that are not acetogens. In Drake HL, editor. Acetogenesis. Chapman & Hall Microbiology Series (Physiology / Ecology / Molecular Biology / Biotechnology). Springer, Boston, MA; 1994. pp. 506–538.

100. Russell MJ, Martin W. The rocky roots of the acetyl-CoA pathway. Trends Biochem Sci. 2004; 29: 358–363. https://doi.org/10.1016/j.tibs.2004.05.007 PMID: 15236743

101. Martin WF, Thauer RK. Energy in ancient metabolism. Cell. 2017; 168: 953–955. https://doi.org/10.1016/j.cell.2017.02.032 PMID: 28283068

102. Semenov SN, Kraft LJ, Ainla A, Zhao M, Baghbanzadeh M, Campbell VE et al. Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. Nature. 2016; 537: 656–660. https://doi.org/10.1038/nature19776 PMID: 27680939

103. Goldford JE, Hartman H, Smith TF, Segrè D. Remnants of an ancient metabolism without phosphate. Cell. 2017; 168: 1126–1134. https://doi.org/10.1016/j.cell.2017.02.001 PMID: 28262353

104. Goldford JE, Segrè D. Modern views of ancient metabolic networks. Curr Opin Syst Biol. 2018; 8: 117–124. https://doi.org/10.1016/j.coisb.2018.01.004

105. Adam PS, Borrel G, Gribaldo S. Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. Proc Natl Acad Sci USA. 2018; 115: E5836–E5837. https://doi.org/10.1073/pnas.1716667115 PMID: 29358391

106. Sato T, Atomi H, Imanaka T. Archaeal type III RuBisCOs function in a pathway for AMP metabolism. Science. 2007; 315: 1003–1006. https://doi.org/10.1126/science.1135999 PMID: 17303759

107. Aono R, Sato T, Imanaka T, Atomi H. A pentose bisphosphate pathway for nucleoside degradation in Archaea. Nat Chem Biol. 2015; 11: 355–360. https://doi.org/10.1038/nchembio.1786 PMID: 25822915

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

109. Castelle CJ, Banfield JF. Major new microbial groups expand diversity and alter our understanding of the tree of life. Cell. 2018; 172: 1181–1197. https://doi.org/10.1016/j.cell.2018.02.016 PMID: 29522741

110. Sousa FL, Preiner M, Martin WF: Native metals, electron bifurcation, and CO_2 reduction in early biochemical evolution. Curr Opin Microbiol. 2018; 43: 77–83. https://doi.org/10.1016/j.mib.2017.12.010 PMID: 29316496

111. Martin W, Russell MJ. On the origin of biochemistry at an alkaline hydrothermal vent. Philos Trans R Soc Lond B. 2007; 362: 1887–1925.

112. Sojo V, Pomiankowski A, Lane N. A bioenergetic basis for membrane divergence in archaea and bacteria. PLoS Biol. 2014; 12: e1001926. https://doi.org/10.1371/journal.pbio.1001926 PMID: 25116890

113. Schlegel K, Leone V, Faraldo-Gómez JD, Müller V. Promiscuous archaeal ATP synthase concurrently coupled to Na^+ and H^+ translocation. Proc Natl Acad Sci USA. 2012; 109: 947–952. https://doi.org/10.1073/pnas.1115796109 PMID: 22219361

114. Buckel W, Thauer RK. Energy conservation via electron bifurcating ferredoxin reduction and proton/Na^+ translocating ferredoxin oxidation. Biochim Biophys Acta. 2013; 1827: 94–113. https://doi.org/10.1016/j.bbabio.2012.07.002 PMID: 22800682

115. Wolfenden R, Lewis CA Jr., Yuan Y, Carter CW Jr.. Temperature dependence of amino acid hydrophobicities. Proc Natl Acad Sci USA. 2015; 112: 7484–7488. https://doi.org/10.1073/pnas.1507565112 PMID: 26034278

116. Stockbridge RB, Lewis CA Jr., Yuan Y, Wolfenden R. Impact of temperature on the time required for the establishment of primordial biochemistry, and for the evolution of enzymes. Proc Natl Acad Sci USA. 2010; 107: 22102–22105. https://doi.org/10.1073/pnas.1013647107 PMID: 21123742

117. Chapelle FH, O'Neill K, Bradley PM, Methé BA, Ciufo SA, Knobel LL, et al. A hydrogen-based subsurface microbial community dominated by methanogens. Nature. 2002; 415: 312–315. https://doi.org/10.1038/415312a PMID: 11797006

118. Lever MA, Heuer VB, Morono Y, Masui N, Schmidt F, Alperin MJ, et al. Acetogenesis in deep subseafloor sediments of the Juan de Fuca Ridge Flank: A synthesis of geochemical, thermodynamic, and gene-based evidence. Geomicrobiol J. 2010; 27: 183–211.

119. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: The unseen majority. Proc Natl Acad Sci USA. 1998; 95: 6578–6583. PMID: 9618454

120. Hordijk W, Steel M. Chasing the tail: The emergence of autocatalytic networks. Biosystems. 2017; 152: 1–10. https://doi.org/10.1016/j.biosystems.2016.12.002 PMID: 28027958

121. Steel M, Kauffmann S. A note on random catalytic branching processes. J Theoret Biol. 2018; 437: 222–224.

122. Koonin EV, Martin W. On the origin of genomes and cells within inorganic compartments. Trends Genet. 2005; 21: 647–654. https://doi.org/10.1016/j.tig.2005.09.006 PMID: 16223546

123. Baross JA, Martin WF. The ribofilm as a concept for life's origins. Cell 2015; 162: 13–15. https://doi.org/10.1016/j.cell.2015.06.038 PMID: 26140586

124. Sutherland JD. Opinion: Studies on the origin of life–the end of the beginning. Nat Rev Chem. 2017; 10: 0012. https://doi.org/10.1038/s41570-016-0012

125. Hartmann RK, Mörl M, Sprinzl M. The tRNA world. RNA. 2004; 10: 344–349. https://doi.org/10.1261/rna.5240904 PMID: 14970379

126. Grosjean H, Breton M, Sirand-Pugnet P, Tardy F, Thiaucourt F, Citti C, et al. Predicting the minimal translation apparatus: Lessons from the reductive evolution of mollicutes. PLoS Genet. 2014; 10: e1004363. https://doi.org/10.1371/journal.pgen.1004363 PMID: 24809820

127. Grosjean H, Gupta R, Maxwell ES. Modified nucleotides in archaeal RNAs. In: Blum P, editor. Archaea: New Models for Prokaryotic Biology. Norfolk, UK: Caister Academic Press; 2008. pp. 171–196.

128. Yarian C, Townsend H, Czestkowski W, Sochacka E, Malkiewicz AJ, Guenther R, et al. Accurate translation of the genetic code depends on tRNA modified nucleosides. J Biol Chem. 2002; 277: 16391–16395. https://doi.org/10.1074/jbc.M200253200 PMID: 11861649

129. Gustilo EM, Vendeix FA, Agris PF. tRNA's modifications bring order to gene expression. Curr Opin Microbiol. 2008; 11: 134–140. https://doi.org/10.1016/j.mib.2008.02.003 PMID: 18378185

130. Decatur WA, Fournier MJ. rRNA modifications and ribosome function. Trends Biochem Sci. 2002; 27: 344–351. PMID: 12114023

131. Schneider C, Becker S, Okamura H, Crisp A, Amatov T, Stadlmeier M, et al. Noncanonical RNA nucleosides as molecular fossils of an early Earth—Generation by prebiotic methylations and carbamoylations. Angwandte Chemie Int Ed. 2018; 57: 1–5.

132. Lawrence JG, Ochman H. Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci USA.1998; 95: 9413–9417. PMID: 9689094

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

134. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev. 2005; 15: 589–594. https://doi.org/10.1016/j.gde.2005.09.006 PMID: 16185861

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

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

137. Kummerfeld SK, Teichmann SA. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. 2005; 21: 25–30. https://doi.org/10.1016/j.tig.2004.11.007 PMID: 15680510

138. Méheust R, Watson AK, Lapointe FJ, Papke RT, Lopez P, Bapteste E. Hundreds of novel composite genes and chimeric genes with bacterial origins contributed to haloarchaeal evolution. Genome Biol. 2018; 19:75. https://doi.org/10.1186/s13059-018-1454-9

139. Henry CS, Lerma-Ortiz C, Gerdes SY, Mullen JD, Colasanti R, Zhukov A, et al. Systematic identification and analysis of frequent gene fusion events in metabolic pathways. BMC Genomics. 2016; 17: 473. https://doi.org/10.1186/s12864-016-2782-3 PMID: 27342196

140. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature. 1999; 402: 86–90. https://doi.org/10.1038/47056 PMID: 10573422

141. Gogarten JP, Deamer D. Is LUCA a thermophilic progenote? Nat Microbiol. 2016; 1: 16229. https://doi.org/10.1038/nmicrobiol.2016.229 PMID: 27886195

142. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA. 1977; 74: 5088–5090. PMID: 270744

143. Weiss MC, Neukirchen S, Roettger M, Mrnjavac N, Nelson-Sathi S, Martin WF, et al. Reply to 'Is LUCA a thermophilic progenote?' Nat Microbiol. 2016; 1: 16230. https://doi.org/10.1038/nmicrobiol.2016.230 PMID: 27886196

144. Martin WF, Weiss MC, Neukirchen S, Nelson-Sathi S, Sousa FL. Physiology, phylogeny, and LUCA. Microb Cell. 2016; 3: 582–587. https://doi.org/10.15698/mic2016.12.545 PMID: 28357330

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

146. Liu Y, Beer LL, Whitman WB. Methanogens: A window into ancient sulfur metabolism. Trends Microbiol. 2012; 20: 251–258. https://doi.org/10.1016/j.tim.2012.02.002 PMID: 22406173

147. Muchowska KB, Varma SJ, Chevallot-Beroux E, Lethuillier-Karl L, Li G, Moran J. Metals promote sequences of the reverse Krebs cycle. Nat Ecol Evol. 2017; 1: 1716–1721. https://doi.org/10.1038/s41559-017-0311-7 PMID: 28970480

148. McCollom TM. Abiotic methane formation during experimental serpentinization of olivine. Proc Natl Acad Sci USA. 2016; 113: 13965–13970. https://doi.org/10.1073/pnas.1611843113 PMID: 27821742

5.6 Origin and phylogenetic relationships of [4Fe-4S]-containing O_2-sensors of bacteria

C. Barth[1], M. C. Weiss[2], M. Roettger[2], W.F. Martin[2] und G. Unden[1]

Affiliations

1 Microbilogy and Wine Research, Institute for Molecular Physiology, Johannes Gutenberg University Mainz, Deutschland.
2 Institut für Molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Deutschland.

Dieser Artikel wurde am 10. September 2018 in *Enviornmental Microbiology* Ausgabe 20 veröffentlicht.

Beitrag von Madeline C. Weiß:

Die gesamte bioinformatische Analyse habe ich durchgeführt. Dazu gehörten eine BLAST-Suche der Proteinsequenzen der Sauerstoffsensoren gegen die Proteinfamilien, welche in Weiss *et al.* (2016b) verwendet wurden, sowie die Erstellung von Multiplen Sequenzalignments und phylogenetischen Bäumen. Des Weiteren wurden von C. Barth und mir die phylogenetischen Bäume, hinsichtlich der Verteilung von Sauerstoffsensoren besitzenden Organismen, ausgewertet. Ein weiterer Teil der bioinformatischen Analyse war die Erstellung der Anwesenheits-Abwesenheits Matrizen, dargestellt in den Abbildungen 2, 5 und 6. Des Weiteren habe ich an dem Text mitgewirkt und die Methoden beschrieben.

environmental microbiology

Environmental Microbiology (2018) **20**(12), 4567–4586 doi:10.1111/1462-2920.14411

Origin and phylogenetic relationships of [4Fe–4S]-containing O_2 sensors of bacteria

C. Barth,[1] M. C. Weiss,[2] M. Roettger,[2] W. F. Martin[2] and G. Unden[1*]

[1]*Microbiology and Wine Research, Institute for Molecular Physiology, Johannes Gutenberg University Mainz, Mainz, Germany.*

[2]*Institute for Molecular Evolution, Heinrich Heine University of Düsseldorf, Düsseldorf, Germany.*

Summary

The advent of environmental O_2 about 2.5 billion years ago forced microbes to metabolically adapt and to develop mechanisms for O_2 sensing. Sensing of O_2 by $[4Fe–4S]^{2+}$ to $[2Fe–2S]^{2+}$ cluster conversion represents an ancient mechanism that is used by FNR_{Ec} (*Escherichia coli*), FNR_{Bs} (*Bacillus subtilis*), $NreB_{Sa}$ (*Staphylococcus aureus*) and $WhiB3_{Mt}$ (*Mycobacterium tuberculosis*). The phylogenetic relationship of these sensors was investigated. FNR_{Ec} homologues are restricted to the proteobacteria and a few representatives from other phyla. Homologues of FNR_{Bs} and $NreB_{Sa}$ are located within the bacilli, of WhiB3 within the actinobacteria. Archaea contain no homologues. The data reveal no similarity between the FNR_{Ec}, FNR_{Bs}, $NreB_{Sa}$ and WhiB3 sensor families on the sequence and structural levels. These O_2 sensor families arose independently in phyla that were already present at the time O_2 appeared, their members were subsequently distributed by lateral gene transfer. The chemistry of [4Fe–4S] and [2Fe–2S] cluster formation and interconversion appears to be shared by the sensor protein families. The type of signal output is, however, family specific. The homologues of FNR_{Ec} and $NreB_{Sa}$ vary with regard to the number of Cys residues that coordinate the cluster. It is suggested that the variants derive from lateral gene transfer and gained other functions.

Received 15 June, 2018; accepted 10 September, 2018. *For correspondence. E-mail unden@uni-mainz.de; Tel (+49) 6131 3923550; Fax (+49) 6131 3922695.

Introduction

In modern microbes, molecular oxygen is important for many aspects of bacterial physiology, mainly catabolism (oxidation, respiration) and stress response. Cyanobacteria started producing O_2 about 2.5 billion years ago (Fischer *et al*., 2016). Prior to that time, there were only anaerobes on Earth. The appearance of an O_2-enriched atmosphere confronted microbes with the presence of a strong oxidant that could readily inactivate or otherwise poison the active site and reactive groups of ancient enzymes, mostly metalloenzymes, that we today view as oxygen sensitive (Raymond and Segre, 2006; Martin and Sousa, 2015). These forced microbes to find mechanisms of detecting and dealing with O_2, mechanisms that have persisted to the present. Access to aerobic or anaerobic niches requires extensive remodelling of energy conserving metabolic and anabolic pathways (see review by Unden and Bongaerts, 1997). Additionally, during aerobic growth reactive oxygen species such as (hydrogen)peroxide, superoxide and hydroxyl radicals are formed which require a protective response by the bacteria (for reviews see Imlay 2002; 2006).

Because of its ecological and physiological importance, most bacteria contain sensors for detecting O_2 and reactive oxygen species. Sensors for O_2 are essentially restricted to facultative anaerobic and (micro)aerobic bacteria. About 2.5×10^9 years before present, atmospheric O_2 rose to estimated levels of about 0.02–0.04 atm (Holland, 2006), close to the Pasteur point (Engelhardt, 1974) for the onset of aerobic metabolism. The Pasteur point coincides with the switch point (approx. 0.5–2% of air saturation) of O_2-regulated genes in facultative metabolism of *Escherichia coli* that respond to FNR_{Ec} and other O_2 sensors (Becker *et al*., 1996; Tseng *et al*., 1996; Becker *et al*., 1997). O_2 levels of 0.02–0.04 atm resulted in mild oxygenation of ocean surface waters, a situation that persisted for almost 2 billion years (Lenton *et al*., 2016) because deep ocean oxygenation was not completed until roughly 580–430 million years ago (Stolper and Keller, 2018). In response to environmental oxygenation, bacteria evolved a variety of O_2-sensing systems, classified as direct and indirect O_2 sensors, that recruited

different biochemical mechanisms for monitoring O_2 levels (Green and Paget, 2004; Unden *et al*., 2010).

Direct O_2 sensors use either O_2-sensitive [4Fe–4S] clusters (such as the FNR proteins of *E. coli* and *Bacillus subtilis* and NreB of *Staphylococcus carnosus*), heme B (such as FixL of *Sinorhzobium meliloti* and Dos of *E. coli*) or FAD (such as NifL of *Azotobacter vinelandii*) as O_2-reactive prosthetic groups (for review see Green and Paget, 2004; Unden *et al*., 2010). In *Pseudomonas* species, a system related to the hypoxia-inducible transcription factor (HIF) of animals has been described that employs O_2-dependent hydroxylation reactions to sense decreased O_2 availability (Scotti *et al*., 2014; Schmidt *et al*., 2016).

Indirect O_2 sensors respond to metabolites and pathway intermediates that change their cellular concentration or redox state in response to O_2 availability. The ArcAB two-component system of *E. coli* responds to the redox state of the respiratory quinones or changed quinone/quinol ratios (Malpica *et al*., 2004). The Rex transcriptional regulator of Gram-positive bacteria on the other hand measures the cellular NADH/NAD ratio, which changes as a function of O_2-reducing respiratory activity (Brekasis and Paget, 2003; Sickmier *et al*., 2005; Wang *et al*., 2008).

O_2-labile [4Fe–4S] clusters as universal cofactors for O_2 sensing

Iron–sulphur clusters are widespread and ancient metal cofactors of proteins that are composed of iron ions and sulfide. The clusters are coordinated via weak covalent bonds with cysteine thiol sidechains of the protein. Two-iron–sulfur [2Fe–2S], three-iron–sulfur [3Fe–4S] and four-iron–sulfur [4Fe–4S] clusters are common in proteins (Beinert *et al*., 1997). Proteins with [4Fe–4S] clusters are most versatile and have roles in protein-bound electron transfer in ferredoxin, fumarate reductase and many other redox or respiratory enzymes (Beinert, 1976; Malkin and Rabinowitz, 1967; Lancaster *et al*., 1999), as catalytically active sites in hydratases such as aconitase (Beinert *et al*., 1996), and in iron and redox responsive regulatory proteins (Beinert *et al*., 1996; Mettert and Kiley, 2015). All direct O_2 sensors with FeS clusters use [4Fe–4S] clusters. For this reason only [4Fe–4S] cluster binding sensors will be considered here.

In aconitase and related enzymes, the [4Fe–4S] cluster is required for binding and activation of the substrate (Beinert *et al*., 1996). In addition to its catalytic role, the cluster has a structural role in aconitase (Beinert *et al*., 1996). Thus, apo-aconitase is catalytically inactive but has a regulatory function (Beinert *et al*., 1996; Rouault *et al*., 1992). Loss of the iron–sulfur cluster causes structural rearrangements that allow binding of the apo-enzyme (iron regulatory protein; IRP) to iron-responsive elements (IREs), which are located in the 5′ region of the mRNA of iron homeostasis genes (Kaptain *et al*., 1991; Tang and Guest, 1999; Tang *et al*., 2005; Commichau and Stulke, 2008). IRE binding controls gene expression through mRNA stability.

By contrast, the transcriptional bacterial regulators FNR_{Ec}, FNR_{Bs}, NreB, WhiB3 and NsrR (Volbeda *et al*., 2017) contain [4Fe–4S] clusters that react chemically with molecular O_2 or NO, which controls the function of the sensors by cluster conversion and modification. The physiological role of FNR_{Ec}, FNR_{Bs}, and NreB is essentially that of O_2 sensors, whereas that of WhiB3 can be either that of an O_2 or NO sensor depending on physiological conditions (Singh *et al*., 2007). For each of the sensors, reaction with O_2 has been characterized in detail (see below). NsrR, in contrast, is essentially an NO sensor and the [4Fe–4S] cluster is used for NO response (Volbeda *et al*., 2017). For this reason in the present work only, FNR_{Ec}, FNR_{Bs}, NreB and WhiB3 will be discussed. The iron–sulfur clusters are surface exposed and are able to react with O_2. FNR_{Ec}, FNR_{Bs} and WhiB3 represent transcriptional regulators (Fig. 1). DNA binding and their function as transcriptional regulators is controlled by the [4Fe–4S] clusters which react directly with O_2. In the FNR proteins, the reaction of the cluster with O_2 causes cluster degradation. As a consequence FNR_{Ec} monomerizes and loses the ability for DNA binding and transcriptional activation (Lazazzera *et al*., 1996), whereas FNR_{Bs} is a permanent dimer but loses DNA binding due to conformational changes after degradation of the FeS cluster (Reents *et al*., 2006b).

In WhiB3, the [4Fe–4S] cluster is required for complex formation with the sigma factor σ^A. Degradation of the FeS cluster disassembles the complex and transcriptional activity (Kudhair *et al*., 2017). In the sensor kinase NreB, the sensory PAS domain controls the activity of the kinase domain in response to O_2 via the iron–sulfur cluster. Auto-phosphorylation of NreB leads to the phosphorylation of the response regulator NreC that activates in the phosphorylated state (NreC-P) the expression of target genes.

The Cys residues that ligate the iron–sulfur clusters in FNR_{Ec}, FNR_{Bs}, WhiB3 and NreB are located in clusters that differ in sequence, spacing and location within the sensors (Fig. 1B). In FNR_{Bs} one of the ligands of the iron–sulfur cluster is replaced by an Asp residue (Gruner *et al*., 2011). FNR_{Ec} was the first O_2 sensor of this type and represents the prototype of this type of O_2 sensors (Shaw and Guest, 1982; Shaw *et al*., 1983; Green and Paget, 2004). The crystal structure of the FNR_{Ec} type FNR_{Af} from *Aliivibrio fischeri* was solved (Volbeda *et al*., 2015). The protein consists of two domains that provide the sensory and the DNA-binding function. The N-

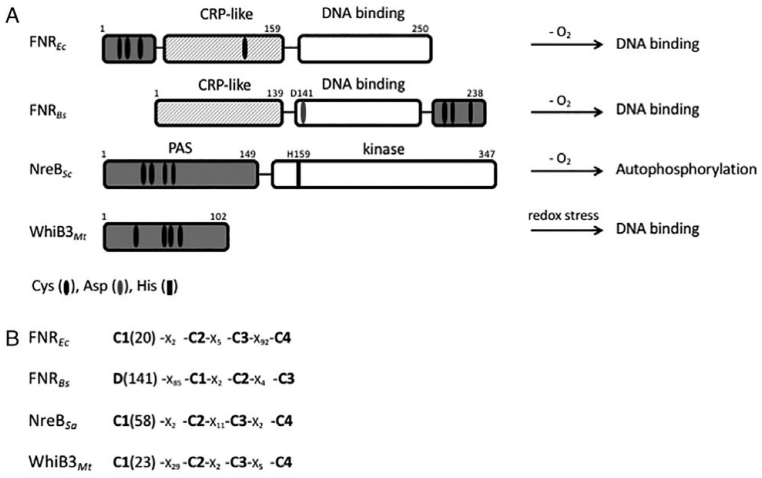


Fig. 1. Domain structure (A) and conserved cluster-binding motif (B) of the $[4Fe–4S]^{2+}$-containing sensors FNR, NreB and WhiB3.A. The cysteine-carrying sensory domains (grey), the output domains and the kind of response (DNA binding or autophosphorylation) are depicted. Approximate positions of cluster coordinating conserved cysteine and aspartate residues as well as phosphorylation residue His159 of sensor kinase NreB of *S. carnosus* are marked. B. The [Fe–S]-cluster binding sequence motifs consisting of conserved cysteine (C) and aspartate (D) residues are shown as well as their position in the protein. Variable amino acid residues between cluster ligands are indicated with x and the corresponding number. Modified after Unden and colleagues (2013).

terminal sensory domain ligates under anoxic conditions a $[4Fe–4S]^{2+}$ cluster and under oxic conditions a $[2Fe–2S]^{2+}$ cluster.

The $[4Fe–4S]^{2+}$ cluster binding sensors FNR_{Ec} and FNR_{Af} are dimers with an α-helical dimer interface, and the dimeric state is required for site specific DNA-binding (Lazazzera *et al*., 1993; Khoroshilova *et al*., 1995; Kiley and Beinert, 1998). Conversion to the $[2Fe–2S]^{2+}$ form results in a rearrangement of the dimer interface causing monomerization and loss of specific DNA-binding (Khoroshilova *et al*., 1995; Volbeda *et al*., 2015). The $[4Fe–4S]^{2+}$/$[2Fe–2S]^{2+}$ cluster conversion was identified by Mössbauer and EPR spectroscopy (Lazazzera *et al*., 1993; Khoroshilova *et al*., 1995; Lazazzera *et al*., 1996; Khoroshilova *et al*., 1997; Kiley and Beinert, 1998). Combination of visible absorbance, EPR spectroscopy and time resolved electrospray ionization mass spectrometry allowed a very detailed analysis of the reactions. The O_2 triggers a reaction in two steps (Crack *et al*., 2006; 2007; 2008):

$$\text{Step 1}: [4Fe-4S]^{2+} + O_2 \rightarrow [3Fe-4S]^{1+} + Fe^{2+} + O_2^-$$
$$\text{Step 2}: [3Fe-4S]^{1+} \rightarrow [2Fe-2S]^{2+} + Fe^{3+} + 2S^{2-}$$

Step 2 apparently involves a second oxidation by O_2 (Crack *et al*., 2017; Zhang *et al*., 2012). The two sulfide ions in step 2 are not released into aqueous solution but oxidized to sulfane (S^0) and form a persulfide with two of the Cys ligands (RS^-) of the cluster. Up to two of the Cys ligands exist then in the persulfide (RSS^-) state. The reaction of step 2 has to be reformulated accordingly (Step 2*):

$$\text{Step 2}^*: [3Fe-4S](RS)_3 + RS^- + O_2 + 4H^+ \rightarrow [2Fe-2S](RS)_2(RSS)_2 + Fe^{3+} + 2H_2O.$$

Formation of the Cys-persulfide provides a mechanism for storing the sulfur released from the iron–sulfur cluster during the [4Fe–4S]/[2Fe–2S] cluster conversion rather than releasing it to the water space. This mechanism allows reversion of the O_2-inactivated FNR and to the anaerobic $[4Fe–4S]^{2+}$ form by reduction and repair without involvement of the iron–sulfur biosynthesis machinery (Zhang *et al*., 2012; Crack *et al*., 2017).

In NreB, a $[4Fe–4S]^{2+}$/$[2Fe–2S]^{2+}$ cluster conversion is the basis for O_2 sensing (Mullner *et al*., 2008), which suggests a reaction sequence similar to FNR_{Ec}. In the anoxic form, FNR_{BS} and WhiB3 contain a $[4Fe–4S]^{2+}$ cluster as well (Jakimowicz *et al*., 2005; Reents *et al*., 2006b; Crack *et al*., 2009). It appears therefore that FNR_{Ec}, FNR_{Bs}, NreB and WhiB3 use the same cofactor for O_2 sensing and that the reactions occurring at the $[4Fe–4S]^{2+}$ during response to O_2 are similar.

Here, we investigate the phylogenetic relationships of bacterial O_2 sensors using $[4Fe–4S]^{2+}$ clusters for sensing. The protein sequences of FNR_{Ec}, FNR_{Bs}, *Staphylococcus aureus* NreB and *Mycobacterium tuberculosis* WhiB3 were used to identify homologues among gene families (clusters) generated from 1981 sequenced prokaryotic genomes and screened for retainment of the consensus Cys clusters for binding the iron–sulfur clusters. The latter represent the most characteristic feature of the proteins in order to characterize their potential for O_2 sensing and their distribution and variation among prokaryotes.

Results

Clustering and phylogenetic distribution of FNR_{Ec}, FNR_{Bs}, $NreB_{Sa}$ and WhiB3 homologues

The presence or absence of matrix in Fig. 2 summarizes the occurrence of FNR_{Ec}, FNR_{Bs}, $NreB_{Sa}$, $WhiB3_{Mt}$ and their homologues in the bacterial taxa indicated. No homologues were detected in archaea. The occurrence of CRP, a global regulator and homologue of FNR_{Ec} that lacks O_2 sensing clusters, is also indicated. Each column represents a cluster, ticks summarizing the results of BLAST searches and clustering to detect homologues (black ticks indicate presence; white ticks indicate absence) from prokaryote genomes in the RefSeq 2012 database. The clusters were generated using the standard Markov cluster algorithm (Enright *et al*., 2002) (MCL) at a 25% global identity threshold as previously described (Nelson-Sathi *et al*., 2015; Weiss *et al*., 2016). Each black tick indicates the presence of one (or more) homologous protein in the corresponding species (rows)

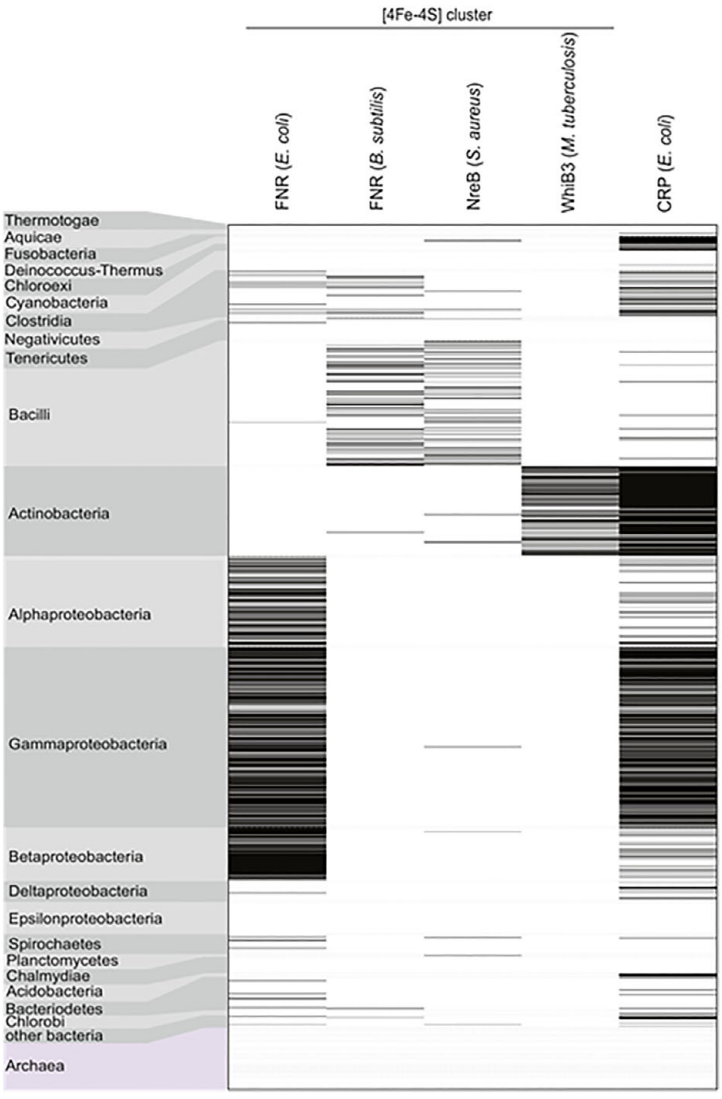


Fig. 2. Matrix of [Fe–S]-containing sensors FNR, NreB, WhiB3 and global regulator CRP.The matrix shows the distribution of the identified homologous proteins for each reference protein during BLAST search. One black line corresponds to one homologous protein in the respective bacteria. On the vertical axis, the predefined bacterial phyla (sometimes classes) are displayed.

with a BLAST hit with ≥ 25% local identity and a pairwise alignment with ≥ 25% global identity. Protein sequence similarity was scored without weighting of specific motifs. To distinguish between homologous proteins that could function as oxygen sensor and false positive sensors, the presence of the cluster-ligating Cys motif was scored. FNR homologues that lack N- or C-terminal Cys motifs were scored as CRP-type regulators due to the absence of the oxygen-responsive FeS cluster required for sensor function. The NreB-type regulators are identified by the similarity of PAS and kinase domain.

FNR of the *E. coli*-type (FNR_{Ec}) predominantly occurs among the α-, β- and γ-Proteobacteria but some homologues are found in the clostridia, spirochaetes, bacteriodetes and other phyla. By contrast, the *B. subtilis*-type FNR (FNR_{Bs}) occurs mainly within the bacilli and has no homologues in the proteobacteria. Few additional BLAST hits are found within the clostridia, negativicutes, bacteriodetes and actinobacteria. The sensor kinase NreB from *S. aureus* and its homologues are mostly restricted to the Bacilli but there are few similar proteins in other firmicutes and in the actinobacteria, proteobacteria, spirochaetes, planctomycetes as well as in the deinococcus-thermus groups. The *M. tuberculosis* WhiB3 sensor is limited to the actinobacteria. The global regulator CRP from *E. coli* that is related to FNR_{Ec} and FNR_{Bs} (Korner *et al*., 2003) has homologues in almost all bacterial phyla. Most of the CRP homologues are found within the actinobacteria and the proteobacteria whereas the bacilli that harbour most of the FNR_{Bs} proteins are mostly devoid of CRP. In general, the oxygen sensors FNR_{Ec}, FNR_{Bs}, NreB and WhiB3 are restricted to specific bacterial phyla whereas CRP occurs in most bacterial phyla. Several phyla lack specific O_2 sensors of this type (aquifex, thermotoga, fusobacteria, ε-proteobacteria, acidobacteria, chalmydiae, chlorobi, cyanobacteria and the green filamentous bacteria) or only show sporadic occurrence (bacteroides, planctomyces, tenericutes, δ-proteobacteria, negativicutes and deinococcus-thermus). Notably, the archaea, including the aerobic haloarchaea (Euryarchaeota) and the (facultatively) aerobic hyperthermophilic Crenarchaeota (*Sulfolobus*, *Acidianus*, *Pyrobaculum*, *Metallosphaera*) are devoid of FNR_{Ec}, FNR_{Bs}, NreB or WhiB3 type sensors.

The iron–sulfur clusters of FNR_{Ec}, FNR_{Bs}, $NreB_{Sc}$ and WhiB3 are co-ordinated by Cys and occasionally Asp residues. FNR_{Ec} and FNR_{Bs} are distant homologues of the CRP protein (Shaw *et al*., 1983; Korner *et al*., 2003). FNR_{Ec} contains a short N-terminal extension to CRP of about 29 AA (Fig. 1A) that are specific for FNR_{Ec} (Shaw *et al*., 1983). Three of the four Cys residues (C1–C3) of FNR_{Ec} for ligating the iron–sulfur cluster are placed in the FNR-typic extension (Fig. 1B), only the fourth residue (C122) is located in the CRP homologous region. FNR_{Bs} that is also distantly homologous to CRP carries the Cys cluster in a C-terminal extension downstream the CRP homologous region. Only the first ligand (D141) is located in the CRP homologous region. Replacement of single residues of the binding clusters by other residues generally inactivates their capacity for O_2 sensing (Melville and Gunsalus, 1990; Green *et al*., 1993; Kamps *et al*., 2004; Gruner *et al*., 2011), demonstrating the significance of the residues for the basic function of the proteins. The proteins of the searches (Fig. 2) were therefore screened for the presence of the respective Cys/Asp clusters and the presence of sensory domain. Refined overviews of FNR_{Ec}, FNR_{Bs}, $NreB_{Sc}$ and WhiB3 homologues are presented in the following sections including an evaluation of their presumptive properties in O_2 sensing.

FNR-Ec-type sensors in the prokaryotic kingdom

FNR_{Ec} is a major regulator of the aerobic/anaerobic switch in *E. coli* and responsible for the induction of genes of anaerobic and microaerobic respiration, fermentation and anaerobiosis related genes. Transcriptional activation by FNR_{Ec} depends on its dimeric state which is controlled by the [4Fe–4S]/[2Fe–2S] cluster conversion in response to O_2 presence. The matrix of Fig. 2 contains altogether 414 FNR_{Ec} homologues after deletion of redundant or closely related hits from related species. From the remaining, 95% of the homologues were located within the α-, β- and γ-proteobacteria (Fig. 3), with the γ-proteobacteria enclosing the largest number. The residual hits (5% of 414) were in diverse phyla outside the proteobacteria, that is in the clostridia (six strains), bacteriodetes (six strains), the spirochaetes (three strains) and others (Fig. 4). The homologues with the 'diverse' origin showed a closed clustering separate from the proteobacteria. For all homologues within the α-, β- and γ-proteobacteria the phylogenetic tree of the homologues agrees with that of their hosts (Supporting Information Fig. S1), suggesting that the proteins co-evolved with their phyla.

The 414 FNR-Ec type homologues were investigated for the conservation of the Cys clusters by recording the number and spacing of Cys residues. Spacing variants delineate subfamilies of FNR proteins, whereas conservation or loss of Cys residues indicates conservation of O_2 sensing. The majority of FNR_{Ec} homologues (79%) retained a 4Cys cluster (Fig. 3), and the remaining 21% were lacking one or more of the Cys residues (Fig. 3). The variants containing less than 4Cys residues of the cluster are mostly scattered in the phyla with some specific preferences (Figs 3 and 4). The 4Cys variants represent the majority of the homologues in the α-, β- and γ-proteobacteria, whereas in the other bacteria, the variants with ≤ 3Cys residues predominate (12 from

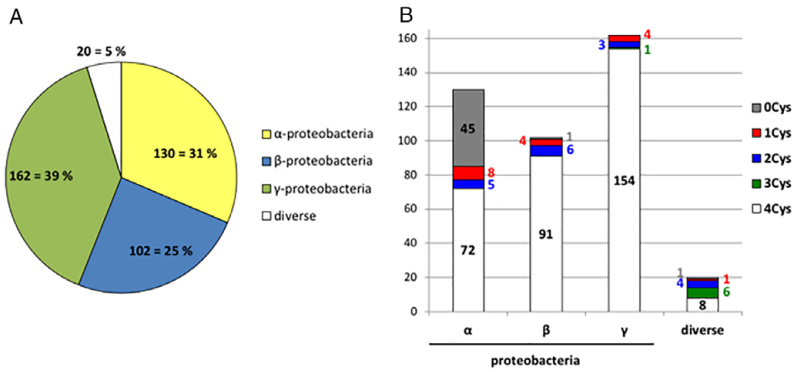


Fig. 3. Distribution of FNR_{Ec} homologues and Cys variants.The homologues (664 proteins) were identified by the BLAST search followed by the removal of close homology (> 80% global identity) in order to delete hits from closely related strains. The resulting homologues (414 proteins) represent the basis for the data in A and B. A. The distribution of the FNR_{Ec} homologues in the bacterial phyla. B. The number of Cys variants within the bacterial phyla from the total of 414 FNR_{Ec} homologues.The α-, β- and γ-proteobacteria of (A) show the same colouring as in the phylogenetic tree of Supporting Information Fig. S1. The Cys variants in B) are labelled by colour code 0Cys (grey), 1Cys (red), 2Cys (blue), 3Cys (green) and 4Cys (white).

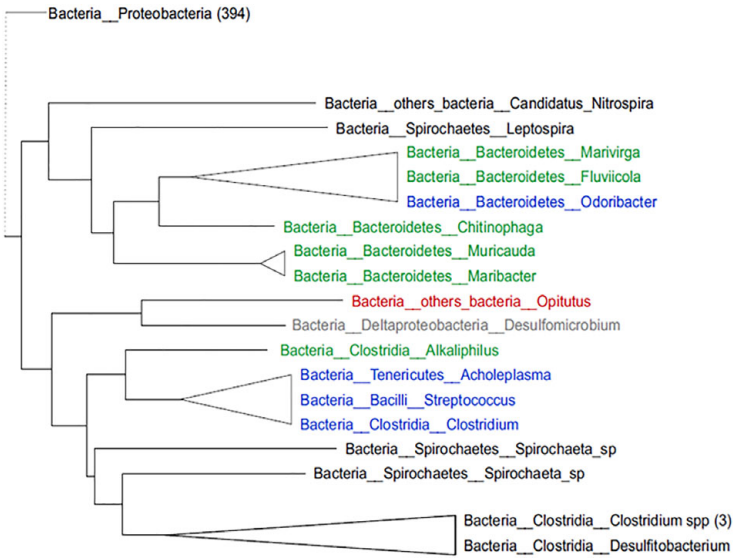


Fig. 4. FNR_{Ec} variants outside the proteobacteria phylogenetic group. Labelled in black, green, blue, red and grey are variants with four, three, two, one or zero conserved Cys residues respectively. The digits in brackets correspond to the amount of FNR homologues in the respective genus or phylum.

20 homologues). The function of these ≤ 3Cys variants has not been studied, but their prevalence suggests a modified function.

The 3Cys variants are rare and are found only in the γ-proteobacteria and the 'diverse lineages' group. The 2Cys and 1Cys variants are found in all protobacterial phyla at a frequency of ≤ 10% of phylum members each. The 0Cys variants are characteristic for the α-proteobacteria and represent 35% of the FNR_{Ec} homologues (Fig. 3). In the following, the FNR_{Ec} Cys variants will be discussed with respect to their possible function, but we note that only 4Cys and 0Cys variants of FNR_{Ec} have been characterized by experiments.

4Cys variants of FNR_{Ec}

Within the 4Cys FNR_{Ec} proteins, three types of spacing are observed. The **C1**–X_2–**C2** spacing is conserved in all types, but for the C2/C3 and C3/C4 pairs some variation can be seen (Table 1 and Supporting Information Figs S2

Table 1. Occurrence and properties of FNR_{Ec} variants in relation to their Cys clusters.

FNR_{Ec} type	Cys clusters	Bacteria (examples)	Function or regulated genes	References
FNR (4Cys)	**C1**–X_2–**C2**–X_5–**C3**–X_{92}–**C4**	γ-Proteobacteria	Transcriptional regulation in response to O_2 (anaerobic and microaerobic respirations; fermentation; related genes)	Galimand and colleagues (1991), Salmon and colleagues (2003), Sawers (1991), Shaw and Guest (1982), Spiro and Guest (1988) and Unden and Bongaerts (1997)
FNR, Anr, EtrA, HlyX, CydR, FnrA	(Alignment, Supporting Information Fig. S2)			
FNR (4Cys)	**C1**–X_2–**C2**–X_5–**C3**–X_{92}–**C4**	β-Proteobacteria	Anaerobic growth; virulence; anaerobic regulation fumarate and nitrate respiration; sugar fermentation	Bartolini and colleagues (2006), Edwards and colleagues (2010) and Wood and colleagues (1998)
Fnr, Btr	(Alignment, Supporting Information Fig. S2)	(*Bordetella pertussis*; *Neisseria meningitides*)		
FNR (4Cys)	**C1**–X_2–**C2**–X_7–**C3**–X_{87}–**C4**	α-Proteobacteria	N_2 fixation, anaerobic respiration	Anthamatten and colleagues (1992), Batut and Boistard (1994), Schluter and colleagues (1992) and Zeilstra-Ryalls and colleagues (1997)
FnrN, FnrL, AadR, $FixK_1$	(Alignment, Supporting Information Fig. S2)	(*Rhizobium* spp., *Rhodo-bacter* spp., *Bradyrhizobium* spp. and others)		
Fnr		γ-Proteobacteria [(*Pseudo-*) *Xanthomonas*]	Not characterized	Not characterized
FNR (4Cys)	**C1**–X_2–**C2**–X_{3-8}–**C3**–X_{88-98}–**C4**	Diverse bacteria	Not characterized	Not characterized
	(Alignment, Supporting Information Fig. S3)	(Clostridia, spirochaeta, leptospira)		
FNR (3Cys)	**C1**–X_2–**C2**–X_{97-98}–**C4**	γ-Proteobacteria	Not characterized	Not characterized
	(Alignment, Supporting Information Fig. S4)	(*Acidithiobacillus caldus*) Diverse bacteria (bacteriodetes, *Clostridium* spp.)		
FNR (2Cys)	**C3**–X_{92}–**C4** (most common)	Proteobacteria	Not characterized	Not characterized
	C2–X_5–**C3**	α-Proteobacteria (*Methylobacterium* spp.)		
	C2–X_{97-98}–**C4**	Diverse bacteria (bacteriodetes, clostridia, bacilli, tenericutes)		
FNR (1Cys)	**C4 (**or **C1)**	α-,β-, γ-Proteobacteria Diverse bacteria (*Opitutus*)	Not characterized	
FNR (0Cys)	No C	α-Proteobacteria (often in addition to 4Cys-FNR)	N_2 fixation (*fix* and *nif* genes), nitrate respiration	Batut and colleagues (1989) and Fischer (1994)
($FixK_2$, FixK)	(Alignment in Supporting Information Fig. S5)	*Bradyrhizobium japonicum* ($FixK_2$), *Rhizobium meliloti* (FixK)		

and S3). The prototypic spacing **C1**–X_2–**C2**–X_5–**C3**–X_{92}–**C4** (Table 1 and Supporting Information Fig. S2) from FNR_{Ec} is present in the FNR proteins of the γ- and β-proteobacteria and represent the largest group of 4Cys FNR_{Ec} proteins. FNR_{Ec} of *E. coli* represents the prototype of 4Cys FNR where all details in O_2 sensing and cluster biochemistry have been studied (Shaw and Guest, 1982; Khoroshilova *et al*., 1995; 1997; Green *et al*., 1996; Crack *et al*., 2017). In addition to FNR_{Ec}, the proteins of other γ-proteobacteria have been verified as FNR_{Ec}-type O_2 sensors, including Fnr or Anr from *Pseudomonas* spp. (Galimand *et al*., 1991; Sawers, 1991; Ibrahim *et al*., 2015), EtrA from *Shewanella* spp. (Maier and Myers, 2001; Cruz-Garcia *et al*., 2011), Fnr from *Vibrio* strains (Septer *et al*., 2010; Kado *et al*., 2017), FnrP of *Pasteurella* (Uhlich *et al*., 1999), HlyX of *Actinobacillus pleuropneumoniae* (MacInnes *et al*., 1990) and Fnr of *Klebsiella pneumoniae* (Grabbe *et al*., 2001) (Table 1). FNR proteins of the β-proteobacteria contain the same Cys spacing as the γ-proteobacteria, but the biochemistry of iron–sulfur cluster has not been studied *in vitro.* The FNR proteins of the β-proteobacteria control the expression of catabolic processes such as anaerobic respiration and fermentation (Table 1) similar to FNR from the γ-proteobacteria and are significant for the virulence of some pathogenic strains like *Bordetella*, *Neisseria* and *Burkholderia* (Bannan *et al*., 1993; Wood *et al*., 1998; Bartolini *et al*., 2006; Edwards *et al*., 2010; Sass *et al*., 2013).

The α-proteobacteria contain 4Cys variants with a slightly different spacing of the C2/C3 and the C3/C4 residues (**C1**–X_2–**C2**–X_7–**C3**–X_{87}–**C4)** (Table 1 and Supporting Information Fig. S2). For some 4Cys FNR_{Ec} variants of the α-proteobacteria, O_2 sensitivity and their capability for complementing FNR_{Ec} function *in vivo* have been verified but detailed biochemical studies on the properties of the modified Cys clusters are missing. The 4Cys FNR_{Ec} homologues of the α-proteobacteria often function as transcriptional activator for genes of anaerobic metabolism many of which are related to bacteria/plant association like nitrate and microaerobic respiration, N_2-fixation, virulence and functionally related genes (Table 1). Representatives of this class have been characterized mainly *in vivo*, such as FnrL of *Rhodobacter capsulatus* (Zeilstra-Ryalls *et al*., 1997), FnrN of *Rhizobium leguminosarum* (Schluter *et al*., 1992), SinR of *Agrobacterium tumefaciens* (Ramey *et al*., 2004), $FixK_1$ of *Bradyrhizobium japonicum* (Anthamatten *et al*., 1992) and AadR of *Rhodpseudomonas palustris (Dispensa et al.,* 1992*).* Remarkably, the 4Cys FNR of the plant pathogenic α-proteobacteria *Xanthomonas* (and *Pseudoxanthomonas*) contain the same Cys cluster indicating a relation of this type of FNR to a plant associated biotope.

The 4Cys FNR_{Ec} proteins of the 'diverse' bacteria group outside the proteobacteria (Fig. 4) differ in the 4Cys cluster slightly from that of the γ/β- and the α-proteobacteria (Table 1 and Supporting Information Fig. S3). The difference in Cys-spacing and their separate phylogenetic clustering stresses their diversification from the proteobacterial FNR_{Ec} proteins. None of the FNR_{Ec}-type proteins of this group has been tested for iron–sulfur cluster properties or function, but the presence of a modified 4Cys cluster suggests that the proteins may be O_2-sensitive transcriptional regulators similar to FNR_{Ec}. Cluster type and the protein sequence give no indication on the phylogenetic origin of this group of FNR_{Ec} proteins.

3Cys variants of FNR_{Ec}

3Cys variants of FNR_{Ec} are scarce (Fig. 3). Most variants (six) are placed within the 'diverse' group non-proteobacteria, and only one further representative is found within the (γ-)proteobacteria (Fig. 3 and Supporting Information Fig. S1). None of the variants has been characterized genetically or biochemically. In the 3Cys variants generally C3 is missing (**C1**–X_2–**C2**–$X_{97–98}$–**C4**; Table 1 and Supporting Information Fig. S4) whereas C1, C2 and C4 are conserved. In some variants, C3 is replaced by a Ser, Asn or Pro residue, or a Cys residue in a modified position. These residues could serve as the fourth ligand for the iron–sulfur cluster, with or without conservation of the function (Muraki *et al*., 2010). The 3Cys variants therefore might be functionally similar to the 4Cys FNR_{Ec} proteins, or represent precursors of other FNR variants.

2Cys variants of FNR_{Ec}

2Cys variants of FNR_{Ec} are present in small numbers (18 representatives in total) in the proteobacterial and the diverse groups (Fig. 3). The 2Cys variants uniformly lack C1. Most of the homologues contain conserved C3 and C4, but other combinations of conserved Cys residues are present as well (Table 1). Thus, in *Burkholderia* (β-proteobacteria) and *Acholeplasma* (Tenericutes), the N-terminal sequence with C1 is deleted, whereas in the variants of *Chromholaobacter* (γ-proteobacteria), *Streptococcus* (bacilli), *Clostridium cellulolyticum* and others, the Cys residues are not conserved.

None of the 2Cys variants has been functionally or biochemically characterized. The Flp proteins from lactic acid bacteria (Gostick *et al*., 1998; Scott *et al*., 2000) serve as models for the function of FNR-like proteins with two Cys residues (Table 1). FlpA from *Lactococcus lactis* is a member of the FNR/CRP family of transcriptional regulators (Korner *et al*., 2003) with low similarity (22% identity with FNR_{Ec}) and is not part of the FNR_{Ec} like protein cluster of Supporting Information Fig. S1. The Cys pair of FlpA (residues C_{15} and C_{112}) assembles in the FlpA dimer an O_2-labile [4Fe–4S] cluster that abolished DNA binding (Scott *et al*., 2000). Flp from *Lactobacillus casei* on the other hand uses the Cys pair for an intramolecular disulphide-dithiol redox switch (Gostick *et al*., 1998). The reactions of the Flp proteins *in vivo* are not clear, however. The Flp proteins respond to oxidative stress and control redox stress reactions, zinc uptake and the arginine deiminase pathway. It is feasible that the 2Cys FNR_{Ec} proteins employ similar reactions or functions.

1Cys variant of FNR-Ec

The 1Cys variants of FNR_{Ec} have high global sequence similarity to FNR_{Ec} and are located within the proteobacteria and *Opitutus terrae* from the diverse group (Fig. 3 and Supporting Information Fig. S1). *Opitutus terrae* and some proteobacteria including *Paracoccus denitrificans* contain only the 1Cys variant of FNR_{Ec}, whereas other strains carry additionally 4Cys FNR_{Ec}. The conserved Cys residue is mostly C4, whereas in some representatives C1 (*Methylobacterium*) or C3 (*O. terrae* and *Variovorax paradoxus*) is conserved. The bacteria with conserved C4 generally lack the N-terminal part of FNR_{Ec} with aa 1–25 and C1 to C3 of FNR_{Ec}.

Presence of a single redox sensitive Cys residue is reminiscent of the *B. subtilis* OhrR protein. The thiolate of Cys15 from OhrR is oxidized (reversibly) to

Cys15-SOH (sulphenic acid) by treatment with hydroperoxide. The oxidation inhibits DNA-binding of OhrR and induces expression of the organic hydroperoxidase gene (Fuangthong and Helmann, 2002). Generally, conservation of single Cys residues suggests redox regulation by reversible protein S-thiolation or thiol-based redox switches (Hillion and Antelmann, 2015; Loi *et al*., 2015).

0Cys variants of FNR-Ec

Most of the 0Cys variants of FNR_{Ec} are located in the α-proteobacteria (Fig. 3) with few examples within the δ-(*Desulfomicrobium*) and β-proteobacteria (*Burkholderia*). Some of the 0Cys variants show deletions in the N-terminal C1 to C3 region of FNR_{Ec}, whereas other cover the region without conservation of the Cys residues. The residual part of the protein is conserved which defines the proteins as FNR_{Ec} homologues.

Prototypes of the 0Cys variants are represented by $FixK_2$ of *Rhodopseudomonas japonicum* and FixK of *Rhizobium meliloti* (Batut *et al*., 1989; Fischer, 1994). $FixK_2$ and FixK are part of an O_2-regulatory cascade, which induces nitrogen fixation and nitrate respiratory genes under anoxic conditions. $FixK_2$ and FixK are no O_2 sensors on their own. Expression of the *fixK*$_2$ and *fix* genes is, however, under the transcriptional regulation of the FixL-FixJ O_2-sensing system that stimulates expression (and function) of $FixK_2$ and FixK under anoxic conditions (for overview see Fischer (1994)). In summary, the 0Cys proteins are members of the FNR_{Ec} family but represent in the form of FixK and $FixK_2$ indirect O_2 sensors without iron–sulfur cluster.

FNR B. subtilis *(FNR_{Bs})*

The search for homologues of FNR_{Bs} yielded 96 hits, which were mostly (82%) in the bacilli phylum (Fig. 1) and among those most (63%) in the genus *Bacillus*. *Bacillus subtilis* FNR_{Bs} coordinates the [4Fe–4S] cluster by Asp D141 and three Cys residues. The Cys residues are located in a C-terminal domain (Gruner *et al*., 2011; Reents *et al*., 2006b) that represents an extension to the CRP homologous region (see Fig. 1). The basic function of this type of FNR in O_2 sensing has been studied for FNR_{Bs} (Gruner *et al*., 2011; Reents *et al*., a,b). FNR_{Bs} and FNR_{Bl} of *Bacillus licheniformis* (Klinger *et al*., 1998; Rey *et al*., 2004) are required for anaerobic induction of the *nar* genes coding for the nitrate respiratory system *narGHJI* and *narK* as well as for *arfM*, a fermentation regulator.

A total of 18% of the FNR_{Bs}-type proteins lack the C-terminal domain and the Cys residues for cluster binding, other variants of the Cys cluster were not identified. Most of the 0Cys variants lack also the conserved Asp residue. The function of the 0-Cys/Asp variants of FNR_{Bs} has not been analysed, but most of the strains are facultatively anaerobic and capable of fermentation or nitrate respiration (Supporting Information Table S1). The 0Cys variants of FNR_{Bs} could represent indirect, iron–sulfur deficient O_2 sensors, such as the 0Cys variants of FNR_{Ec} (compare Table 1), or other regulators of the FNR-CRP family (Korner *et al*., 2003). The FNR_{Bs} variant from aerobic *N. koreensis* has the conserved Asp residue, but lacks the C-terminal Cys cluster of FNR_{Bs}. Instead, it contains, an N-terminal Cys-cluster similar to that of FNR_{Ec}. The protein appears to be an FNR_{Bs} homologue with an FNR_{Bs}/FNR_{Ec} hybrid iron–sulfur binding 3Cys/Asp site.

Fig. 5B shows a matrix for the FNR_{Bs} proteins (80 variants) excluding 0Cys-FNR_{Bs} variants. The occurrence of the FNR_{Bs} homologues is presented in a simplified tree where strains are combined on the genus level (Fig. 5A). FNR_{Bs} is confined to the Bacilli phylum, with few exceptions of homologues in *Clostridium botulinum* (clostridia) and *Selenomonas ruminantium* (negativicutes). FNR_{Bs} is therefore restricted to the firmicutes with strong predominance in the genus *Bacillus.* Homologues are present in most of the Bacilli genera, including *Paenibacillus*, *Geobacillus*, *Exiguobacterium*, *Brevibacillus*, *Lysinibacillus*, *Macrococcus*, *Staphylococcus* and *Anoxybacillus.*

The bacteria of the matrix of Fig. 5 and the corresponding FNR_{Bs} proteins can be functionally sub-grouped by the controlled genes. (i) The typical form represented by FNR_{Bs} of *B. subtilis* and *B. licheniformis* controls anaerobic induction of the nitrate respiratory system (Cruz Ramos *et al*., 1995; Klinger *et al*., 1998; Rey *et al*., 2004; Reents *et al*., 2006a). The branch comprising 17 non-pathogenic strains with *B. subtilis*, *B. licheniformis*, *Bacillus artrophaeus* and the plant growth promoting rhizobacterium *Bacillus amyloliquefaciens* (Chen *et al*., 2007; He *et al*., 2012) and *Bacillus* sp. (Song *et al*., 2012) appears to be part of this class. FNR from the *Geobacilli* and *Bacillus megaterium* could be part of the same cluster due to the presence of nitrate respiration and colocalization with *nar* genes (Feng *et al*., 2007; Muhd Sakaff *et al*., 2012; Brumm *et al*., 2015). Interestingly, *Paenibacillus terrae*, *Paenibacillus polymyxa* and *Paenibacillus mucilaginosus* contain two FNR-like proteins each. (ii) FNR_{Bc} of the pathogenic *Bacillus cereus* modulates under anaerobic conditions glucose fermentation and other catabolic genes in a carbohydrate-dependent manner but is dispensable for nitrate respiration. FNR_{Bc} also activates the expression of enterotoxins (Zigha *et al*., 2007; Messaoudi *et al*., 2010; Esbelin *et al*., 2012). (iii) The FNR homologues of *Bacillus selenitireducens* are more distantly related and not part of the FNR_{Bs} and FNR_{Bc} branches. The bacteria are non-pathogenic and not nitrate respiring (Switzer Blum *et al*., 1998; Eppinger *et al*., 2011), suggesting that their FNR proteins

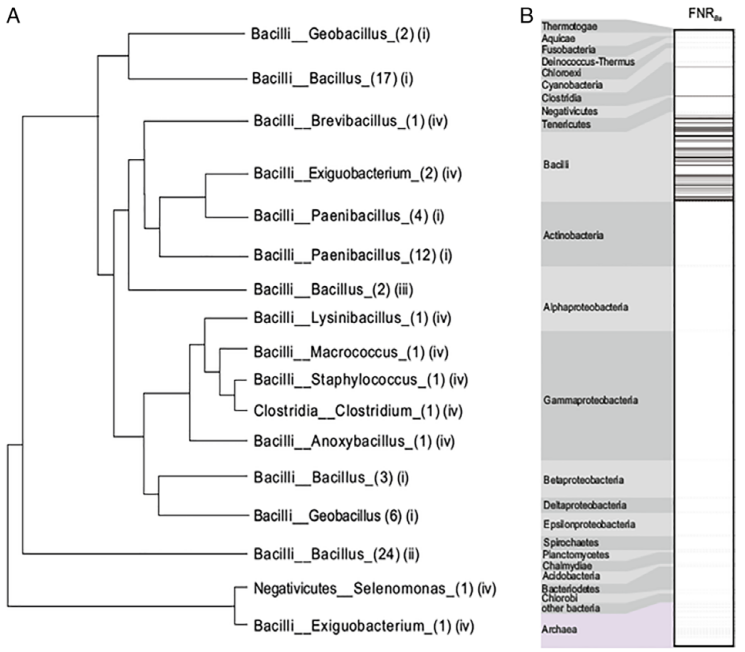


Fig. 5. Schematic tree of *B. subtilis* FNR_{Bs} (A) and matrix for strains carrying FNR_{Bs}-like proteins (B). In (B), strains are listed after removing the 0Cys/D variants (see main text), other details as for Fig. 2. In (A), the clades were collapsed on genus level. The digits in brackets correspond to the amount of FNR homologues in the respective genus.

fulfil a different function in regulation. (iv) The role of FNR_{Bs}-type proteins in bacteria outside the genus *Bacillus* is mostly unknown, e.g., in *Exiguobacterium sibiricum* and *Anoxybacillus flavithermus* (Rodrigues *et al*., 2008; Saw *et al*., 2008). FNR of *A. flavithermus* FNR might control arginine metabolism that of *Macrococcus caseolyticus* and *Staphylococcus pseudintermedius* nitrate respiration (Baba *et al*., 2009). The latter strains also contain a homologue of the NreABC system, which regulates in *Staphylococcus* nitrate respiration in response to O_2 availability (Fedtke *et al*., 2002; Schlag *et al*., 2008). The role of FNR_{Bs} in the bacteria is therefore not clear. The role of FNR_{Bs} in the pathogenic *C. botulinum* and the nonpathogenic *S. ruminantium* is also not known. Both bacteria grow anaerobically and both proteins are most similar to the FNR_{Bs} such as proteins of *S. pseudintermedius* and *Exiguobacterium* with unknown function.

NreB

For *S. aureus* NreB, 73 homologous proteins were identified in the BLAST search. Most (84%) of the homologues occur among the bacilli and of those 54% among the staphylococci, documenting predominance of NreB in bacilli and staphylococci. The sensor kinase NreB coordinates the sensory iron–sulfur cluster by four conserved Cys residues of the N-terminal PAS domain (Fig. 1; Mullner *et al*., 2008). Proteins of the matrix (12% of the homologous) that showed similarity only in the kinase domain of HisKA_3 sensor kinases (Huynh *et al*., 2010) but not in the sensory PAS domain and were deleted in the revised matrix (Fig. 6B). The revised matrix includes 65 NreB-like proteins with similarity including the PAS domain. NreB is a part of the NreABC two component system (*nreA nreB nreC* gene cluster) (Schlag *et al*., 2008). The *nreB* genes of the revised matrix were accompanied mostly by *nreA* and *nreC* genes. Among the 65 strains with NreB-like proteins, 47 encoded the complete Cys cluster for binding the iron–sulfur cluster, 12 with only one of the Cys residues and few others with two (4) or three (2) conserved Cys residues. Most of the Cys variants are found outside the staphylococci or bacilli genera and lack also the *nreA* and *nreC* genes. Figure 6 shows the schematic tree of NreB proteins after combining branches on genus level. Branches were only combined for proteins with the same amount of conserved Cys residues. The Cys residues are essential for the binding of the [4Fe–4S] cluster and function of NreB as an O_2 sensor (Kamps *et al*., 2004; Mullner *et al*., 2008). Therefore conservation of the Cys cluster was inspected as described above for FNR_{Ec}. Notably, all homologues having the complete Cys motif are grouped on one main branch of the tree whereas the variants (highlighted in

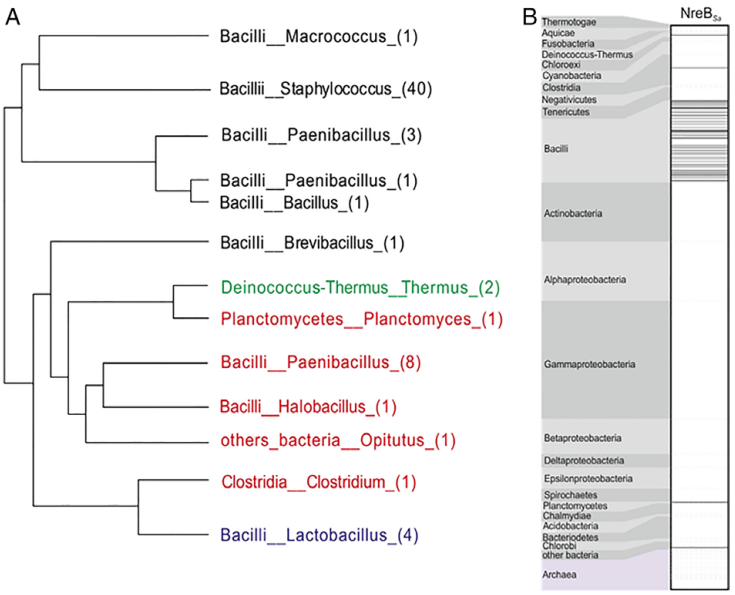


Fig. 6. Schematic tree of *S. aureus* NreB (A) and matrix of strains carrying NreB$_{Sa}$ like proteins (B). In the matrix, the 0Cys variants were deleted (compare main text). Clades were collapsed on genus level. The digits in brackets correspond to the amount of FNR homologues in the respective genus. Homologues with three conserved residues of the [Fe–S]-binding motif are highlighted in green, whereas two cysteine residues are marked in blue and only one cysteine residue is marked in red.

green, blue and red) are dispersed, with *Brevibacillus brevis* as the only exception.

4Cys variants of NreB

Most homologues of NreB (40 from 47 proteins with 4 Cys) are found in the genus *Staphylococcus*. Studies on the regulatory properties and on the biochemistry of the NreABC system were performed in the non-pathogenic *S. carnosus* and in *S. aureus* (Fedtke *et al*., 2002; Mullner *et al*., 2008; Schlag *et al*., 2008; Unden *et al*., 2013).

Some bacteria encode NreB and FNR of the Bs-type in parallel, which are responsible for anaerobic regulation of nitrate respiration in *Staphylococcus* and *Bacillus* respectively (Cruz Ramos *et al*., 1995; LaCelle *et al*., 1996; Fedtke *et al*., 2002; Schlag *et al*., 2008). The differential roles for both sensors in bacteria like *M. caseolyticus* is not known (Baba *et al*., 2009). *Bacillus clausii* encodes NreABC but lacks FNR$_{Bs}$. It was suggested that NreABC regulates expression of nitrate respiration in this *Bacillus* strain (Mullner *et al*., 2008). NreB homologues are found also in some paenibacilli including *Paenibacillus* sp., *P. terrae* and *P. polymyxa*. Their *nreB* gene is located near the *nar* genes. *Paenibacillus* sp. that encodes FNR of the FNR-Bs type and NreABC is capable of nitrate respiration (Mead *et al*., 2012). Gene clustering suggests that NreABC rather than FNR$_{Bs}$ regulates expression of nitrate respiration. *Brevibacillus brevis* representing a member of the Paenibacillaceae (Chen *et al*., 2012) that also carries FNR$_{Bs}$ and NreB, in contrast, lacks genes for the (nitrate sensory component) NreA and for dissimilatory nitrate reduction, suggesting a different role of NreB in the bacteria.

3Cys and 2Cys variants of NreB

Of the 65 NreB-like proteins, six showed variation of the Cys cluster with two or three conserved residues (Fig. 6 and Supporting Information Fig. S6). None of the Cys variants has been studied genetically or biochemically. The 3Cys-homologues of NreB contain an alternative Cys residue close to the N-terminus and may represent 4Cys variants (Supporting Information Fig. S6). The role of the 3Cys NreB of the aerobic *Thermus* species lacking anaerobic nitrate respiration (Henne *et al*., 2004) is not known.

The 2Cys variants are represented by NreB homologues in *Lactobacillus reuteri*, *Lactobacillus plantarum* and *Lactobacillus fermentum* (Fig. 6 and Supporting Information Fig. S6). In the 2Cys variants, only the C-terminal Cys pair is conserved, but the proteins share high sequence similarity with NreB PAS and kinase domains. The lactobacilli possess genes for dissimilatory nitrate reduction (Kleerebezem *et al*., 2003; Morita *et al*., 2008). *Lactobacillus fermentum* and *L. plantarum* also encode the nitrate binding NreA-like protein (Mullner *et al*., 2008; Unden *et al*., 2013). It can be speculated that

a dimer of the 2Cys variant of NreB is required for coordinating one iron–sulfur cluster as hypothesized for nitrogenase NflD subunit in *Methanocaldococcus jannaschii* (Staples *et al*., 2007) or FlpA (Scott *et al*., 2000). Alternatively, the 2Cys variants function by using a regulatory inter- or intra-molecular thiol/disulfide switch (Gostick *et al*., 1998; Scott *et al*., 2000).

1Cys variants of NreB

Twelve NreB variants with only one conserved cysteine residue are present in bacilli (*Paenibacillus* and *Halobacillus*) and *Clostridium* sp., *Planctomyces limnophilus* and *O. terrae*. Within the paenibacilli strains with 4Cys or 1Cys NreB or FNR_{Bs}-like proteins are found, indicating different functions for the variants. The 1Cys-homologues are supposed to fulfil a role in redox-sensing like the 1Cys FNR_{Ec} proteins. C1 or C2 or C4 of the Cys cluster can be conserved. In most systems NreB is not associated with NreA and NreC, which argues against involvement in transcriptional and anaerobic nitrate regulation.

WhiB3 M. tuberculosis

BLAST search identified 124 proteins that are homologues of WhiB3 of *M. tuberculosis*. The WhiB3 homologues were exclusively present in the actinobacteria (Fig. 1). Approximately one-third of the homologues were detected in the mycobacteria, and about 10% belong each to *Streptomyces* and corynebacteria. The remaining homologues are widespread in the actinobacteria including *Rhodococcus, Frankia*, *Nocardia* and *Amycolatopsis* with broad branching of the phylogenetic tree (not shown).

In general, actinobacteria contain multiple WhiB-like proteins which share similar structures including the Cys and the DNA-binding motif but have only moderate sequence similarity. Most mycobacteria possess seven WhiB-like proteins (WhiB1 to B7) which control expression of different gene clusters (Geiman *et al*., 2006; Soliveri *et al*., 2000; Saini *et al*., 2012). WhiB3 to WhiB7 respond to redox stress (O_2 and NO), WhiB1 only to NO, and the function of WhiB2 is redox-independent. WhiB5 is lacking in nonpathogenic mycobacteria (Saini *et al*., 2012). The response of WhiB3 to O_2 is very similar to that of FNR_{Ec} converting the $[4Fe–4S]^{2+}$ cluster first to $[3Fe–4S]^{1+}$ with a concomitant release of Fe^{2+} and one electron which are then used for the two-electron reduction of O_2 to H_2O_2 (Singh *et al*., 2007). In a following step, the $[3Fe–4S]^{1+}$ is converted in a nonredox reaction to $[2Fe–2S]^{2+}$. Overall, the reactions are very similar to those of Step 1 and 2 for FNR_{Ec} (see above) with an overall reaction $[4Fe–4S]^{2+} + O_2 + 2H^+ \rightarrow [2Fe–2S]^{2+} + 2Fe^{3+} + 2S^{2-} + H_2O_2$. According to Singh and colleagues (2007), the H_2O_2 may destroy the [2Fe–2S] cluster.

The structure of WhiB1, which is NO responsive, shows that WhiB1 is a four-helix bundle (Kudhair *et al*., 2017). The core of the protein is formed by three α-helices that are held together by the [4Fe–4S] cluster, which is required for formation of a complex with the major sigma factor σ^A. Reaction of the cluster with NO disassembles the complex and DNA binding as well as gene expression at target genes.

The redox sensor WhiB3 of *M. tuberculosis* controls expression of genes that are important for maintenance of intracellular redox homeostasis as well as virulence, pathogenesis and persistence (Singh *et al*., 2007; 2009). A total of 5% of the WhiB3 homologues lack the first Cys residue of the cluster-binding motif due to a N-terminal deletion of 20–30 amino acids, whereas the residual part of the protein is homologous to $WhiB3_{Mt}$, e.g., in the proteins of *Mycobacterium intracellulare*, *Mycobacterium* sp., *Streptomyces hygroscopicus*, *Amycolicicoccus subflavus* and *Thermonospora curvata*. WhiB4 of *M. intracellulare* having only three cysteine residues was proposed to take over a role in the adaptation to peroxide stress comparable to that of OxyR (Lewis and Falkinham, 2015). WhiB3 homologues with complete cluster-binding motif are found in pathogenic (*Mycobacterium bovis* and *Mycobacterium leprae*) and non-pathogenic (*Mycobacterium smegmatis*) mycobacteria. Function and regulation of WhiB3 appear to be similar in all *Mycobacterium* species.

The WhiD protein of non-pathogenic *Streptomyces coelicolor* is homologous to $WhiB3_{Mt}$ and functions in sporulation and septum formation (Molle *et al*., 2000; Jakimowicz *et al*., 2005). The degradation of native WhiD [4Fe–4S] cluster to apo-WhiD by O_2 or by peroxide stress is very slow, and no [2Fe–2S] cluster is formed as intermediate (Crack *et al*., 2009). *In vitro* the [4Fe–4S] cluster degrades similar to that of FNR_{Ec} with a [2Fe–2S] intermediate (Jakimowicz *et al*., 2005). Therefore, WhiD, such as WhiB3, acts as redox sensor but regulates different cellular responses. Nonpathogenic *Corynebacterium glutamicum* only possesses four WhiB proteins called WhcA, WhcB, WhcD and WhcE that regulate various cell functions in response to oxidative stress (Kim *et al*., 2005; Choi *et al*., 2009; Lee *et al*., 2012, 2018).

Overall, WhiB-like proteins are restricted to actinobacteria, which are all aerobic, indicating the evolution of WhiB proteins with the [4Fe–4S] cluster after accumulation of oxygen on earth.

Discussion

Origin and relation of the sensor proteins FNR_{Ec}, FNR_{Bs}, NreB and WhiB3

Sequence comparison and distribution indicate ancient origin and independent evolution of FNR_{Ec}, FNR_{Bs}, NreB and WhiB3/D. FNR_{Ec} and FNR_{Bs} proteins are derived

from the ancient carbon-regulator CRP (Saier *et al*., 1996). The global identity between FNR and CRP is low (17.9% between FNR_{Ec} and CRP_{Ec}), and the proteins fall for this reason into separate clusters and alignments. The role of CRP as the precursor and the FNR proteins as the derivatives are suggested by the broad distribution of CRP in many phyla and its full-length conservation in FNR_{Ec} and FNR_{Bs}. Korner and colleagues (2003) and Green and colleagues (2001) describe the CRP-FNR family of transcriptional regulators that includes in addition to CRP and FNR also other redox regulators like YeiI and Flp. The latter have low global identity with FNR and CRP and are not considered in the present study.

FNR_{Ec} and FNR_{Bs} are distinguished from CRP by the presence of short N- and C-terminal sequences, respectively, containing each three Cys ligands for ligation of the iron–sulfur clusters. Together with an internal fourth Cys (C4) or Asp residue, the Cys residues of the terminal extensions ligate the [4Fe–4S] cluster in FNR_{Ec} or FNR_{Bs}. In addition, FNR_{Ec} and FNR_{Bs} differ by the mode of signal output. In FNR_{Ec}, the signal output is produced by monomerization of the protein after reaction with O_2 and the oligomerization state then affects DNA binding (Lazazzera *et al*., 1993). Within the CRP/FNR family of transcriptional regulators, inactivation by monomerization is unique for FNR_{Ec} whereas other members such as CRP (Anderson *et al*., 1971; Takahashi *et al*., 1980) and FNR_{Bs} retain their dimeric state in the presence of their effector or in the active or inactive state respectively. Thus, FNR_{Bs} is a permanent dimer and DNA binding is regulated by conformational changes within FNR_{Bs} (Reents *et al*., 2006b). The broad overlap of co-occurrence of CRP with FNR_{Ec} in the γ-proteobacteria suggests the origin for FNR_{Ec} within this group.

NreB belongs to the large family of histidine sensor kinases that show broad variation in their domain composition. In NreB and related His kinases sensing occurs by an N-terminal PAS domain. The PAS domain of the closely related FixL sensor kinase binds hemeB at a position homologous to the [4Fe–4S]-binding site of NreB (Mullner *et al*., 2008; Unden *et al*., 2013). Remarkably, both NreB and FixL represent direct O_2 sensors. NreB and FixL are present, however, essentially within the Bacilli and the proteobacteria respectively. Therefore, the hemeB/[4Fe–4S]-binding sites appear to have a common origin and an occasional but rare lateral transfer to other bacterial phyla could be detected.

The WhiB3/D proteins are restricted to actinobacteria. The actinobacteria contain no FNR_{Ec}, FNR_{Bs} or NreB-type O_2 regulators but a significant variation in the WhiB3/D regulators, which take over multiple functions. Their unique presence in aerobic actinobacteria suggests development of the WhiB3/D proteins after separation or development of the group and after establishment of oxic conditions.

Recent reports for the structures of FNR_{Ec}- and Wbl-type proteins show binding and interaction of the [4Fe–4S] clusters with the respective proteins (Kudhair *et al*., 2017; Volbeda *et al*., 2015). In FNR_{Ec}, binding of the [4Fe–4S] cluster is achieved by a pocket formed from two α-helices, one β-sheet and the N-terminal loop of FNR whereas in WhiB1 proteins the [4Fe–4S] cluster is coordinated by three α-helices that are part of a four-helix bundle. WhiB1 is basically of the NO sensing type, but the overall structure is conserved in the Wbl/Whi family. NreB accommodates the cluster in a PAS domain that binds the [4Fe–4S] cluster by the α-helical PAS core and the α-helical connector which links this region to the β-scaffold and the kinase domain (Miyatake *et al*., 2000; Mullner *et al*., 2008). Therefore, the structural arrangements around the iron–sulfur clusters and their binding share no obvious similarity apart from the Cys ligands, which is again strong indication for their independent origin and evolution.

Functional diversity and diversification

Despite general restriction of FNR_{Ec}, FNR_{Bs}, NreB and WhiB3/D to diverse bacterial phyla, some bacteria contain multiple sensor proteins of these types in various combinations (Table 2). (i) Some bacteria contain multiple sensors of the same type. Thus *Pseudomonas putida* comprises three FNR-Ec type proteins that differ in their O_2 sensitivity (Ibrahim *et al*., 2015). Actinobacteria contain up to seven WhiB3/D proteins that differ in their response to O_2 but also in the target promoters for the individual WhiB3/D proteins (Molle *et al*., 2000; Steyn *et al*., 2002; Jakimowicz *et al*., 2005; Kim *et al*., 2005; Choi *et al*., 2009; Crack *et al*., 2009; Lee *et al*., 2012, 2018). There are also strains such as *P. mucilaginosus* with more than one 4Cys-FNR_{Bs} variant. (ii) Other bacteria such as *B. japonicum* and *R. meliloti* contain multiple variants of the same type of regulator, such as AadR (4Cys FNR_{Ec}, a direct O_2 sensor) and $FixK_2$ or FixK (0Cys FNR_{Ec}, an indirect O_2 regulator (Fischer, 1994) (Table 1 and Supporting Information Table S4). (iii) Other strains contain different types of predicted O_2 sensors such as 4Cys-FNR_{Bs} and 4Cys-NreB that are predicted to form functional O_2 sensors. The function and dual roles in O_2 sensing have not been verified, however. This list is extended by bacteria with 4Cys-FNR_{Bs} (or 4Cys-FNR_{Ec}) proteins that are combined with 1Cys- or 0Cys-variants of an alternative sensor (FNR_{Bs}, FNR_{Ec} and NreB). (iv) Other strains contain FNR hybrid sensors with FNR_{Ec} /FNR_{Bs} hybrid Cys clusters. Thus the FNR_{Bs} homologous protein of *Clostridium perfringens* lacks the D/Cys cluster of FNR_{Bs}-type sensors but contains an N-terminal Cys-cluster of the FNR_{Ec}-type including the central C4 residue. The spacing of the 4 Cys residues is slightly different from that of FNR_{Ec}.

Table 2. Species containing multiple forms of FNR_{Ec}, FNR_{Bs} and NreB.

Species	FNR_{Ec}	FNR_{Bs}	NreB
(i) Multiple sensors within one species			
Pseudomonas (γ)	4Cys[a] (3x)		
Paenibacillus spp. (B)		4Cys (2x)	
(ii) Multiple variants of one sensor type			
Bradyrhizobium japonicum (α)	4Cys[a], 0Cys[a]		
Rhizobium meliloti	4Cys[a], 0Cys[a]		
Rhodoferax ferrireducens (β)	4Cys[a], 1Cys		
Paenibacillus spp. (B)			4Cys[a], 1Cys
(iii) Different sensor classes in one species			
M. caseolyticus (B)	–	4Cys	4Cys[a]
Paenibacillus sp. JDR2	–	4Cys	1Cys
S. ruminantium (Neg)	–	4Cys[a]	0Cys
D. hafniense DCB2	4Cys	0Cys	–
Shewanella piezotolerans (γ)	4Cys[a]	–	0Cys
(iv) Hybrid			
C. perfringens str 13	–	No D/Cys, but FNR-Ec type N-terminal Cys cluster with variant spacing C5–x_4–C–x_3–C–x_{98}–C113	

The type of FNR or NreB (4Cys, 1Cys or 0Cys variant) is indicated. Abbreviations for the bacterial phyla in the species column: B, bacilli, Neg, negativicutes; C, clostridia; α, β or γ, α-, β- or γ-proteobacteria.
[a]Homologues of FNR or NreB that are supposed to regulate anaerobic (nitrate) respiration.

Origin of the [4Fe–4S] clusters and [4Fe–4S]/[2Fe–2S] cluster conversion in O_2 sensing: the role of the cluster biochemistry

The FNR_{Ec}, FNR_{Bs}, WhiB3/D and NreB proteins use $[4Fe–4S]^{2+}$ clusters for O_2 sensing (Khoroshilova *et al*., 1997; Jakimowicz *et al*., 2005; Reents *et al*., 2006b; Singh *et al*., 2007; Crack *et al*., 2009; 2017; Mullner *et al*., 2008; Zhang *et al*., 2012). For FNR_{Ec}, NreB and to some extent also for WhiB3/D, the response to O_2 appears to be similar by the conversion to a $[2Fe–2S]^{2+}$ cluster, despite the use of unrelated protein types and Cys clusters for ligation that differ in spacing of the Cys residues and sequence (Fig. 1). It is not known whether the similarity extends to details such as the formation of an [4Fe–3S] intermediate and of the Cys persulfides as described for FNR_{Ec} (Crack *et al*., 2017).

Cluster conversion is the basis for signal transmission to the output domains. The response triggers conversion of dimeric to monomeric FNR_{Ec} and loss of specific DNA binding (Lazazzera *et al*., 1996; Volbeda *et al*., 2015), activation of the kinase domain in NreB (Mullner *et al*., 2008; Nilkens *et al*., 2014) and loss of DNA binding by the permanent dimer FNR_{Bs} (Reents *et al*., 2006b). It appears therefore that the response of the iron–sulfur cluster depends to a large extent on the chemistry and properties of the cluster rather than on the protein, whereas the signal output is governed by the surrounding protein.

Iron–sulfur clusters can be produced in protic and aproteinogenic systems, and $[4Fe–4S]^{2+}$ and $[2Fe–2S]^{2+}$ clusters have been produced under anoxic conditions in the thiol-ligated form with various thiol ligands (for review see Venkateswara Rao and Holm, 2004). The cubane-type $[4Fe–4S]^{2+}$-thiolate clusters are very common. Formation of the $[4Fe–4S]^{2+}$ cluster represents a thermodynamic sink and the $[4Fe–4S]^{2+}$ cluster is among the most stable under anoxic conditions (Ogino *et al*., 1998; Venkateswara Rao and Holm, 2004). The clusters can be incorporated in cysteinyl peptides (Ohno *et al*., 1991; Ueyama *et al*., 1985). The [2Fe–2S] cluster, in particular the more stable $[2Fe–2S]^{2+}$ form, can also be generated in many thiolates and cysteinyl peptides (references in (Hagen *et al*., 1983; Venkateswara Rao and Holm, 2004). The $[2Fe–2S]^{1+}$ cluster is less stable, but can be produced by reduction of the $[2Fe–2S]^{2+}$ cluster. Biotin and lipoate synthases, anaerobic ribonucleotide reductase and the pyruvate formate-lyase activating enzyme contain [4Fe–4S] clusters that serve as electron donors in the radical catalysis of the enzymes (Buis and Broderick, 2005; Duin *et al*., 1997; Ollagnier-De Choudens *et al*., 2000; Ugulava *et al*., 2000; Lotierzo *et al*., 2005). The enzymes are sensitive to inactivation by O_2 that causes a side reaction with $[4Fe–4S]^{2+}$ to $[2Fe–2S]^{2+}$ conversion (compare Venkateswara Rao and Holm, 2004), resembling the reaction at the iron–sulfur cluster of FNR_{Ec}. Thus, the information from cluster (bio)chemistry is compatible with the observation that $[4Fe–4S]^{2+}$ and $[2Fe–2S]^{2+}$ clusters are present and interconverted in different protein environments. The conformational changes in cluster conversion can then be used for largely different responses, depending on the surrounding protein. The $[4Fe–4S]^{2+}$ clusters therefore represent a unique molecular device that can be modulated in evolution to bring forth O_2 sensors in different protein background and different output reactions. Remarkably, this mode of O_2 sensing that has been used for

developing independent lines of O_2 sensors in bacteria is apparently missing in archaea and eukaryotes.

Experimental procedures

Sequences

Protein sequences involved in oxygen sensing and containing a [4Fe–4S] cluster were downloaded in July 2016 from UniProt (Consortium, 2017) and GenBank (Benson *et al*., 2005): fumarate and nitrate reduction regulatory protein (FNR) from *E. coli* (UniProt: P0A9E5) and *B. subtilis* (GenBank: KIX83509), oxygen sensor histidine kinase NreB from *S. aureus* (GenBank: EFW34334), redox-responsive transcriptional regulator WhiB3 from *M. tuberculosis* (GenBank: KPU49338) and cAMP-activated global transcriptional regulator (CRP) from *E. coli* (UniProt: P0ACJ8). Furthermore, all protein sequences of 1.981 complete prokaryotic genomes were downloaded from the NCBI RefSeq database (version June 2012) and clustered into protein families as previously described (Nelson-Sathi *et al*., 2015; Weiss *et al*., 2016).

Identification of homologous protein families

Homologous sequences were identified by sequence comparisons of the five reference sequences involved in oxygen sensing, without any weighting of specific residues or motifs, with all 6.1 million proteins of the 1.981 complete prokaryotic genomes using BLASTp (Altschul *et al*., 1997) with an *E*-value threshold ≤ 10^{-10} and local identity cutoff ≥ 25%. Remaining BLAST hits with global amino acid identity not smaller than 25%, calculated with needle from EMBOSS 6.6.0 (Rice *et al*., 2000), were subsequently compared to the predefined protein families. The protein family showing the highest number of BLAST hits for the respective reference sequence was identified as the homologous protein family.

Identification of the number of cysteine residues

Conserved cysteine residues linked to the [4Fe–4S] cluster were counted manually in sequence comparisons performed with Clustal Omega v1.2.4 (Sievers *et al*., 2011) for each protein. Because of their low sequence conservation, *B. subtilis* FNR- and NreB-like proteins without conserved cysteine residues were excluded alignments used to generate trees.

Multiple sequence alignment and phylogenetic tree reconstruction

In the identified homologous protein family for FNR (*E. coli*), for all sequences showing global identity ≥ 80% on the species level, only the longest sequence was retained in the protein family. Sequences in each identified homologous protein family were aligned together with the respective reference sequence using MAFFT v7.299b (Katoh and Standley, 2013) with the options *–localpair, –maxiterate = 1000* and *–anysymbol*. Maximum likelihood trees were reconstructed using RAxML v8.2.8 (Stamatakis, 2014) using the PROTCATWAG model. The clades in the trees for NreB and FNR (*B. subtilis*) were collapsed on the genus level using Figtree v1.3.1, when the corresponding [4Fe–4S] binding motif contained the same number of cysteine residues. No schematic tree was generated for WhiB3, where all sequences of the protein family belong to the taxonomic group actinobacteria.

Acknowledgements

Financial support by Deutsche Forschungsgemeinschaft to GU (grant UN 49/18-1) and by the European Research Council (grant 666053 to WM) as well as the Volkswagen Foundation (Grant 93046 to WM) is gratefully acknowledged.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Anderson, W. B., Schneider, A. B., Emmer, M., Perlman, R. L., and Pastan, I. (1971) Purification of and properties of the cyclic adenosine 3′,5′-monophosphate receptor protein which mediates cyclic adenosine 3′,5′-monophosphate-dependent gene transcription in Escherichia coli. *J Biol Chem* **246**: 5929–5937.

Anthamatten, D., Scherb, B., and Hennecke, H. (1992) Characterization of a fixLJ-regulated *Bradyrhizobium japonicum* gene sharing similarity with the Escherichia coli fnr and *Rhizobium meliloti* fixK genes. *J Bacteriol* **174**: 2111–2120.

Baba, T., Kuwahara-Arai, K., Uchiyama, I., Takeuchi, F., Ito, T., and Hiramatsu, K. (2009) Complete genome sequence of *Macrococcus caseolyticus* strain JCSCS5402, [corrected] reflecting the ancestral genome of the human-pathogenic staphylococci. *J Bacteriol* **191**: 1180–1190.

Bannan, J. D., Moran, M. J., MacInnes, J. I., Soltes, G. A., and Friedman, R. L. (1993) Cloning and characterization of btr, a *Bordetella pertussis* gene encoding an FNR-like transcriptional regulator. *J Bacteriol* **175**: 7228–7235.

Bartolini, E., Frigimelica, E., Giovinazzi, S., Galli, G., Shaik, Y., Genco, C., *et al*. (2006) Role of FNR and FNR-regulated, sugar fermentation genes in *Neisseria meningitidis* infection. *Mol Microbiol* **60**: 963–972.

Batut, J., and Boistard, P. (1994) Oxygen control in *Rhizobium*. *Antonie Van Leeuwenhoek* **66**: 129–150.

Batut, J., Daveran-Mingot, M. L., David, M., Jacobs, J., Garnerone, A. M., and Kahn, D. (1989) fixK, a gene homologous with fnr and crp from Escherichia coli, regulates nitrogen fixation genes both positively and negatively in *Rhizobium meliloti*. *EMBO J* **8**: 1279–1286.

Becker, S., Holighaus, G., Gabrielczyk, T., and Unden, G. (1996) O2 as the regulatory signal for FNR-dependent gene regulation in *Escherichia coli*. *J Bacteriol* **178**: 4515–4521.

Becker, S., Vlad, D., Schuster, S., Pfeiffer, P., and Unden, G. (1997) Regulatory O2 tensions for the synthesis of fermentation products in *Escherichia coli* and relation to aerobic respiration. *Arch Microbiol* **168**: 290–296.

Beinert, H. (1976) Iron-sulfur proteins, the most numerous and most diversified components of the mitochondrial electron transfer system. *Adv Exp Med Biol* **74**: 137–149.

Beinert, H., Kennedy, M. C., and Stout, C. D. (1996) Aconitase as Ironminus signSulfur protein, enzyme, and iron-regulatory protein. *Chem Rev* **96**: 2335–2374.

Beinert, H., Holm, R. H., and Munck, E. (1997) Iron-sulfur clusters: nature's modular, multipurpose structures. *Science* **277**: 653–659.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2005) GenBank. *Nucleic Acids Res* **33**: D34–D38.

Brekasis, D., and Paget, M. S. (2003) A novel sensor of NADH/NAD+ redox poise in *Streptomyces coelicolor* A3(2). *EMBO J* **22**: 4856–4865.

Brumm, P. J., Land, M. L., and Mead, D. A. (2015) Complete genome sequence of *Geobacillus thermoglucosidasius* C56-YS93, a novel biomass degrader isolated from obsidian hot spring in Yellowstone National Park. *Stand Genomic Sci* **10**: 73.

Buis, J. M., and Broderick, J. B. (2005) Pyruvate formate-lyase activating enzyme: elucidation of a novel mechanism for glycyl radical formation. *Arch Biochem Biophys* **433**: 288–296.

Chen, X. H., Koumoutsi, A., Scholz, R., Eisenreich, A., Schneider, K., Heinemeyer, I., *et al*. (2007) Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Nat Biotechnol* **25**: 1007–1014.

Chen, W., Wang, Y., Li, D., Li, L., Xiao, Q., and Zhou, Q. (2012) Draft genome sequence of *Brevibacillus brevis* strain X23, a biocontrol agent against bacterial wilt. *J Bacteriol* **194**: 6634–6635.

Choi, W. W., Park, S. D., Lee, S. M., Kim, H. B., Kim, Y., and Lee, H. S. (2009) The whcA gene plays a negative role in oxidative stress response of *Corynebacterium glutamicum*. *FEMS Microbiol Lett* **290**: 32–38.

Commichau, F. M., and Stulke, J. (2008) Trigger enzymes: bifunctional proteins active in metabolism and in controlling gene expression. *Mol Microbiol* **67**: 692–702.

Consortium, T. U. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**: D158–d169.

Crack, J. C., Green, J., Le Brun, N. E., and Thomson, A. J. (2006) Detection of sulfide release from the oxygen-sensing [4Fe–4S] cluster of FNR. *J Biol Chem* **281**: 18909–18913.

Crack, J. C., Green, J., Cheesman, M. R., Le Brun, N. E., and Thomson, A. J. (2007) Superoxide-mediated amplification of the oxygen-induced switch from [4Fe–4S] to [2Fe–2S] clusters in the transcriptional regulator FNR. *Proc Natl Acad Sci U S A* **104**: 2092–2097.

Crack, J. C., Gaskell, A. A., Green, J., Cheesman, M. R., Le Brun, N. E., and Thomson, A. J. (2008) Influence of the environment on the [4Fe–4S]2+ to [2Fe–2S]2+ cluster switch in the transcriptional regulator FNR. *J Am Chem Soc* **130**: 1749–1758.

Crack, J. C., den Hengst, C. D., Jakimowicz, P., Subramanian, S., Johnson, M. K., Buttner, M. J., *et al*. (2009) Characterization of [4Fe–4S]-containing and cluster-free forms of *Streptomyces* WhiD. *Biochemistry* **48**: 12252–12264.

Crack, J. C., Thomson, A. J., and Le Brun, N. E. (2017) Mass spectrometric identification of intermediates in the O2-driven [4Fe–4S] to [2Fe–2S] cluster conversion in FNR. *Proc Natl Acad Sci U S A* **114**: E3215–E3223.

Cruz Ramos, H., Boursier, L., Moszer, I., Kunst, F., Danchin, A., and Glaser, P. (1995) Anaerobic transcription activation in *Bacillus subtilis*: identification of distinct FNR-dependent and -independent regulatory mechanisms. *EMBO J* **14**: 5984–5994.

Cruz-Garcia, C., Murray, A. E., Rodrigues, J. L., Gralnick, J. A., McCue, L. A., Romine, M. F., *et al*. (2011) Fnr (EtrA) acts as a fine-tuning regulator of anaerobic metabolism in *Shewanella oneidensis* MR-1. *BMC Microbiol* **11**: 64.

Dispensa, M., Thomas, C. T., Kim, M. K., Perrotta, J. A., Gibson, J., and Harwood, C. S. (1992) Anaerobic growth of *Rhodopseudomonas palustris* on 4-hydroxybenzoate is dependent on AadR, a member of the cyclic AMP receptor protein family of transcriptional regulators. *J Bacteriol* **174**: 5803–5813.

Duin, E. C., Lafferty, M. E., Crouse, B. R., Allen, R. M., Sanyal, I., Flint, D. H., and Johnson, M. K. (1997) [2Fe–2S] to [4Fe–4S] cluster conversion in *Escherichia coli* biotin synthase. *Biochemistry* **36**: 11811–11820.

Edwards, J., Cole, L. J., Green, J. B., Thomson, M. J., Wood, A. J., Whittingham, J. L., and Moir, J. W. (2010) Binding to DNA protects *Neisseria meningitidis* fumarate and nitrate reductase regulator (FNR) from oxygen. *J Biol Chem* **285**: 1105–1112.

Engelhardt, W. A. (1974) On the dual role of respiration. *Mol Cell Biochem* **5**: 25–33.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002) An ancient algorithm for large scale detection of protein families. *Nucleic Acids Res* **30**: 1575–1584.

Eppinger, M., Bunk, B., Johns, M. A., Edirisinghe, J. N., Kutumbaka, K. K., Koenig, S. S., *et al*. (2011) Genome sequences of the biotechnologically important *Bacillus megaterium* strains QM B1551 and DSM319. *J Bacteriol* **193**: 4199–4213.

Esbelin, J., Jouanneau, Y., and Duport, C. (2012) *Bacillus cereus* Fnr binds a [4Fe–4S] cluster and forms a ternary complex with ResD and PlcR. *BMC Microbiol* **12**: 125.

Fedtke, I., Kamps, A., Krismer, B., and Gotz, F. (2002) The nitrate reductase and nitrite reductase operons and the narT gene of *Staphylococcus carnosus* are positively controlled by the novel two-component system NreBC. *J Bacteriol* **184**: 6624–6634.

Feng, L., Wang, W., Cheng, J., Ren, Y., Zhao, G., Gao, C., *et al*. (2007) Genome and proteome of long-chain alkane

degrading *Geobacillus thermodenitrificans* NG80-2 isolated from a deep-subsurface oil reservoir. *Proc Natl Acad Sci U S A* **104**: 5602–5607.
Fischer, H. M. (1994) Genetic regulation of nitrogen fixation in rhizobia. *Microbiol Rev* **58**: 352–386.
Fischer, W. W., Hemp, J., and Johnson, J. E. (2016) Evolution of oxygenic photosynthesis. *Annu Rev Earth Planet Sci* **44**: 647–683.
Fuangthong, M., and Helmann, J. D. (2002) The OhrR repressor senses organic hydroperoxides by reversible formation of a cysteine-sulfenic acid derivative. *Proc Natl Acad Sci U S A* **99**: 6690–6695.
Galimand, M., Gamper, M., Zimmermann, A., and Haas, D. (1991) Positive FNR-like control of anaerobic arginine degradation and nitrate respiration in *Pseudomonas aeruginosa*. *J Bacteriol* **173**: 1598–1606.
Geiman, D. E., Raghunand, T. R., Agarwal, N., and Bishai, W. R. (2006) Differential gene expression in response to exposure to antimycobacterial agents and other stress conditions among seven *Mycobacterium tuberculosis* whiB-like genes. *Antimicrob Agents Chemother* **50**: 2836–2841.
Gostick, D. O., Green, J., Irvine, A. S., Gasson, M. J., and Guest, J. R. (1998) A novel regulatory switch mediated by the FNR-like protein of lactobacillus casei. *Microbiology* **144**: 705–717.
Grabbe, R., Klopprogge, K., and Schmitz, R. A. (2001) Fnr is required for NifL-dependent oxygen control of nif gene expression in *Klebsiella pneumoniae*. *J Bacteriol* **183**: 1385–1393.
Green, J., and Paget, M. S. (2004) Bacterial redox sensors. *Nat Rev Microbiol* **2**: 954–966.
Green, J., Sharrocks, A. D., Green, B., Geisow, M., and Guest, J. R. (1993) Properties of FNR proteins substituted at each of the five cysteine residues. *Mol Microbiol* **8**: 61–68.
Green, J., Bennett, B., Jordan, P., Ralph, E. T., Thomson, A. J., and Guest, J. R. (1996) Reconstitution of the [4Fe–4S] cluster in FNR and demonstration of the aerobic-anaerobic transcription switch in vitro. *Biochem J* **316**: 887–892.
Green, J., Scott, C., and Guest, J. R. (2001) Functional versatility in the CRP-FNR superfamily of transcription factors: FNR and FLP. *Adv Microb Physiol* **44**: 1–34.
Gruner, I., Fradrich, C., Bottger, L. H., Trautwein, A. X., Jahn, D., and Hartig, E. (2011) Aspartate 141 is the fourth ligand of the oxygen-sensing [4Fe–4S]2+ cluster of *Bacillus subtilis* transcriptional regulator Fnr. *J Biol Chem* **286**: 2017–2021.
Hagen, K. S., Watson, A. D., and Holm, R. H. (1983) Synthetic routes to iron sulfide (Fe2S2, Fe3S4, Fe4S4, and Fe6S9), clusters from the common precursor tetrakis(ethanethiolate)ferrate(2-) ion ([Fe(SC2H5)4]2-): structures and properties of [Fe3S4(SR)4]3- and bis(ethanethiolate)nonathioxohexaferrate(4-) ion ([Fe6S9(SC2H5)2]4-), examples of the newest types of Fe-S-SR clusters. *J Am Chem Soc* **105**: 3905–3913.
He, P., Hao, K., Blom, J., Ruckert, C., Vater, J., Mao, Z., *et al*. (2012) Genome sequence of the plant growth promoting strain *Bacillus amyloliquefaciens* subsp. plantarum B9601-Y2 and expression of mersacidin and other secondary metabolites. *J Biotechnol* **164**: 281–291.
Henne, A., Bruggemann, H., Raasch, C., Wiezer, A., Hartsch, T., Liesegang, H., *et al*. (2004) The genome sequence of the extreme thermophile *Thermus thermophilus*. *Nat Biotechnol* **22**: 547–553.
Hillion, M., and Antelmann, H. (2015) Thiol-based redox switches in prokaryotes. *Biol Chem* **396**: 415–444.
Holland, H. D. (2006) The oxygenation of the atmosphere and oceans. *Philos Trans R Soc Lond B: Biol Sci* **361**: 903–915.
Huynh, T. N., Noriega, C. E., and Stewart, V. (2010) Conserved mechanism for sensor phosphatase control of two-component signaling revealed in the nitrate sensor NarX. *Proc Natl Acad Sci U S A* **107**: 21140–21145.
Ibrahim, S. A., Crack, J. C., Rolfe, M. D., Borrero-de Acuna, J. M., Thomson, A. J., Le Brun, N. E., *et al*. (2015) Three *Pseudomonas putida* FNR family proteins with different sensitivities to O2. *J Biol Chem* **290**: 16812–16823.
Imlay, J. A. (2002) How oxygen damages microbes: oxygen tolerance and obligate anaerobiosis. *Adv Microb Physiol* **46**: 111–153.
Imlay, J. A. (2006) Iron–sulphur clusters and the problem with oxygen. *Mol Microbiol* **59**: 1073–1082.
Jakimowicz, P., Cheesman, M. R., Bishai, W. R., Chater, K. F., Thomson, A. J., and Buttner, M. J. (2005) Evidence that the *Streptomyces* developmental protein WhiD, a member of the WhiB family, binds a [4Fe–4S] cluster. *J Biol Chem* **280**: 8309–8315.
Kado, T., Kashimoto, T., Yamazaki, K., and Ueno, S. (2017) Importance of fumarate and nitrate reduction regulatory protein for intestinal proliferation of *Vibrio vulnificus*. *FEMS Microbiol Lett* **364**: pii: fnw274. doi: 10.1093/femsle/fnw274.
Kamps, A., Achebach, S., Fedtke, I., Unden, G., and Gotz, F. (2004) Staphylococcal NreB: an O(2)-sensing histidine protein kinase with an O(2)-labile iron-sulphur cluster of the FNR type. *Mol Microbiol* **52**: 713–723.
Kaptain, S., Downey, W. E., Tang, C., Philpott, C., Haile, D., Orloff, D. G., *et al*. (1991) A regulated RNA binding protein also possesses aconitase activity. *Proc Natl Acad Sci U S A* **88**: 10109–10113.
Katoh, K., and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.
Khoroshilova, N., Beinert, H., and Kiley, P. J. (1995) Association of a polynuclear iron-sulfur center with a mutant FNR protein enhances DNA binding. *Proc Natl Acad Sci U S A* **92**: 2499–2503.
Khoroshilova, N., Popescu, C., Munck, E., Beinert, H., and Kiley, P. J. (1997) Iron–sulfur cluster disassembly in the FNR protein of *Escherichia coli* by O2: [4Fe–4S] to [2Fe–2S] conversion with loss of biological activity. *Proc Natl Acad Sci USA* **94**: 6087–6092.
Kiley, P. J., and Beinert, H. (1998) Oxygen sensing by the global regulator, FNR: the role of the iron-sulfur cluster. *FEMS Microbiol Rev* **22**: 341–352.
Kim, T. H., Park, J. S., Kim, H. J., Kim, Y., Kim, P., and Lee, H. S. (2005) The whcE gene of *Corynebacterium glutamicum* is important for survival following heat and oxidative stress. *Biochem Biophys Res Commun* **337**: 757–764.
Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O. P., Leer, R., *et al*. (2003) Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci USA* **100**: 1990–1995.

Klinger, A., Schirawski, J., Glaser, P., and Unden, G. (1998) The fnr gene of *Bacillus licheniformis* and the cysteine ligands of the C-terminal FeS cluster. *J Bacteriol* **180**: 3483–3485.

Korner, H., Sofia, H. J., and Zumft, W. G. (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiol Rev* **27**: 559–592.

Kudhair, B. K., Hounslow, A. M., Rolfe, M. D., Crack, J. C., Hunt, D. M., Buxton, R. S., *et al*. (2017) Structure of a Wbl protein and implications for NO sensing by *M. tuberculosis*. *Nat Commun* **8**: 2280.

LaCelle, M., Kumano, M., Kurita, K., Yamane, K., Zuber, P., and Nakano, M. M. (1996) Oxygen-controlled regulation of the flavohemoglobin gene in *Bacillus subtilis*. *J Bacteriol* **178**: 3803–3808.

Lancaster, C. R., Kroger, A., Auer, M., and Michel, H. (1999) Structure of fumarate reductase from *Wolinella succinogenes* at 2.2 A resolution. *Nature* **402**: 377–385.

Lazazzera, B. A., Bates, D. M., and Kiley, P. J. (1993) The activity of the *Escherichia coli* transcription factor FNR is regulated by a change in oligomeric state. *Genes Dev* **7**: 1993–2005.

Lazazzera, B. A., Beinert, H., Khoroshilova, N., Kennedy, M. C., and Kiley, P. J. (1996) DNA binding and dimerization of the Fe–S-containing FNR protein from *Escherichia coli* are regulated by oxygen. *J Biol Chem* **271**: 2762–2768.

Lee, J. Y., Park, J. S., Kim, H. J., Kim, Y., and Lee, H. S. (2012) Corynebacterium glutamicum whcB, a stationary phase-specific regulatory gene. *FEMS Microbiol Lett* **327**: 103–109.

Lee, D.S., Kim, P., Kim, E.S., Kim, Y., and Lee, H.S. (2018) *Corynebacterium glutamicum* WhcD interacts with WhiA to exert a regulatory effect on cell division genes. *Antonie Van Leeuwenhoek* 111: 641–648.

Lenton, T. M., Dahl, T. W., Daines, S. J., Mills, B. J., Ozaki, K., Saltzman, M. R., and Porada, P. (2016) Earliest land plants created modern levels of atmospheric oxygen. *Proc Natl Acad Sci USA* **113**: 9704–9709.

Lewis, A. H., and Falkinham, J. O., 3rd. (2015) Microaerobic growth and anaerobic survival of *Mycobacterium avium*, *Mycobacterium intracellulare* and *Mycobacterium scrofulaceum*. *Int J Mycobacteriol* **4**: 25–30.

Loi, V. V., Rossius, M., and Antelmann, H. (2015) Redox regulation by reversible protein S-thiolation in bacteria. *Front Microbiol* **6**: 187.

Lotierzo, M., Tse Sum Bui, B., Florentin, D., Escalettes, F., and Marquet, A. (2005) Biotin synthase mechanism: an overview. *Biochem Soc Trans* **33**: 820–823.

MacInnes, J. I., Kim, J. E., Lian, C. J., and Soltes, G. A. (1990) *Actinobacillus pleuropneumoniae* hlyX gene homology with the fnr gene of *Escherichia coli*. *J Bacteriol* **172**: 4587–4592.

Maier, T. M., and Myers, C. R. (2001) Isolation and characterization of a *Shewanella putrefaciens* MR-1 electron transport regulator etrA mutant: reassessment of the role of EtrA. *J Bacteriol* **183**: 4918–4926.

Malkin, R., and Rabinowitz, J. C. (1967) Nonheme iron electron-transfer proteins. *Annu Rev Biochem* **36**: 113–148.

Malpica, R., Franco, B., Rodriguez, C., Kwon, O., and Georgellis, D. (2004) Identification of a quinone-sensitive redox switch in the ArcB sensor kinase. *Proc Natl Acad Sci USA* **101**: 13318–13323.

Martin, W. F., and Sousa, F. L. (2015) Early microbial evolution: the age of anaerobes. *Cold Spring Harb Perspect Biol* **8**: a018127.

Mead, D. A., Lucas, S., Copeland, A., Lapidus, A., Cheng, J. F., Bruce, D. C., *et al*. (2012) Complete genome sequence of *Paenibacillus* strain Y4.12MC10, a novel *Paenibacillus lautus* strain isolated from obsidian hot spring in Yellowstone National Park. *Stand Genomic Sci* **6**: 381–400.

Melville, S. B., and Gunsalus, R. P. (1990) Mutations in fnr that alter anaerobic regulation of electron transport-associated genes in *Escherichia coli*. *J Biol Chem* **265**: 18733–18736.

Messaoudi, K., Clavel, T., Schmitt, P., and Duport, C. (2010) Fnr mediates carbohydrate-dependent regulation of catabolic and enterotoxin genes in *Bacillus cereus* F4430/73. *Res Microbiol* **161**: 30–39.

Mettert, E. L., and Kiley, P. J. (2015) Fe–S proteins that regulate gene expression. *Biochim Biophys Acta* **1853**: 1284–1293.

Miyatake, H., Mukai, M., Park, S. Y., Adachi, S., Tamura, K., Nakamura, H., *et al*. (2000) Sensory mechanism of oxygen sensor FixL from *Rhizobium meliloti*: crystallographic, mutagenesis and resonance Raman spectroscopic studies. *J Mol Biol* **301**: 415–431.

Molle, V., Palframan, W. J., Findlay, K. C., and Buttner, M. J. (2000) WhiD and WhiB, homologous proteins required for different stages of sporulation in *Streptomyces coelicolor* A3(2). *J Bacteriol* **182**: 1286–1295.

Morita, H., Toh, H., Fukuda, S., Horikawa, H., Oshima, K., Suzuki, T., *et al*. (2008) Comparative genome analysis of *Lactobacillus reuteri* and *Lactobacillus fermentum* reveal a genomic Island for reuterin and cobalamin production. *DNA Res* **15**: 151–161.

Muhd Sakaff, M. K., Abdul Rahman, A. Y., Saito, J. A., Hou, S., and Alam, M. (2012) Complete genome sequence of the thermophilic bacterium *Geobacillus thermoleovorans* CCB_US3_UF5. *J Bacteriol* **194**: 1239.

Mullner, M., Hammel, O., Mienert, B., Schlag, S., Bill, E., and Unden, G. (2008) A PAS domain with an oxygen labile [4Fe–4S](2+) cluster in the oxygen sensor kinase NreB of *Staphylococcus carnosus*. *Biochemistry* **47**: 13921–13932.

Muraki, N., Nomata, J., Ebata, K., Mizoguchi, T., Shiba, T., Tamiaki, H., *et al*. (2010) X-ray crystal structure of the light-independent protochlorophyllide reductase. *Nature* **465**: 110–114.

Nelson-Sathi, S., Sousa, F. L., Roettger, M., Lozada-Chavez, N., Thiergart, T., Janssen, A., *et al*. (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**: 77–80.

Nilkens, S., Koch-Singenstreu, M., Niemann, V., Gotz, F., Stehle, T., and Unden, G. (2014) Nitrate/oxygen co-sensing by an NreA/NreB sensor complex of *Staphylococcus carnosus*. *Mol Microbiol* **91**: 381–393.

Ogino, H., Inomata, S., and Tobita, H. (1998) Abiological iron–sulfur clusters. *Chem Rev* **98**: 2093–2122.

Ohno, R., Ueyama, N., and Nakamura, A. (1991) Influence of the distal para substituent through NH–S hydrogen

bonds on the positive shift of the reduction potentials of [Fe4S4(Z-cys-Gly-NHC6H4-p-X)4]2– (X = H, OMe, F, Cl, CN) complexes. *Inorg Chem* **30**: 4887–4891.

Ollagnier-De Choudens, S., Sanakis, Y., Hewitson, K. S., Roach, P., Baldwin, J. E., Munck, E., and Fontecave, M. (2000) Iron-sulfur center of biotin synthase and lipoate synthase. *Biochemistry* **39**: 4165–4173.

Ramey, B. E., Matthysse, A. G., and Fuqua, C. (2004) The FNR-type transcriptional regulator SinR controls maturation of *Agrobacterium tumefaciens* biofilms. *Mol Microbiol* **52**: 1495–1511.

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

Reents, H., Munch, R., Dammeyer, T., Jahn, D., and Hartig, E. (2006a) The Fnr regulon of *Bacillus subtilis*. *J Bacteriol* **188**: 1103–1112.

Reents, H., Gruner, I., Harmening, U., Böttger, L. H., Layer, G., Heathcote, P., *et al*. (2006b) *Bacillus subtilis* Fnr senses oxygen via a [4Fe–4S] cluster coordinated by three cysteine residues without change in the oligomeric state. *Mol Microbiol* **60**: 1432–1445.

Rey, M. W., Ramaiya, P., Nelson, B. A., Brody-Karpin, S. D., Zaretsky, E. J., Tang, M., *et al*. (2004) Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related *Bacillus* species. *Genome Biol* **5**: R77.

Rice, P., Longden, I., and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

Rodrigues, D. F., Ivanova, N., He, Z., Huebner, M., Zhou, J., and Tiedje, J. M. (2008) Architecture of thermal adaptation in an *Exiguobacterium sibiricum* strain isolated from 3 million year old permafrost: a genome and transcriptome approach. *BMC Genomics* **9**: 547.

Rouault, T. A., Haile, D. J., Downey, W. E., Philpott, C. C., Tang, C., Samaniego, F., *et al*. (1992) An iron–sulfur cluster plays a novel regulatory role in the iron-responsive element binding protein. *Biometals* **5**: 131–140.

Saier, M. H., Ramseier, T. M., and Reizer, J. (1996)Regulation of carbon utilization . In *Escherichia coli* and *Salmonella:* Cellular and Molecular Biology, Neidhardt, F. C. (ed). Washington, C: ASM Press, pp. 1325–1343.

Saini, V., Farhana, A., and Steyn, A. J. (2012) *Mycobacterium tuberculosis* WhiB3: a novel iron-sulfur cluster protein that regulates redox homeostasis and virulence. *Antioxid Redox Signal* **16**: 687–697.

Salmon, K., Hung, S. P., Mekjian, K., Baldi, P., Hatfield, G. W., and Gunsalus, R. P. (2003) Global gene expression profiling in *Escherichia coli* K12. The effects of oxygen availability and FNR. *J Biol Chem* **278**: 29837–29855.

Sass, A. M., Schmerk, C., Agnoli, K., Norville, P. J., Eberl, L., Valvano, M. A., and Mahenthiralingam, E. (2013) The unexpected discovery of a novel low-oxygen-activated locus for the anoxic persistence of *Burkholderia cenocepacia*. *ISME J* **7**: 1568–1581.

Saw, J. H., Mountain, B. W., Feng, L., Omelchenko, M. V., Hou, S., Saito, J. A., *et al*. (2008) Encapsulated in silica: genome, proteome and physiology of the thermophilic bacterium *Anoxybacillus flavithermus* WK1. *Genome Biol* **9**: R161.

Sawers, R. G. (1991) Identification and molecular characterization of a transcriptional regulator from *Pseudomonas aeruginosa* PAO1 exhibiting structural and functional similarity to the FNR protein of *Escherichia coli*. *Mol Microbiol* **5**: 1469–1481.

Schlag, S., Fuchs, S., Nerz, C., Gaupp, R., Engelmann, S., Liebeke, M., *et al*. (2008) Characterization of the oxygen-responsive NreABC regulon of *Staphylococcus aureus*. *J Bacteriol* **190**: 7847–7858.

Schluter, A., Patschkowski, T., Unden, G., and Priefer, U. B. (1992) The *Rhizobium leguminosarum* FnrN protein is functionally similar to *Escherichia coli* Fnr and promotes heterologous oxygen-dependent activation of transcription. *Mol Microbiol* **6**: 3395–3404.

Schmidt, A., Hammerbacher, A. S., Bastian, M., Nieken, K. J., Klockgether, J., Merighi, M., *et al*. (2016) Oxygen-dependent regulation of c-di-GMP synthesis by SadC controls alginate production in *Pseudomonas aeruginosa*. *Environ Microbiol* **18**: 3390–3402.

Scott, C., Guest, J. R., and Green, J. (2000) Characterization of the *Lactococcus lactis* transcription factor FlpA and demonstration of an in vitro switch. *Mol Microbiol* **35**: 1383–1393.

Scotti, J. S., Leung, I. K., Ge, W., Bentley, M. A., Paps, J., Kramer, H. B., *et al*. (2014) Human oxygen sensing may have origins in prokaryotic elongation factor Tu prolyl-hydroxylation. *Proc Natl Acad Sci USA* **111**: 13331–13336.

Septer, A. N., Bose, J. L., Dunn, A. K., and Stabb, E. V. (2010) FNR-mediated regulation of bioluminescence and anaerobic respiration in the light-organ symbiont *Vibrio fischeri*. *FEMS Microbiol Lett* **306**: 72–81.

Shaw, D. J., and Guest, J. R. (1982) Nucleotide sequence of the fnr gene and primary structure of the Fnr protein of *Escherichia coli*. *Nucleic Acids Res* **10**: 6119–6130.

Shaw, D. J., Rice, D. W., and Guest, J. R. (1983) Homology between CAP and Fnr, a regulator of anaerobic respiration in *Escherichia coli*. *J Mol Biol* **166**: 241–247.

Sickmier, E. A., Brekasis, D., Paranawithana, S., Bonanno, J. B., Paget, M. S., Burley, S. K., and Kielkopf, C. L. (2005) X-ray structure of a Rex-family repressor/NADH complex insights into the mechanism of redox sensing. *Structure* **13**: 43–54.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., *et al*. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**: 539.

Singh, A., Guidry, L., Narasimhulu, K. V., Mai, D., Trombley, J., Redding, K. E., *et al*. (2007) Mycobacterium tuberculosis WhiB3 responds to O2 and nitric oxide via its [4Fe–4S] cluster and is essential for nutrient starvation survival. *Proc Natl Acad Sci U S A* **104**: 11562–11567.

Singh, A., Crossman, D. K., Mai, D., Guidry, L., Voskuil, M. I., Renfrow, M. B., and Steyn, A. J. (2009) Mycobacterium tuberculosis WhiB3 maintains redox homeostasis by regulating virulence lipid anabolism to modulate macrophage response. *PLoS Pathog* **5**: e1000545.

Soliveri, J. A., Gomez, J., Bishai, W. R., and Chater, K. F. (2000) Multiple paralogous genes related to the *Streptomyces coelicolor* developmental regulatory gene whiB are

present in *Streptomyces* and other actinomycetes. *Microbiology* **146**: 333–343.
Song, J. Y., Kim, H. A., Kim, J. S., Kim, S. Y., Jeong, H., Kang, S. G., *et al*. (2012) Genome sequence of the plant growth-promoting rhizobacterium *Bacillus* sp. strain JS. *J Bacteriol* **194**: 3760–3761.
Spiro, S., and Guest, J. R. (1988) Inactivation of the FNR protein of *Escherichia coli* by targeted mutagenesis in the N-terminal region. *Mol Microbiol* **2**: 701–707.
Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
Staples, C. R., Lahiri, S., Raymond, J., Von Herbulis, L., Mukhophadhyay, B., and Blankenship, R. E. (2007) Expression and association of group IV nitrogenase NifD and NifH homologs in the non-nitrogen-fixing archaeon *Methanocaldococcus jannaschii*. *J Bacteriol* **189**: 7392–7398.
Steyn, A. J., Collins, D. M., Hondalus, M. K., Jacobs, W. R., Jr., Kawakami, R. P., and Bloom, B. R. (2002) *Mycobacterium tuberculosis* WhiB3 interacts with RpoV to affect host survival but is dispensable for in vivo growth. *Proc Natl Acad Sci U S A* **99**: 3147–3152.
Stolper, D. A., and Keller, C. B. (2018) A record of deep-ocean dissolved O2 from the oxidation state of iron in submarine basalts. *Nature* **553**: 323–327.
Switzer Blum, J., Burns Bindi, A., Buzzelli, J., Stolz, J. F., and Oremland, R. S. (1998) *Bacillus arsenicoselenatis*, sp. nov., and *Bacillus selenitireducens*, sp. nov.: two haloalkaliphiles from Mono Lake, California that respire oxyanions of selenium and arsenic. *Arch Microbiol* **171**: 19–30.
Takahashi, M., Blazy, B., and Baudras, A. (1980) An equilibrium study of the cooperative binding of adenosine cyclic 3′,5′-monophosphate and guanosine cyclic 3′,5′-monophosphate to the adenosine cyclic 3′,5′-monophosphate receptor protein from *Escherichia coli*. *Biochemistry* **19**: 5124–5130.
Tang, Y., and Guest, J. R. (1999) Direct evidence for mRNA binding and post-transcriptional regulation by *Escherichia coli* aconitases. *Microbiology* **145**: 3069–3079.
Tang, Y., Guest, J. R., Artymiuk, P. J., and Green, J. (2005) Switching aconitase B between catalytic and regulatory modes involves iron-dependent dimer formation. *Mol Microbiol* **56**: 1149–1158.
Tseng, C. P., Albrecht, J., and Gunsalus, R. P. (1996) Effect of microaerophilic cell growth conditions on expression of the aerobic (cyoABCDE and cydAB) and anaerobic (narGHJI, frdABCD, and dmsABC) respiratory pathway genes in *Escherichia coli*. *J Bacteriol* **178**: 1094–1098.
Ueyama, N., Kajiwara, A., Terakawa, T., Ueno, S., and Nakamura, A. (1985) Redox potentials of oligopeptide/Fe4S42+ complexes. Remarkable positive shift of the redox potential with (benzyloxycarbonyl)-L-Cys-Gly-L-Ala-L-Cys-OMe as chelating ligands. *Inorg Chem* **24**: 4700–4704.
Ugulava, N. B., Gibney, B. R., and Jarrett, J. T. (2000) Iron–sulfur cluster interconversions in biotin synthase: dissociation and reassociation of iron during conversion of [2Fe–2S] to [4Fe–4S] clusters. *Biochemistry* **39**: 5206–5214.
Uhlich, G. A., McNamara, P. J., Iandolo, J. J., and Mosier, D. A. (1999) Cloning and characterization of the gene encoding *Pasteurella haemolytica* FnrP, a regulator of the *Escherichia coli* silent hemolysin sheA. *J Bacteriol* **181**: 3845–3848.
Unden, G., and Bongaerts, J. (1997) Alternative respiratory pathways of *Escherichia coli*: energetics and transcriptional regulation in response to electron acceptors. *Biochim Biophys Acta* **1320**: 217–234.
Unden, G., Müllner, M., and Reinhard, F. (2010)Sensing of oxygen by bacteria . In *Bacterial Signaling*, Krämer, R., and Jung, K. (eds). Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA.
Unden, G., Nilkens, S., and Singenstreu, M. (2013) Bacterial sensor kinases using Fe–S cluster binding PAS or GAF domains for O2 sensing. *Dalton Trans* **42**: 3082–3087.
Venkateswara Rao, P., and Holm, R. H. (2004) Synthetic analogues of the active sites of iron–sulfur proteins. *Chem Rev* **104**: 527–559.
Volbeda, A., Darnault, C., Renoux, O., Nicolet, Y., and Fontecilla-Camps, J. C. (2015) The crystal structure of the global anaerobic transcriptional regulator FNR explains its extremely fine-tuned monomer-dimer equilibrium. *Sci Adv* **1**: e1501086.
Volbeda, A., Dodd, E. L., Darnault, C., Crack, J. C., Renoux, O., Hutchings, M. I., *et al*. (2017) Crystal structures of the NO sensor NsrR reveal how its iron-sulfur cluster modulates DNA binding. *Nat Commun* **8**: 15052.
Wang, E., Bauer, M. C., Rogstam, A., Linse, S., Logan, D. T., and von Wachenfeldt, C. (2008) Structure and functional properties of the *Bacillus subtilis* transcriptional repressor Rex. *Mol Microbiol* **69**: 466–478.
Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W. F. (2016) The physiology and habitat of the last universal common ancestor. *Nat Microbiol* **1**: 16116.
Wood, G. E., Khelef, N., Guiso, N., and Friedman, R. L. (1998) Identification of Btr-regulated genes using a titration assay. Search for a role for this transcriptional regulator in the growth and virulence of *Bordetella pertussis*. *Gene* **209**: 51–58.
Zeilstra-Ryalls, J. H., Gabbert, K., Mouncey, N. J., Kaplan, S., and Kranz, R. G. (1997) Analysis of the fnrL gene and its function in *Rhodobacter capsulatus*. *J Bacteriol* **179**: 7264–7273.
Zhang, B., Crack, J. C., Subramanian, S., Green, J., Thomson, A. J., Le Brun, N. E., and Johnson, M. K. (2012) Reversible cycling between cysteine persulfide-ligated [2Fe–2S] and cysteine-ligated [4Fe–4S] clusters in the FNR regulatory protein. *Proc Natl Acad Sci U S A* **109**: 15734–15739.
Zigha, A., Rosenfeld, E., Schmitt, P., and Duport, C. (2007) The redox regulator Fnr is required for fermentative growth and enterotoxin synthesis in *Bacillus cereus* F4430/73. *J Bacteriol* **189**: 2813–2824.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1.
Appendix S2.

6 Zusammenfassung der Ergebnisse

Der letzte universelle Vorfahre aller Organismen (engl. *last universal common ancestor*, LUCA) ist das Verbindungsstück der abiotischen mit der biotischen Welt (Fox *et al.*, 1980). Da die evolutionären Aufzeichnungen durch lateralen Gentransfer (LGT, Kannan *et al.* 2013) und Sequenzdivergenz gelöscht oder verändert wurden, muss anhand von modernen Organismen ein Rückschluss auf den Ursprung gezogen werden (Weiss *et al.*, 2016b; Martin *et al.*, 2017, 2016; Weiss *et al.*, 2018). In vielen Analysen von LUCAs Genom wurde von unterschiedlichen Gruppen ausschließlich nach universell vohandenen Genen in Bakterien und Archaeen gesucht (Koonin, 2003; Ouzounis *et al.*, 2006; Kannan *et al.*, 2013). Dabei konnten 30-100 Gene identifiziert werden, die wahrscheinlich von LUCA stammen. Die gefundenen Gene sind überwiegend ribosomale Gene und an der Translation beteiligt.

In der für diese Arbeit durchgeführten Analyse wurden zwei Kriterien angewendet, um LGT weitesgehend aus dem verwendeten Datensatz zu filtern und gleichzeitig Proteine zu identifizieren, die auf LUCA zurückzuführen sind. Zum einen muss ein Protein in mindestens zwei Taxa der Bakterien und Archaeen vertreten sein und zweitens muss der phylogenetische Stammbaum des Proteins eine Monophylie der Domänen aufweisen (Weiss *et al.*, 2016b; Martin *et al.*, 2017, 2016; Weiss *et al.*, 2018). Proteine, die beide Kriterien erfüllen, können als mögliche LUCA Proteine deklariert werden und wurden wahrscheinlich horizontal vererbt. Dadurch, dass nicht nur universelle Proteine in dieser Analyse betrachtet wurden, können die Ergebnisse Rückschlüsse auf die Physiologie und das Habitat von LUCA geben (Weiss *et al.*, 2016b; Martin *et al.*, 2017, 2016; Weiss *et al.*, 2018). Allerdings können falsch-positive Treffer nie komplett ausgeschlossen werden, da es keine fossilen Rückstände von LUCA gibt (Weiss *et al.*, 2016b; Martin *et al.*, 2016). Die Analyse kann durch mehrfache LGT Ereignisse, Sequenzkonservierung über die Jahre und die begrenzte Anzahl an Organismen, beziehungsweise der geringen Anzahl an Sequenzen beeinflusst werden und zu Widersprüchen führen.

Durch die erwähnten Kriterien konnten aus 286.514 Proteinfamilien 355 Proteine identifiziert werden, die wahrscheinlich bereits in LUCA vorhanden waren. Unter anderem sind in diesen LUCA Proteinen 19 Proteine enthalten, die zur Ribosombiosynthese gehören, 8 Proteine, die mit der Aminoacyl tRNA Synthetase in Verbindung stehen, sowie einige Proteine, die zum Grundsatz der Proteine gehören, die für die Informationsweitergabe innerhalb der Zelle zuständig sind. Diese Ergebnisse spiegeln die bereits bekannten Kenntnisse über LUCA aus vorherigen Analysen wieder. Die weiterhin gefundenen Proteine wurden in vorherigen Analysen nicht gefunden und sind sehr divergent auf unterschiedliche Funktionen aufgeteilt.

Anhand der genomischen Daten kann gezeigt werden, dass LUCA ein anaerober autotropher Organismus (Say und Fuchs, 2010) war, der den Wood-Ljungdahl (WL) Stoff-

wechselweg nutzte (Fuchs, 2011) und in einer Umgebung lebte, die Hydrothermalquellen ähnelten (Baross und Hoffman, 1985; Russell und Hall, 1997). Die in den Proteinen enthaltene Reverse Gyrase weist darauf hin, das LUCA hyperthemophil gelebt hat (Déclais *et al.*, 2000). Es wurden in dieser Analyse keine Proteine gefunden, die auf eine chemoorganotrophe Lebensweise schließen lassen. Allerdings konnten Proteine identifiziert werden, die auf eine chemolithoautotrophe Lebensweise hindeuten. Des Weiteren sind zwei Enzyme des Energiestoffwechsels in den Ergebnissen zu finden. Zum einen Terephthalsäure und zum anderen ATP-Synthase. Außerdem waren Proteine des WL-Weges, Nitrogenasen und Hydrogenasen in der Proteinliste von LUCA enthalten. Diese Proteine deuten auf einen anaeroben Lebensraum hin, da sie sauerstoffsensitiv sind. Des Weiteren war LUCA wahrscheinlich abhängig von Wasserstoff, Kohlenstoffdioxid und Stickstoff.

Kofaktoren, die bei LUCA eine große Rolle spielten, waren Pterine, Molybdän-Cofaktor, Cobalamine, Sirohäm, Thiaminpyrophosphat, Coenzym M und F420. Außerdem sind viele gefundene Proteine S-adenosyl methionine (SAM) abhängig. Ein weiterer Hinweis auf den Lebensraum von LUCA geben die Transitionsmetalle und Eisen-Schwefel bzw. Eisen-Nickel-Schwefel Cluster, die in den LUCA Enzymen vorhanden sind. Diese weisen auf einen metallreichen Lebensraum hin (Weiss *et al.*, 2016b; Martin *et al.*, 2017, 2016; Weiss *et al.*, 2018).

In den in der Analyse erstellten phylogenetischen Stammbäumen wurden die geringsten Distanzen zwischen den Bakterien und Archaeen berechnet, um die ursprünglichsten Organismen, also die Organismen die wahrscheinlich als erstes aus LUCA entstanden sind, zu identifizieren. Dabei wurde deutlich, dass die acetogenen Clostridien, sowie die Methanogenen die ursprünglichsten Prokaryoten sind. Beide Phyla besitzen den WL-Stoffwechselweg und sind in thermophilen Habitaten zu finden (Weiss *et al.*, 2016b; Martin *et al.*, 2017, 2016; Weiss *et al.*, 2018).

Der für die Analyse von LUCA verwendete Datensatz eignet sich auch für die Analyse von anderen ursprünglichen Proteinen. Durch die Verwendung von phylogenomischen Analysemethoden konnte der Ursprung und der phylogenetische Zusammenhang von Sauerstoffsensoren in Bakterien analysiert werden. Dazu wurden Sauerstoffsensoren, welche Eisenschwefel Cluster enthielten (FNR, NreB und WhiB3) mit den Proteinfamilien aus (Weiss *et al.*, 2016b) verglichen und phylogenetische Stammbäume erstellt. Anhand der Stammbäume konnte festgestellt werden, dass die Sauerstoffsensoren unabhägig voneinander entstanden sind und durch LGT weiter verbreitet wurden.

Literaturverzeichnis

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. und Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**:403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. und Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**:3389–3402.

Bada, J. L. (2004). How life began on Earth: A status report. *Earth and Planetary Science Letters*, **226**:1–15.

Balch, W. E., Fox, G. E., Magrum, L. J., Woese, C. R. und Wolfe, R. S. (1979). Methanogens: Reevaluation of a unique biological group. *Microbiological Reviews*, **43**:260–296.

Baross, J. und Hoffman, S. (1985). Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Origins Of Life*, **15**:327–345.

Barth, C., Weiss, M. C., Roettger, M., Martin, W. F. und Unden, G. (2018). Origin and phylogenetic relationships of [4Fe–4S]-containing O_2 sensors of bacteria. *Environmental Microbiology*, **20**:4567–4586.

Boykin, L. M., Kubatko, L. S. und Lowrey, T. K. (2010). Comparison of methods for rooting phylogenetic trees: A case study using Orcuttieae (Poaceae: Chloridoideae). *Molecular Phylogenetics and Evolution*, **54**:687–700.

Broderick, J. B., Duffus, B. R., Duschene, K. S. und Shepard, E. M. (2014). Radical S-adenosylmethionine enzymes. *Chemical Reviews*, **114**:4229–4317.

Bruneman, P. (1971). The recovery of trees from measure of dissimilarity. In: **Hadison, F. R., Kendall, D. G. und Taurau, P., (Hrsg.)**, *Mathematics and the archeological and historical sciences.*, S. 387–395. Edinburgh University Press, Edinburgh.

Cavalli-Sforza, L. L. und Edwards, A. W. F. (1967). Phylogenetic analysis models and estimation procedures. *The American Journal of Human Genetics*, **19**:233–257.

Charlebois, R. L. und Doolittle, W. F. (2004). Computing prokaryotic gene ubiquity: Rescuing the core from extinction. *Genome Research*, **14**:2469–2477.

Dagan, T. und Martin, W. (2006). The tree of one percent. *Genome Biology*, **7**:118.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray, London, 1. Ausgabe.

de Duve, C. (2005). The onset of selection. *Nature*, **433**:581–582.

De Soete, G. (1984). Ultrametric tree representations of incomplete dissimilarity data. *Quality and Quantity*, **18**:387–393.

Déclais, A. C., Marsault, J., Confalonieri, F., Bouthier De La Tour, C. und Duguet, M. (2000). Reverse gyrase, the two domains intimately cooperate to promote positive supercoiling. *Journal of Biological Chemistry*, **275**:19498–19504.

Delsuc, F., Brinkmann, H. und Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, **6**:361–375.

Dongen, S. V. (2000). A cluster algorithm for graphs. *Information Systems*, **-**:1–40.

Dumas, J.-P. und Ninio, J. (1982). Efficient algorithms for folding and comparing nucleic acid sequences. *Nucleic Acids Research*, **10**:197–206.

Enright, A. J., Van Dongen, S. und Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, **30**:1575–1584.

Farris, J. S. (1972). Estimating phylogenetic trees from distance matrices. *The American Naturalist*, **106**:645–668.

Farris, J. S., Kluge, A. G. und Eckardt, M. J. (1970). A numerical approach to phylogenetic systematics. *Systematic Biology*, **19**:172–189.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **27**:401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**:368–376.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology*, **266**:418–27.

Feng, D.-F. und Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, **25**:351–360.

Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H. und Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini.

Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R., Balch, W. E., Tanner, R. S., Magrum, L. J., Zablen, L. B., Blakemore, R., Grupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen,

K. R., Chen, K. N. und Woese, C. R. (1980). The phylogeny of prokaryotes. *Science*, **209**:457–463.

Fuchs, G. (2011). Alternative Pathways of Carbon Dioxide Fixation: Insights into the Early Evolution of Life? *Annual Review of Microbiology*, **65**:631–658.

Gabaldón, T. (2005). Evolution of proteins and proteomes: A phylogenetics approach. *Evolutionary Bioinformatics Online*, **1**:51–61.

Graur, D. und Li, W. H. (2000). Dynamics of genes in populations. In: *Fundamentals of Molecular Evolution*. Sinaur Associates, Inc., Sunderland, Massachusetts.

Haeckel, E. (1866). *Generelle Morphologie der Organismen.* Verlag von Georg Reimer, 1. Ausgabe.

Haldane, J. (1929). The origin of life. *Rationalist Annual*, **148**:242–249.

Hasegawa, M. und Fujiwara, M. (1993). Reltive efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Molecular Phylogenetics and Evolution*, **2**:1–5.

Hasegawa, M., Kishino, H. und Saitou, N. (1991). On the maximum likelihood method in molecular phylogenetics. *Journal of Molecular Evolution*, **32**:443–445.

Hennig, W. (1950). *Grundzüge einer Theorie der phylogenetischen Systematik*. Deutscher Zentralverlag, Berlin.

Hennig, W. (1966). *Phylogenetic Systematics.* University of Illionois Press.

Jain, A. K., Murty, M. N. und Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, **31**:264–323.

Kandler, O. (1995). Cell wall biochemistry in Archaea and its phylogenetic implications. *Journal of Biological Physics*, **20**:165–169.

Kannan, L., Li, H., Rubinstein, B. und Mushegian, A. (2013). Models of gene gain and gene loss for probabilistic reconstruction of gene content in the last universal common ancestor of life. *Biology Direct*, **8**:1–12.

Kelley, D. S., Baross, J. A. und Delaney, J. R. (2002). Volcanoes, fluids, and life at mid-ocean ridge spreading centers. *Annual Review of Earth and Planetary Sciences*, **30**:385–491.

Kishino, H., Miyata, T. und Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, **31**:151–160.

Koonin, E. V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, **1**:127–136.

Lane, N., Allen, J. F. und Martin, W. (2010). How did LUCA make a living? Chemiosmosis in the origin of life. *BioEssays*, **32**:271–280.

Lipman, D. J. und Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, **227**:1435–1441.

Maddison, W. P., Donoghue, M. J. und Maddison, D. R. (1984). Outcrop analysis and parsimony. *Systematic Zoology*, **33**:83–103.

Martin, W., Baross, J., Kelley, D. und Russell, M. J. (2008). Hydrothermal vents and the origin of life. *Nature Reviews Microbiology*, **6**:805–814.

Martin, W. und Russell, M. J. (2003). On the origins of cells: A hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **358**:59–85.

Martin, W. und Russell, M. J. (2007). On the origin of biochemistry at an alkaline hydrothermal vent. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **362**:1887–1925.

Martin, W., Sousa, F. und Lane, N. (2014). Energy at life's origin. *Science*, **344**:1092–1093.

Martin, W. F. und Kowallik, K. V. (1999). Über Natur und Ursprung der Chromatophoren im Pflanzenreiche [On the nature and origin of chromatophores in the plant kingdom]. *European Journal of Phycology*, **34**:287–295.

Martin, W. F., Weiss, M. C., Neukirchen, S., Nelson-Sathi, S. und Sousa, F. L. (2016). Physiology, phylogeny, and LUCA. *Microbial Cell*, **3**:582–587.

Martin, W. F., Zimorski, V. und Weiss, M. C. (2017). Wo lebten die ersten Zellen – und wovon?: Frühe Evolution. *Biologie in Unserer Zeit*, **47**:186–192.

Mereschkowsky, K. (1905). Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biologisches Centralblatt. Bd. 25*, **25**:593–604.

Miller, S. L. (1953). A production of amino acids under possible primitive earth conditions. *Science*, **117**:528–529.

Miyamoto, M. M. und Cracraft, J. (1991). Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics. In: **Miyamoto, M. M. und Cracraft, J., (Hrsg.)**, *Phylogenetic analysis of DNA sequences*, Kapitel 1, S. 3–17. Oxford University Press, New York, 1. Ausgabe.

Napier, W. M. (2004). A mechanism for interstellar panspermia. *Monthly Notices of the Royal Astronomical Society*, **348**:46–51.

Needleman, S. B. und Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**:443–453.

Nelson-Sathi, S., Sousa, F. L., Roettger, M., Lozada-Chávez, N., Thiergart, T., Janssen, A., Bryant, D., Landan, G., Schoenheit, P., Siebers, B., McInerney, J. O. und Martin, W. F. (2015). Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature*, **517**:77–80.

Oparin, A. I. (1952). *The Origin of Life*. New York: Academic Press, New York, 1. Ausgabe.

Orgel, L. E. (2004). Prebiotic chemistry and the origin of the RNA world. *Critical Reviews in Biochemistry and Molecular Biology*, **39**:99–123.

Ouzounis, C. A., Kunin, V., Darzentas, N. und Goldovsky, L. (2006). A minimal estimate for the gene content of the last universal common ancestor - Exobiology from a terrestrial perspective. *Research in Microbiology*, **157**:57–68.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology*, **183**:63–98.

Pearson, W. R. und Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**:2444–2448.

Penny, D. (1982). Towards a basis for classification: the incompleteness of distance measures, incompatibility analysis and phenetic classification. *Journal of Theoretical Biology*, **96**:129–142.

Poehlein, A., Schmidt, S., Kaster, A. K., Goenrich, M., Vollmers, J., Thürmer, A., Bertsch, J., Schuchmann, K., Voigt, B., Hecker, M., Daniel, R., Thauer, R. K., Gottschalk, G. und Müller, V. (2012). An ancient pathway combining carbon dioxide fixation with the generation and utilization of a sodium ion gradient for ATP synthesis. *PLoS ONE*, **7**:e33439.

Puigbò, P., Wolf, Y. I. und Koonin, E. V. (2010). The tree and net components of prokaryote evolution. *Genome Biology and Evolution*, **2**:745–756.

Raymann, K., Brochier-Armanet, C. und Gribaldo, S. (2015). The two-domain tree of life is linked to a new root for the Archaea. *Proceedings of the National Academy of Sciences of the United States of America*, **112**:6670–6675.

Russell, M. J. (2007). The alkaline solution to the emergence of life: Energy, entropy and early evolution. *Acta Biotheoretica*, **55**:133–179.

Russell, M. J. und Hall, A. J. (1997). The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. *Journal of the Geological Society*.

Saitou, N. und Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**:406–25.

Sanger, F., Nicklen, S. und Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, **74**:5463–5467.

Say, R. F. und Fuchs, G. (2010). Fructose 1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. *Nature*, **464**:1077–1081.

Schaeffer, S. E. (2007a). Graph clustering. *Computer Science Review*, **1**:27–64.

Schaeffer, S. E. (2007b). Graph clustering by flow simulation. *Computer Science Review*, **1**:27–64.

Schleifer, K. H. und Kandler, O. (1972). Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriological Reviews*, **36**:407–477.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D. und Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, **7**:539.

Smith, T. und Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Evolution*, **147**:195–197.

Sneath, P. H. und Sokal, R. R. (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. W.H. Freeman & Co Ltd, San Francisco, first edit. Ausgabe.

Sokal, R. R. und Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, **38**:1409–1438.

Sørensen, T. (1948). A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and its Application to Analyses of the Vegetation on Danish Commons. *Det Kongelige Danske Videnskabernes Selskab Biologiske Skrifter*, **5**:1–34.

Sousa, F. L., Thiergart, T., Landan, G., Nelson-Sathi, S., Pereira, I. A. C., Allen, J. F., Lane, N. und Martin, W. F. (2013). Early bioenergetic evolution. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **368**:20130088.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**:1312–1313.

Takahashi, K. und Nei, M. (2000). Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Molecular Biology and Evolution*, **17**:1251–1258.

Thauer, R. K. (1998). Biochemistry of methanogenesis: A tribute to Marjory Stephenson. 1998 Marjory Stephenson Prize Lecture. *Microbiology (Reading, England)*, **144**:2377–2406.

Tria, F. D. K., Landan, G. und Dagan, T. (2017). Phylogenetic rooting using minimal ancestor deviation. *Nature Ecology & Evolution*, **1**:0193.

van Dongen, S. (2000). *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht.

Wächtershäuser, G. (2006). From volcanic origins of chemoautotrophic life to Bacteria, Archaea and Eukarya. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **361**:1787–1806.

Watson, J. und Crick, F. (1953). A structure for deoxyribose nucleic acid. *Nature*, **171**:737–738.

Weiss, M. C., Neukirchen, S., Roettger, M., Mrnjavac, N., Nelson-Sathi, S., Martin, W. F. und Sousa, F. L. (2016a). Reply to 'Is LUCA a thermophilic progenote?'.

Weiss, M. C., Preiner, M., Xavier, J. C., Zimorski, V. und Martin, W. F. (2018). The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genetics*, **14**:1–19.

Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-sathi, S. und Martin, W. F. (2016b). The physiology and habitat of the last universal common ancestor. *Nature Microbiology*, **1**:1–8.

Wheeler, W. C. (2012). *Systematics: A Course of Lectures*. Wiley-Blackwell, West Sussex, 1. Ausgabe.

Wilbur, W. J. und Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences of the United States of America*, **80**:726–730.

Williams, T. A., Foster, P. G., Cox, C. J. und Embley, T. M. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, **504**:231–236.

Woese, C. R. und Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, **74**:5088–5090.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V. und Koonin, E. V. (2002). Genome trees and the tree of life. *Trends in Genetics*, **18**:472–479.

Zahn, C. T. (1971). Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers*, **C-20**:68–86.

Zuckerkandl, E. und Pauling, L. (1962). Molecular Disease, Evolution and genic heterogeneity. In: **Kasha, M. und Pullman, B., (Hrsg.)**, *Horizons in biochemistry*, S. 189–225. Academic Press, New York.

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt und durch meine Unterschrift, dass die vorliegende Arbeit von mir selbstständig ohne fremde Hilfe angefertigt worden ist. Inhalte und Passagen, die aus fremden Quellen stammen und direkt oder indirekt übernommen worden sind, wurden als solche kenntlich gemacht. Ferner versichere ich, dass ich keine andere, außer der im Literaturverzeichnis angegebenen Literatur verwendet habe. Diese Versicherung bezieht sich sowohl auf Textinhalte sowie alle enthaltenen Abbildungen, Skizzen und Tabellen. Die Arbeit wurde bisher keiner Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Madeline Chantal Weiß, Düsseldorf, 2020

Danksagung

Mein ganz besonderer Dank gilt Pro. Dr. William Martin dafür, dass er mich in seinem Institut aufgenommen hat und es mir damit ermöglichte, dass überaus interessante Thema im Rahmen meiner Dissertation unter seiner Leitung zu bearbeiten. Sein umfassendes Wissen und seine Begeisterung haben meine Arbeit begleitet und gefördert. Er ermöglichte mir außerdem zahlreiche außergewöhnliche Reisen zu internationalen Fachkonferenzen, welche meine Arbeit bereichert haben.

Bei Prof. Dr. Laura Rose möchte ich mich für die Bereitschaft bedanken, als zweiter Berichterstatter zu fungieren.

Dr. Mayo Röttger und Dr. Nicole Grünheit möchte ich mich für die große Hilfsbereitschaft bei Fragen und für die Ratschläge zur Verbesserung meiner Programme, sowie für die konstruktive Kritik an dieser Arbeit bedanken.

I thank Dr. Shijulal Nelson-Sathi, Dr. Filipa Sousa and Dr. Joana Xavier for the good advices and their provided assistance with the publications.

Ich bedanke mich bei allen Koautoren der im Zusammenhang mit dieser Arbeit entstandenen Publikationen.

Ein weiterer Dank geht an meine Korrekturleser Renate Weiß, Dr. Nicole Grünheit, Jennifer Andres, Dr. Verena Zimorski und Andrea Alexa. Sie haben mir mit ihren Verbesserungsvorschlägen und konstruktiven Kritiken immer Anregungen geben meine Arbeit zu verfeinern.

Den Mitarbeitern des Instituts für Molekulare Evolution der Heinrich-Heine-Universität Düsseldorf möchte ich für die freundliche und motivierende Arbeitsatmosphäre danken.

Als letztes möchte ich noch meiner Familie, Freunden und den coolen Kids danken. Ohne eure Unterstützung und euren guten Zuspruch wäre ich nicht soweit gekommen. Danke für die Motivation und die Ablenkung von der Arbeit, wenn ich sie brauchte.