Untersuchung von evolutionären Netzwerken und Signalkonflikt in phylogenetischen Studien

Inaugural-Dissertation

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

vorgelegt von

Thorsten Thiergart
aus Düsseldorf

Düsseldorf, November 2014

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

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

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

Tag der mündlichen Prüfung: 12. Januar 2015

Publikationen, die im Rahmen der Dissertation veröffentlich wurden

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

Sousa FL, **Thiergart T**, Landan G, Nelson-Sathi S, Pereira IAC, Allen JF, Lane N, Martin WF (**2013**) Early bioenergetic evolution. *Philosophical Transactions of the Royal Society B*, 368:20130088.

Thiergart T, Landan G, Martin WF (**2014**) Concatenated alignments and the case of the disappearing tree. *BMC Evolutionary Biology*. Eingereicht in August 2014.

Weitere Publikationen

Martin WF, Röttger M, Klösges T, **Thiergart T**, Woehle C, Gould S, Dagan T (**2012**) Modern endosymbiotic theory: Getting lateral gene transfer into the equation. *Journal for Endocytobiosis and Cell Research*, 1:5.

Thiergart T, Landan G, Martin WF, Dagan T (**2014**) Application and comparative performance of network modularity algorithms to ecological communities classification. *Acta Societas Botanicorum Poloniae*, 83:93-102.

Nelson-Sathi S, Sousa FL, Röttger M, Lozada-Chavez N, **Thiergart T**, Janssen A, Bryant D, Landan G, Schönheit P, Siebers B, McInerney JO, Martin WF (**2014**) Origins of major archael clades correspond to gene acquisitions from bacteria. *Nature*, 517:77-80

Inhaltsverzeichnis

1 Zusammenfassung

Vergleiche von DNS- bzw. Aminosäuresequenzen um 1980 von Eukaryoten und Prokaryoten lieferten Anhaltspunkte für eine engere Verwandtschaft der Eukaryoten mit Archaebakterien, als mit Eubakterien. Mit der Verfügbarkeit von größeren Datensätzen und mit Hilfe von alternativen Betrachtungsweisen häuften sich Hinweise, dass Eukaryoten lediglich zu einem Teil archaebakteriellen Ursprungs sind, und sich auch eine Verbindung zu Eubakterien findet. Dies war in Übereinstimmung mit der Idee, wonach die Mitochondrien der Eukaryoten früher einmal freilebende Eubakterien waren und erheblich zu der Entwicklung der Eukaryoten beigetragen haben. Da ein Großteil der Studien, welche sich mit dieser Frage beschäftigen auf phylogenetischen Bäumen beruht, ist es von großem Interesse die Zuverlässigkeit dieser Bäume zu untersuchen.

Der erste Teil dieser Arbeit beschäftigt sich mit der Evolution von Eukaryoten und Prokaryoten unter der Hypothese, dass ihr verwandtschaftliches Verhältnis nicht unbedingt mit herkömmlichen phylogenetischen Darstellungen wiedergegeben werden kann. Dazu wurden 27 eukaryotische und 994 prokaryotische Genome verglichen und die Verteilung der eukaryotischen Gene auf die verschiedenen prokaryotischen Gruppen dargestellt. In phylogenetischen Bäumen wurde bestimmt, welche prokaryotische Gruppe in diesen den Eukaryoten am nächsten steht. Lediglich die Archaebakterien, hier besonders die Euryarchaeota und die α-Proteobakterien, wurden dabei überdurchschnittlich häufig ermittelt. Dabei sind die Rickettsiales, eine Untergruppe der α-Proteobakterien, im Gegensatz zu früheren Analysen, nicht besonders hervorgetreten.

Im zweiten Teil dieser Arbeit wurden die Ursachen und das Ausmaß von widersprüchlichen phylogenetischen Signalen untersucht. Repräsentative Datensätze für verschiedene Gruppen von Eukaryoten und Prokaryoten wurden verglichen und verschiedene Parameter auf ihre Einflussnahme getestet. Für alle Datensätze konnte gezeigt werden, dass phylogenetische Bäume, welche aus verketteten Sequenzdaten erstellt wurden, nur bedingt die Signale der einzelnen Bäume wiedergeben. Die Länge der untersuchten Sequenzen hat sich als ein entscheidender Einfluss herausgestellt. Längere Sequenzen erzeugen wesentlich ähnlichere phylogenetische Bäume als kürzere Sequenzen.

Der letzte Teil dieser Arbeit untersucht Alternativen zu klassischen phylogenetischen Analyse. Es wurden Eisen-Schwefel-Cluster formende Proteine als evolutionäre Marker genutzt, um die Ursprünglichkeit einzelner Gruppen innerhalb der Prokaryoten zu untersuchen. Diese Proteine haben wahrscheinlich eine entscheidende Rolle bei der Entstehung der ersten Organismen gespielt und ihre Anwesenheit kann als Indiz für sehr ursprüngliche Genome gedeutet werden. Die höchste Anzahl dieser Proteine pro Genom wurden in methanogenen Archaebakterien und in acetogenen bzw. sulfatreduzierenden Eubakterien gefunden.

2 Summary

Comparisons of DNA- or amino acid sequences frm the 1980s indicated that eukaryotes are more closely related to archaebacteria than to other prokaryotes. The availability of larger datasets and alternative perspectives gave rise to hints that eukaryotes are probably only to a certain amount of archaebacterial origin and have a connection to eubacteria as well. This was in line with the idea that the eukaryotic mitochondria were once free-living eubacteria and considerably contributed to the evolution of the eukaryotes. Since the majority of studies that address that question are based on the phylogenetic trees, it is of interest to investigate the reliability of these trees.

The first part of this thesis deals with the evolutionary relationship between eukaryotes and prokaryotes considering the hypotheses, that their relationship cannot necessarily be described with a phylogenetic tree. For this, 27 eukaryotic and 994 prokaryotic genomes were compared, and the distribution of the eukaryotic genes on the different prokaryotic groups was illustrated. Within phylogenetic trees it was determind which prokaryotic group was found next to the eukaryotes. Solely the archaebacteria, especially the euryarchaeota, and the α-proteobacteria these were found above average. The rickettsiales, a suborder of the α-proteobacteria, did not stand out, in contrast to former studies.

In the second part of this thesis investigates the causes and the extent of conflicting phylogenetic signals. Representative datasets for different eukaryotic and prokaryotic groups were compared and different parameters were tested for their influence. For all datasets it could been shown, that phylogenetic trees, which were made out of concatenated sequence data, only partly reflect the signal of the single gene trees. The length of the examined sequences emerged as a crucial influence. Longer sequences resulted in trees that are much more similar to each other than shorter sequences.

The last part is investigates alternatives to classical phylogenetic analysis. Iron-sulfur-cluster forming proteins were used as evolutionary markers to examine the antiquity of individual groups within prokaryotes. These proteins might had a crucial role for the formation of the first organisms and their presence could be interpreted as a indication for very primordial genomes. The highest frequency of these proteins per genome was found within methanogenic archaebacteria, acetogenic and sulfate reducing eubacteria, respectively.

3 Einleitung

3.1 Evolution der Eukaryoten und Prokaryoten

Es ist immer noch schwierig grundlegende Fragen über die Herkunft und Evolution der Prokaryoten, sowie ihre Beziehung zu den Eukaryoten, detailliert zu beantworten, selbst mit Hilfe der heutigen Menge an Sequenzdaten (Puigbò et al. 2009, Gribaldo et al. 2010, Brochier-Armanet et al. 2011). Es wurden in neueren Studien viele Parameter aufgedeckt, welche die eigentlichen phylogenetischen Untersuchungen erschweren (Jeffroy et al. 2006, Castresana 2007, Salichos & Rokas 2013). Manche davon sind methodischer Natur, meist betreffen sie die eigentliche Konstruktion der phylogenetischen Bäume (siehe Kapitel 1.4), andere haben wahrscheinlich biologische Ursachen, wie z.B. lateralen Gentransfer (siehe Kapitel 1.2 / 1.3). Daher sind die methodischen Einzelheiten bei der Rekonstruktion von phylogenetischen Bäumen immer noch wichtige Forschungsthemen. Zusätzlich wird versucht, auf phylogenetische Analysen im herkömmlichen Sinne zu verzichten oder sie durch andere Konzepte, wie Netzwerke, zu ergänzen (Dagan et al. 2010, Alvarez-Ponce et al. 2012).

3.1.1 Ursprung der Eukaryoten

Die Rekonstruktion von phylogenetischen Stammbäumen ist in der Biologie die vorherrschende Methode zur Beschreibung von Abstammungsverhältnissen. Die erste phylogenetische Analyse, welche die Prokaryoten in die beiden Reiche Archaea und Eubacteria einteilte, wurde 1977 von Woese & Fox veröffentlicht. Anhand ihrer Positionen in einem phylogenetischen Baum, gab es, zusammen mit den Eukaryoten, nunmehr drei Domänen, in welche Organismen eingeteilt wurden. Vorherige Einteilungen wurden meist anhand von zellulären Merkmalen getroffen. Diese Einteilung, welche anhand von 16S rRNS bestimmt wurde, hat bis heute ihre Gültigkeit behalten. Phylogenetische Bäume, die alle drei Domänen umfassen, wurden später mit Hilfe von Paaren von paralogen Genen gewurzelt, die auch in allen drei Domänen vorkommen (Iwabe et al. 1989, Gogarten et al. 1989). Dabei handelte es sich um die Elongationsfaktoren EF-tu/1 und EF-G/2 (Iwabe et al. 1989) und um die V- und F-type ATPasen (Gogarten et al. 1989). Dies ist möglich, da die Duplikation dieser Gene vor der Entstehung der drei Gruppen aufgetreten ist. Bei jeder dieser Analysen lag die Wurzel innerhalb der Eubakterien. Eine neuere Studie, welche einen auf phylogenetischen Netzwerken basierenden Ansatz zugrunde legte, sieht die Wurzel allerdings zwischen Archae- und Eubakterien (Dagan et al. 2010).

Die Idee, dass Eukaryoten von den „einfacheren“ Prokaryoten abstammen, wurde schon früh in der Literatur diskutiert (Allsopp 1969). Dass alle Eukaryoten bestimmte Merkmale teilen, wie das Vorhandensein von Organellen und eines Zellkerns, legte die Vermutung nahe, die Entstehung aller Eukaryoten könnte auf einen gemeinsamen Vorfahren zurückgehen (Allsopp 1969). Phylogenetische Analysen über die Verbindungen von Eukaryoten zu Prokaryoten kamen zu dem Schluss, dass Eukaryoten mit Archaebakterien in Verbindung stehen (Woese & Fox 1977, Woese et al. 1990). Die zuletzt genannte Studie setzte bei ihren Analysen ausschließlich auf rRNS Daten. Spätere Studien, die sich nicht nur auf die Analyse ribosomaler Gene beschränkten, kamen dagegen zu grundlegend anderen Schlussfolgerungen. Rivera et al. (1998) verglichen zwei eubakterielle Genome, ein archaebakterielles Genom und ein Eukaryotengenom (Hefe), miteinander. Innerhalb des Hefegenoms wurden ca. 350 Gene gefunden, die vergleichbare Homologe innerhalb der Prokaryoten hatten. Alle Gene, welche als informationsverarbeitende Gene (*informational genes*) eingestuft wurden, also Gene, deren Produkte an Translation, Trankskription und Replikation beteiligt sind, schienen fast ausschließlich von den Archaebakterien zu stammen. Wohingegen Gene die nicht an DNS-Verarbeitung beteiligt sind (*operational genes,* hier Stoffwechselgene genannt), in allen Prokaryotengruppen gefunden wurden. Später zeigte sich, dass es sogar wesentlich mehr eubakterielle Homologe in Eukaryoten gibt, als solche mit archaebakteriellem Hintergrund (Esser et al. 2004). Von allen Proteinen in der Hefe hatten ca. 380 ausschließlich Homologe in Eubakterien. Es wurden lediglich 111 Proteine gefunden, die nur Homologe in Archaebakterien hatten. Ungefähr 75% aller Hefeproteine, welche ein prokaryotisches Homolog besaßen, hatten eine höhere Sequenzähnlichkeit zu Eubakterien als zu Archaebakterien. Auch hier gehörten archaebakterielle Homologe eher zu der Gruppe der informationsverarbeitenden Gene, und eubakterielle Homologe eher zu der Gruppe der Stoffwechselgene, was die früheren Befunde bestätigte (Rivera et al. 1998). Folglich schien der Ursprung der Eukaryoten nicht nur mit den Archaebakterien assoziert zu sein, sondern mit beiden prokaryotischen Gruppen. Eine entsprechende Hypothese vermutet, dass es eine Vorläuferzelle gab, ähnlich den heutigen Archaebakterien, welche eine Endosymbiose mit einer Eubakterienzelle eingegangen ist (Esser et al. 2004, Rivera & Lake 2004). Dieser Endosymbiont hätte sein Genom im Laufe der Zeit reduziert, so dass er von der Wirtszelle abhängig wurde und sich schließlich weiter reduziert hat bis er kein eigenständiger Organismus mehr war (Timmis et al. 2004). Im Laufe dieses Prozesses sind Gene aus dem

Genoms des Endosymbioten in das Genom der Wirtszelle gewandert (Timmis et al. 2004). Aus diesen Endosymbionten sind die heutigen Mitochondrien entstanden (Gray 1993). Ein ähnliches Szenario wird für die Entstehung der Plastiden postuliert (Gray 1993).

Es ist nicht mit letzter Sicherheit geklärt, welche rezente prokaryotische Spezies mit dem damaligen Endosymbionten am nächsten verwandt ist. Wahrscheinlich stammen Plastidengene von Cyanobakterien und mitochondriale Gene von α-Proteobakterien ab (Margulis 1970, Gray 1993, Timmis et al. 2004). Dass nur wenige Gene in den Plastiden bzw. Mitochondrien zurückgeblieben sind und die meisten funktionellen Gene im Kern codiert werden (Timmis et al 2004), erschwert die Analysen zusätzlich, da man nicht mit Sicherheit sagen kann, welche Gene aus den Plastiden stammen. Die Zahl der im Mitochondrium kodierten Gene liegt zwischen drei für verschiedene *Plasmodium* Arten (Feagin 2000) und 97 in *Reclinomonas americana* (Lang et al. 1997). Bei *R. americana* handelte es sich überwiegend um Gene, deren Proteine am Elektronentransport, der ATP-Synthese und am Aufbau ribosomaler Strukturen beteiligt sind. Die restlichen Gene, welche vor der Endosymbiose im Genom des α-Proteobakteriums vorhanden waren, sind entweder verloren gegangen oder heute im Kerngenom der Wirtszelle kodiert. Warum vereinzelte Gene in den Mitochondrien zurückgeblieben sind, ist nicht vollständig geklärt. Es wird vermutet, dass manche Gene aufgrund von Unterschieden im genetischen Code oder aufgrund ihrer Hydrophobizität nicht im Kern kodiert werden können (de Grey 2005) oder dass ihre Expression direkt vom Redoxzustand ihrer Genprodukte reguliert werden muss (Allen 2003).

Studien, die sich mit dem Ursprung der Mitochondrien beschäftigten, ermittelten oftmals die Gruppe der Rickettsiales (α-Proteobakterien) als mitochondrielle Vorfahren (Sicheritz-Ponten et al. 1998, Williams et al. 2007), wobei nur die tatsächlichen, im Mitochondrium kodierten Gene untersucht wurden. Das Genom der Rickettsiales hat sich im Laufe der Zeit stark verändert und ist heute sehr klein im Vergleich zu anderern Prokaryoten (Darby et al. 2007, Merhej & Raoult 2010). Diese Tatsache könnte in Frage stellen, ob ein Vorfahre der heutigen Rickettsiales große Ähnlichkeit mit heutigen Vertretern hat, da sie einen stärkeren Wandel ausgesetzt sind als andere Prokaryoten (McCutcheon & Moran 2012). Ähnlichkeiten zu Mitochondrien könnten auch daher rühren, dass beide ähnlichen Prozessen ausgesetzt waren, die ihr Genom verkleinert haben, da Rickettsiales endozelluläre Parasiten sind, welche bekanntermaßen ein reduziertes Genom besitzen (Andersson & Kurland 1998, Moran et al. 2008). Mitochondriale Proteine aus *Chlamydomonas reinhardtii*, die im Kern kodiert sind zeigten ein anderes Bild (Atteia et al. 2009). Die betrachteten 212 Gene deuten auf eine stärkere Ähnlichkeit mit der Gruppe der Rhizobiales hin, eine weitere Untergruppe der α-

Proteobakterien. Eine neuere Studie, in welcher die Vorfahren der Mitochondrien genauer beschrieben werden sollte, kam zu dem Schluss, dass dies aufgrund von mehrdeutigen Ergebnissen sehr schwierig ist (Abhishek et al. 2011). Auch wenn frühere Studien für sich in Anspruch nahmen, zumindest die Gruppen relativ genau eingegrenzt zu haben.

Die Suche nach dem prokaryotischen Vorfahren der Chloroplasten gestaltet sich ähnlich schwer. Aufgrund variierender Ergebnisse scheint es, dass keine einzelne cyanobakterielle Gruppe in qualitativer oder quantitativer Weise eine besonders ausgeprägte Ähnlichkeit mit in Pflanzen kodierten plastidären Genen hat (Dagan et al. 2012, McFadden 2014).

3.1.2 Lateraler Gentransfer in Prokaryoten

Prokaryotische Genome sind davon geprägt, dass sie große Teile ihres genetischen Repertoires dynamisch untereinander austauschen können (Syvanen 1985, Ochmann et al. 2000). Diese horizontale Weitergabe der Gene (auch horizontaler oder lateraler Gentransfer genannt, kurz LGT) kann auf verschiedenen Wegen erfolgen. Die folgenden sind bekannt: 1) Transformation (Chen & Dubnau 2004), 2) Konjugation (Wozniak & Waldor 2010), 3) Transduktion (Zinder und Lederberg 1952) und 4) Übertragung per *gene transfer agents* (Lang et al. 2012). Die Vielzahl an möglichen Übertragungswegen ist ein Indiz dafür, wie weit verbreitet dieser Mechanismus in Prokaryoten ist.

Der statistisch ermittelte Anteil an Genen pro Prokaryotengenom, die durch lateralen Transfer beeinflusst wurden, unterscheidet sich je nach betrachteter Gruppe. Nakamura et al. (2004) fanden in dem prokaryotischen, endozellulären Parasiten *Buchnera sp.* nur 0,5% von LGT betroffene Gene, in dem Euryarchaeoten *Methanosarcina acetivorans* waren es hingegen 25%. In einer späteren Studie von Dagan et al. (2008) wurden ähnliche Schwankungen beobachtet. Dort wurden je nach Phylum Werte zwischen 4% (Chlamydiae) und 34% (δ-proteobakterien) gemessen. Somit liegen beide Studien in etwa in ähnlichen Bereichen, was das Ausmaß an LGT in Prokaryotengenomen betrifft. Außerdem zeigen beide Studien keine konstanten Raten für LGT. Diese beobachtete Bandbreite der Ergebnisse legt nahe, dass nicht alle Gene gleich stark betroffen sind und nicht alle Prokaryoten mit derselben Häufigkeit Gene austauschen.

Insbesondere Gene die an informationsverarbeitenden Prozessen beteiligt sind, werden wesentlich seltener übertragen als Stoffwechselgene (Jain et al. 1999). Im Falle der genannten Studie war die Transferrate für Stoffwechselgene viermal höher als die für informationsverarbeitende Gene. Als Erklärung dafür diente die eingeschränkte Funktionalität eines einzelnen übertragenen informationsverarbeitenden Genes. Da bei metabolischen Funktionen einzelne Proteine ihre Aufgaben oft direkt durchführen, aber die Proteine von informationsverarbeitenden Genen später oft Proteinkomplexe wie z.B. Ribosomen bilden, oder allgemein mehr Interaktionspartner haben, sind diese alleine nicht funktionsfähig (Jain et al. 1999). Neuere Studien bestätigen diesen Trend allerdings nur teilweise. Kanhere & Vingron (2009) beschrieben, dass Gene, welche zwischen Archaebakterien und Eubakterien übertragen wurden, tatsächlich vorwiegend Stoffwechselgene waren, wohingegen Gene, die zwischen Eubakterien übertragen wurden, gleichhäufig aus beiden Gruppen stammten. Als weitere Gründe für unterschiedliche Übertragungsraten sind verschiedene Ursachen im Gespräch. Laut Popa et al. (2011) werden Gene mit höherer Wahrscheinlichkeit zwischen Spezies übertragen, deren Gehalt an den Basen Guanosin und Cytosin (GC-Gehalt) ähnlich ist und deren Genom allgemein eine höhere Sequenzähnlichkeit aufweist. Außerdem erfordern die meisten Übertragungswege für genetisches Material eine enge räumliche Verbindung.

3.2 Phylogenetische Analysen

3.2.1 Einfluss von Lateralen Gentransfer

Da ein großer Teil der Gene in Prokaryoten von LGT betroffen ist, werden auch phylogenetische Rekonstruktionen stark davon beeinflusst (Philippe & Douady 2003, Zhaxybayeva et al. 2004). Zur Bestimmung der nächsten prokaryotischen Verwandten der Eukaryoten wird eine aussagekräftige Phylogenie benötigt. Viele Analysen hatten Schwierigkeiten nur einen einzelnen prokaryotischen Vorfahren der Euakryoten zu ermitteln. Meist wurden immer Ähnlichkeiten zu verschiedenen prokaryotischen Gruppen gefunden (Esser et al. 2004, Dagan et al. 2012, Atteia et al. 2009). Ob diese Ähnlichkeiten auf viele separate Fälle von LGT zurückzuführen sind, oder ob es sich lediglich um Artefakte aufgrund fehlerhafter Analysen handelt, ist nicht geklärt.

Doolittle (1998) erklärt die Anwesenheit prokaryotischer Gene innerhalb der Eukaryoten, welche nicht von den beiden wahrscheinlichsten Vorgängern, Archaebakterien und α-Proteobakterien, stammen folgendermaßen: der Eukaryotenvorgänger hat einzelne Prokaryoten als Nahrung aufgenommen und die innerhalb des Cytosols freigesetzten DNS-Fragmente wurden teilweise in das Genom integriert. Dieser stetig ablaufende Prozess ersetzte dann über die Zeit einen Großteil der Gene im Kerngenom. Aber schon damals wurde die Frage gestellt, warum dann überhaupt archaebakterielle bzw. α-Proteobakterielle Gene zurück geblieben sind. Außerdem fehlten noch Hinweise, dass alle heutigen Eukaryoten ein Mitochondrium, oder eine reduzierte Form derer, wie Mitosomen oder Hydrogenosomen besitzen, dies wurde erst später bewiesen (Martin 2005, Hackstein et al. 2006). Dies spricht dafür, dass es wahrscheinlich nie einen Eukaryoten ohne Mitochondrium gegeben hat und deshalb Protisten, welche sich z.T. phagotroph ernähren (Loftus et al. 2005, Gaudet et al. 2008), keine Modelle für einen amitochondrialen eukaryotischen Vorgänger sind.

Die Frage, wie stark LGT in heutigen Eukaryoten ausgeprägt ist, ist gerechtfertigt, um abzuschätzen, wie hoch die Wahrscheinlichkeit für eine Integration von prokaryotischen Genen in das Genom eines eukaryotischen Vorgängers ist. Es muss zwischen Transfer von Prokaryoten zu Eukaryoten und Transfer zwischen Eukaryoten untereinander unterschieden werden. Letzterer kann jedoch keine Erklärung für das Vorhandensein der vielen verschiedenen prokaryotischen Gene in den heutigen Eukaryoten liefern. Alsmark et al. (2013) haben die Genome von 13 einzelligen Eukaryoten auf Spuren von lateralem Transfer von prokaryotischen Genen untersucht. Je nach Organismus wurden zwischen einem (0,16%)

(*Encephalitozoon cuniculi*) und 63 (0,96%) (*Leishmania major*) Genen der jeweiligen Organismen als potenzielle LGTs eingestuft. All diese Funde wurden mit Hilfe von phylogenetischen Analysen belegt. Wenn explizite Gen-Spender ausgemacht werden konnten, stammten diese meist aus den Gruppen der Proteobacteria, Bacteroidetes und Firmicutes. Zwei der untersuchten Organismen (*Entamoeba histolytica*, *Dictyolstelium discoideum*) sind phagotroph, nehmen also Bakterien als Nahrung auf (Loftus et al. 2005, Gaudet et al. 2008). Dies ist einer der wenigen Mechanismen, durch den Eukaryoten prokaryotische Gene aufnehmen könnten (Keeling & Palmer 2008). In diesen Organismen sollte sich deswegen eine wesentlich höhere Anzahl an lateral übertragenen Genen finden, falls Phagotrophie ein wesentlicher Mechanismus ist, über den prokaryotische Gene in das Genom heutiger Eukaryoten gelangen könnten, was in der entsprechenden Studie eindeutig nicht der Fall war.

Auch phylogenetische Analysen zu dem Ursprung der Prokaryoten sind problematisch. Abby et al. (2012) haben diverse phylogenetische Bäume für unterschiedliche Prokaryotengruppen rekonstruiert und errechnet, wie stark die einzelnen Äste der Bäume von LGT betroffen sind. Die einzelnen Prokaryotengruppen waren erwartungsgemäß unterschiedlich stark betroffen. Innerhalb eines einzelnen phylogenetischen Baumes einer prokaryotischen Gruppe variiert die Anzahl an Ästen, die von LGT betroffen sind, zwischen 0,1% und 5%. Innerhalb eines Baumes, der auf 157 Genen aus verschiedenen prokaryotischen Gruppen basiert, variieren die Raten für einzelne Äste zwischen 0% und 32%, wobei besonders starker Einfluss von LGT (>24%) lediglich bei den Aktinobakterien und den Gammaproteobakterien gefunden wurde.

Lateraler Gentransfer in Prokaryoten, als entscheidender Mechanismus in evolutionärer Sicht, erschwert es, die genaue Herkunft einzelner Gene, seien es prokaryotische oder eukaryotische, zu rekonstruieren. Er führt zu Topologien in phylogenetischen Bäumen führt die nicht die tatsächlichen Verwandtschaftsverhältnisse der einzelnen Taxa widerspiegelt (Philippe & Douady 2003, Bapteste et al. 2004). Die Frage, ob es überhaupt möglich ist, ist noch nicht vollständig beantwortet, denn neben LGT gibt es noch weitere Faktoren, die phylogenetische Analysen erschweren, diese werden im nächsten Kapitel erläutert.

3.2.2 Fehlerquellen

Die wesentlichen Fehlerquellen bei phylogenetischen Analysen sind der unabsichtliche Vergleich von Sequenzen, welche nicht gleichen Ursprunges sind (nicht orthologe

Sequenzen), stochastische Fehler aufgrund von einer ungenügenden Menge an Information in den verwendeten Sequenzdaten und systematische Fehler (Jeffroy et al. 2006, Rodríguez-Ezpeleta et al. 2007). Darunter fallen größtenteils Probleme die sich aus der Funktionsweise der Rekonstruktionsmethoden ergeben (Bergsten 2005, Rodríguez-Ezpeleta et al. 2007, Cox et al. 2008).

Mit Hilfe verschiedener Methoden soll ein versehentlicher Vergleich von nicht-orthologen Sequenzen mit orthologen Sequenzen verhindert werden. Wird mit prokaryotischen Sequenzen gearbeitet, kann auf Datenbanken zurückgegriffen werden, in denen die Gene bereits in orthologe Gruppen eingeteilt wurden. Eine davon ist die COG Datenbank (*Cluster of orthologous groups*) (Tatusov et al. 1997, 2003). Die Gruppen wurden so konstruiert, dass Gene nur als ortholog angesehen werden, wenn diese von zwei weiteren Genen aus anderen Spezies als ähnlichste Sequenzen identifiziert wurden. Dies muss auch umgekehrt gelten. Auf diese Weise soll verhindert werden, dass paraloge Sequenzen in eine Gruppe aufgenommen werden. Außerdem wurden COGs, die aus mehreren Genen zu bestehen schienen, getrennt.

Die Frage, wie stochastische Fehler wirksam vermieden werden, ist noch nicht vollständig beantwortet. Generell wird angenommen, Fehler dieser Art ausschließen zu können, wenn eine ausreichende Menge an Genen für die Analyse genutzt wird (Jeffroy et al. 2006). Weitere Taxa hinzuzufügen scheint das phylogenetische Signal eher negativ zu beeinflussen (Rokas & Caroll 2005). Die meist genutzte Methode, um die Zahl der analysierten Gene zu erhöhen, ist die Verkettung von Alignments einzelner Gene zu einen einzigen langen Alignment (Baldauf et al. 2000, Brown et al. 2001, Cicarelli et al. 2006, Cox et al. 2008). Da sich die phylogenetischen Bäume der einzelnen Gene allerdings stark unterscheiden können, stellt sich die Frage, wie vertrauenswürdig der abgeleitete verkettete Baum ist. Wird ein Baum aus verketteten Alignments mit den Bäumen der einzelnen Gene verglichen, so sind die einzelnen Kanten innerhalb des verketteten Baumes selten bis überhaupt nicht in den einzelnen Bäumen zu finden (Salichos & Rokas 2013). Es muss besonders beachtet werden, dass die verketteten Bäume meist hohe Bootstrapwerte haben. Diese zeigen die Übereinstimmung mit Bäumen, die aus Stichproben des eigentlichen Alignments erstellt wurden an und sollten auf eine statistische Glaubwürdigkeit dieser Bäume hindeuten (Felsenstein 1985). Die genannten Ergebnisse stellen den Nutzen dieser Werte allerdings in Frage. Mehrere Arbeiten fanden insbesondere für die tieferen Knoten innerhalb von phylogenetischen Bäumen eine mangelnde statistische Unterstützung, sowohl bei

Eukaryoten (Salichos & Rokas 2013) als auch bei Prokaryoten (Creevey et al. 2004, Bapteste et al. 2008).

Systematische Fehler bei phylogenetischen Rekonstruktionen, verursacht z.B. durch starke Unterschiede der Basen oder in der Aminosäurezusammensetzung der untersuchten Sequenzen oder durch multiple Substitutionen an der gleichen Stelle, lassen sich nicht durch das Verbinden von mehreren Datensätzen verhindern (Jeffroy et al. 2006). Die Vielzahl an möglichen Fehlerquellen und die Tatsache, dass wahrscheinlich nicht alle durch einen einzigen Ansatz ausgeschlossen werden können, legt nahe, auch Alternativen zur klassischen Rekonstruktion phylogenetischer Stammbäume zu suchen (Creevey et al. 2004, Bapteste et al. 2005, Galtier & Daubin 2008).

3.3 Alternativen zu phylogenetischen Bäumen

3.3.1 Metalloproteine als evolutionäre Marker

Eine zusammenfassende Theorie besagt, dass die ersten lebenden Organismen wahrscheinlich chemolithoautotroph, thermophil und anaerob lebend waren (Wächtershäuser 1998, Martin et al. 2008, Fuchs 2011). Im Sinne dieser Hypothese haben sich erste Stoffwechselwege und nachfolgende Organismen an der Oberfläche von im Ozean liegenden warmen Quellen formiert. Das dort vorkommende Eisen hat dabei wahrscheinlich eine wichtige Rolle in frühen Stoffwechselwegen gespielt (Barros & Hoffman 1985, Wächtershäuser 1998, Martin et al. 2008, Fuchs 2011). In einer Umgebung in der es noch keine Proteine gab, könnte es bei Redoxreaktionen eine Rolle gespielt haben und dann später in die entsprechenden protein-abhängigen Prozesse eingebunden worden sein (Martin & Russell 2003). Ein Vorteil der so entstandenen Metalloproteine ist, dass sie sehr klein und einfach zu synthetisieren sind (Hall et al. 1971). Sie könnten bei höheren Temperaturen, wie sie wahrscheinlich in der Umgebung in der die ersten Organismen entstanden sind herrschten (Wächtershäuser 1998, Martin et al. 2008), die Aufgaben von NAD/P(H) übernommen haben (Daniel & Danson 1995), welches bei diesen Temperaturen nicht funktionsfähig wäre. Eisenhaltige Proteine stellen also eine sehr ursprüngliche Art von Proteinen dar und sind deshalb aus evolutionärer Sicht sehr interessant.

Eine Datenbanksuche unter allen Enzymen mit bekannter Struktur ergab, dass ungefähr 40% dieser Enzyme als katalytisches Zentrum ein Metallion besitzen (Andreini et al. 2008). Unter diesen wurden besonders häufig, in absteigender Zahl Magnesium, Zink, Eisen und Mangan gefunden. Zink ist innerhalb von Prokaryoten besonders in Enzymen enthalten (83%) und nur zu einem geringen Anteil in Proteinen, welche die Transkription regeln (8%). Eisen-Proteine sind häufig in Oxireduktasen enthalten und am Elektronentransfer (51%), an anderen Enzymaktivitäten (35%) und nur in geringem Maße an Gen-Expression und Reparatur (7%) beteiligt (Andreini et al. 2009). Major et al. (2004) haben die Verteilung von Eisen-Schwefel-Cluster formenden Proteinen innerhalb von 120 Prokaryoten untersucht. Diese, auch Ferredoxine genannten Proteine, enthalten Eisen-Schwefel-Cluster (Fe-S-Cluster), welche als Elektronentransporter bei Redoxreaktionen dienen (Fuchs 2011). Neben einer starken Korrelation von Genomgröße und Anzahl an Fe-S-Cluster bildenden Proteinen fanden sie heraus, dass methanogene Archaebakterien eine wesentlich höhere Anzahl an diesen Proteinen besitzen, im Gegensatz zu den anderen prokaryotischen Gruppen. Da diese

Fe-S-Cluster bildenden Proteine eine sehr ursprüngliche Form von Proteinen darstellen, könnte ihr Vorhandensein in einem Organismus auch auf ein sehr ursprüngliches Proteom/Genom hindeuten (Hall 1971, Liu et al. 2012). Somit wäre es möglich, eine grobe Einteilung von prokaryotischen Gruppen vorzunehmen, ohne einen phylogenetischen Baum zu benötigen.

3.3.2 Phylogenetische Netzwerke

Alle Netzwerke bestehen aus zwei Grundelementen, den Knoten und den Kanten. Dabei sind Knoten durch Kanten miteinander verknüpft (Newman 2003). In der Biologie lassen sich unzählige Beispiele für Systeme finden, die als Netzwerke dargestellt werden können. Wobei biologische Entitäten, wie Gene oder Organismen, meist als Knoten dienen und ihre Beziehung zueinander über die Kanten repräsentiert wird. Die Interaktionen von Proteinen untereinander lassen sich z.B. als Netzwerk darstellen und können so gegebenenfalls Aufschluss über unbekannte Funktionen geben (Vazquez et al. 2003). Auch auf der Genebene lassen sich Netzwerke erstellen, um z.B. Rückschlüsse zu erhalten wie und ob sie miteinander interagieren (Brazhnik et al. 2002).

Da in phylogenetischen Bäumen, welche auch eine Art von simplen Netzwerk darstellen, ein Blatt nur zu einem direkten Vorfahren zurückzuführen ist, sind phylogenetische Bäume für die Darstellung nicht-baumartiger evolutionärer Prozesse, wie Endosymbiose oder lateralem Gentransfer, nicht geeignet. Auch ist es nicht möglich, widersprüchliche phylogenetische Signale in einem bifurzierenden Baum darzustellen. Netzwerke die nicht zwingend bifurzierend aufgebaut sind, bieten die Möglichkeit, einen Knoten, welcher ein einzelnes Taxon oder Gen repräsentiert, mit mehreren anderen Knoten in Verbindung zu setzen. Daher können sie sowohl nicht-baumartigen Prozesse darstellen als auch widersprüchliche Signale aus mehreren unterschiedlichen Bäumen in einer Abbildung vereinen. Es gibt mehrere Methoden solche Netzwerke zu konstruieren. Entweder werden einem phylogenetischen Baum zusätzliche Äste hinzugefügt (Makarenkov & Legendre 2004, Bryant & Moulton 2003) oder es wird auf eine baumartige Struktur verzichtet und nur die tatsächlich vorhandenen Ähnlichkeiten dargestellt (*similarity* Netzwerke) (Popa et al. 2011, Dagan et al. 2008). Letztere werden benutzt, um LGT-Ereignisse darzustellen bzw. zu rekonstruieren. Sie ersetzen keine phylogenetischen Analysen. Wird auf einen baumartigen Unterbau verzichtet, verschwindet auch die Information über diesen Teil der evolutionären Vorgeschichte.

Ob phylogenetische Netzwerke einen Ersatz für phylogenetische Bäume darstellen können, hängt meist von der zugrundeliegenden Fragestellung ab. Meist werden Widersprüche in den Daten aufgezeigt und keine eindeutigen phylogenetischen Zusammenhänge beschrieben (Bryant & Moulton 2003). Eine Analyse von Netzwerken, welche mit einer Vielzahl von Taxa erstellt wurden, zeigte, dass diese Ergebnisse oft schwer zu interpretieren sind, besonders wenn sich die Bäume stark unterscheiden und/oder viele Taxa enthalten (Layeghifard et al. 2013). Da aber Netzwerke in vielen verschiedenen naturwissenschaftlichen Disziplinen angewendet werden, gibt es eine entsprechend große Menge an Ressourcen, die genutzt werden können um Netzwerke und ihren Aufbau zu analysieren (Strogatz 2001, Fortunato 2010). Darunter fallen Analysen die eigentliche Struktur und Eigenschaften eines Netzwerkes analysieren und es ermöglichen diese mit anderen Netzwerken zu vergleichen (Strogatz 2001). Ein weiteres großes Teilgebiet ist die Analyse von *communities* (engl. Gemeinschaften) in Netzwerken. Dabei handelt es sich um Gruppen von Knoten in einem Netzwerk die mehr Verbindungen untereinander haben, als zu dem restlichen Netzwerk (Fortunato 2010).

4 Zielsetzung

Viele Analysen zum Ursprung der Eukaryoten wurden mit Hilfe von kleinen Datensätzen insbesondere mit einer geringen Anzahl an eukaryotischen Genomen durchgeführt (Esser et al. 2004, Woese 1990). In vorherigen Studien wurden teilweise auch lediglich Sequenzvergleiche benutzt und keine phylogenetischen Analysen durchgeführt (Esser et al. 2004). Deshalb war es das Ziel dieser Arbeit, eine möglichst große Zahl an Genen und Organismen mit in die Analysen einzubeziehen. Es sollte außerdem dem chimären Charakter der Eukaryoten Rechnung getragen und ein Netzwerk erstellt werden, welches die Verbindungen der Eukaryoten zu den einzelnen prokaryotischen Gruppen darstellt, die höchstwahrscheinlich zu deren Entstehung beigetragen haben (Esser et al. 2004, Rivera & Lake 2004).

Ein Großteil der Analysen, die den Ursprung der Eukaryoten oder Prokaryoten betrachten, benutzen Bäume, welche aus verketteten Alignments erstellt wurden (Baldauf et al. 2000, Ciccarelli et al. 2006, Cox et al. 2008). Ein Ziel dieser Dissertation war es herauszufinden, wie vertrauenswürdig diese Bäume sind, um spezifische Aussagen über ihre Topologien treffen zu können. Die zugrunde liegenden Mechanismen, welche Einfluss auf die entstehenden Topologien haben, wurden untersucht. Die Ergebnisse widersprechen sich jedoch (Rokas et al. 2003, Galtier 2007). Es wurden phylogenetische Bäume von Eu- und Prokaryoten, sowie mehrere Parameter, welche Einfluss auf mögliche Inkongruenzen nehmen könnten, wie Sequenzlänge, Alignmentlänge oder Sequenzidentität, verglichen.

Als Alternative zu phylogenetischen Methoden, besonders wenn sehr lang zurückliegende Ereignisse beschrieben werden sollen, wurde unter anderem das Vorhandensein von Eisen-Schwefel-Cluster bildenden Proteinen in Prokaryoten analysiert. Dies geschah unter der Prämisse, dass diese Eigenschaft besonders in alten prokaryotischen Gruppen übermäßig ausgeprägt ist (Hall 1971, Liu et al. 2012). Ältere Studien hierzu hatten nur einen beschränken Ausschnitt aller prokaryotischen Gruppen untersucht (Major et al. 2004). Vor diesen Hintergrund sind die hier vorgestellten Untersuchungen erfolgt.

5 Publikationen

5.1 An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin

Thorsten Thiergart[1], Giddy Landan[1], Marc Schenk[1], Tal Dagan[1], William F. Martin[1]

[1] Institut für molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Deutschland

Dieser Artikel wurde 2012 in *Genome Biology and Evolution* veröffentlich Ausgabe 4

Beitrag von Thorsten Thiergart:

Ich habe sämtliche Daten, die für Analysen benötigt wurden, erstellt (Einteilung der Proteine in funktionelle Gruppen, Zusammenstellung der homologen Gruppen, Erstellung der phylogenetischen Bäume). Mit diesen Daten habe ich die Analysen zur Einteilung der phylogenetischen Bäume (Abbildung 1) und die Bestimmung der prokaryotischen Schwestergruppen durchgeführt (Abbildung 2, 5 & 7). Diese Ergebnisse wurden von mir näher untersucht, um die jeweiligen funktionellen Gruppen mit einzubeziehen (Abbildung 4) und die Ergebnisse anhand eines phylogenetischen Netzwerkes zu verdeutlichen (Abbildung 8). Die kompletten Kapitel, welche die verwendeten Methoden erläutern und der Teil, welcher die Ergebnisse beschreibt wurden von mir und Prof. Dr. Tal Dagan verfasst. Zusätzlich bin ich Koautor der restlichen Kapitel.

An Evolutionary Network of Genes Present in the Eukaryote Common Ancestor Polls Genomes on Eukaryotic and Mitochondrial Origin

Thorsten Thiergart, Giddy Landan, Marc Schenk, Tal Dagan, and William F. Martin*

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

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

Accepted: 14 February 2011

Abstract

To test the predictions of competing and mutually exclusive hypotheses for the origin of eukaryotes, we identified from a sample of 27 sequenced eukaryotic and 994 sequenced prokaryotic genomes 571 genes that were present in the eukaryote common ancestor and that have homologues among eubacterial and archaebacterial genomes. Maximum-likelihood trees identified the prokaryotic genomes that most frequently contained genes branching as the sister to the eukaryotic nuclear homologues. Among the archaebacteria, euryarchaeote genomes most frequently harbored the sister to the eukaryotic nuclear gene, whereas among eubacteria, the α-proteobacteria were most frequently represented within the sister group. Only 3 genes out of 571 gave a 3-domain tree. Homologues from α-proteobacterial genomes that branched as the sister to nuclear genes were found more frequently in genomes of facultatively anaerobic members of the rhiozobiales and rhodospirilliales than in obligate intracellular ricketttsial parasites. Following α-proteobacteria, the most frequent eubacterial sister lineages were γ-proteobacteria, δ-proteobacteria, and firmicutes, which were also the prokaryote genomes least frequently found as monophyletic groups in our trees. Although all 22 higher prokaryotic taxa sampled (crenarchaeotes, γ-proteobacteria, spirochaetes, chlamydias, etc.) harbor genes that branch as the sister to homologues present in the eukaryotic common ancestor, that is not evidence of 22 different prokaryotic cells participating at eukaryote origins because prokaryotic "lineages" have laterally acquired genes for more than 1.5 billion years since eukaryote origins. The data underscore the archaebacterial (host) nature of the eukaryotic informational genes and the eubacterial (mitochondrial) nature of eukaryotic energy metabolism. The network linking genes of the eukaryote ancestor to contemporary homologues distributed across prokaryotic genomes elucidates eukaryote gene origins in a dialect cognizant of gene transfer in nature.

Key words: endosymbiosis, eukaryotes, phylogenomics, lateral gene transfer, horizontal gene transfer, endosymbiotic gene transfer.

Introduction

Although the evolutionary details of the prokaryote-to-eukaryote transition are still incompletely resolved (Brown and Doolittle 1997; Koonin 2012), the crucial role that mitochondria played in that transition is becoming increasingly evident (Lane and Martin 2010; Lane 2011). Presently, two main categories of competing hypotheses address the prokaryote-to-eukaryote transition: autogenous and symbiogenic (Maynard-Smith and Szathmáry 1995; Embley and Martin 2006; Pisani et al. 2007; Lane 2009). Autogenous models posit that eukaryotes arose from a single ancestral lineage via mutation in a gradualist type of evolutionary process. Symbiogenic models posit that eukaryotes arose via a symbiotic association of divergent prokaryotic cells, with symbiosis (and gene transfer from endosymbiont to host in some formulations) forging the prokaryote-to-eukaryote transition, with phases of evolutionary innovation marked by distinctly non-gradualist characteristics. Both the autogenous and the symbiogenic categories harbor a number of specific competing alternative hypotheses, respectively, each of which in turn generates testable predictions about the phylogenetic affinities of eukaryotic genes to prokaryotic homologues.

Among the autogenous models, three are currently discussed. The neomuran hypothesis (Cavalier-Smith 1975) argued in its original formulation that eukaryotes arose from cyanobacteria through conventional mutation and selection processes. In more modern formulations, the neomuran hypothesis posits that eukaryotes arose from actinobacteria (Cavalier-Smith 2002); hence, it predicts that eukaryotic genes should uncover detectable affinities to homologues encountered in contemporary actinobacterial genomes. A second and more recent—but far less explicit—autogenous model has it that eukaryotes are descended from planctomycetes or the planctomycete–verrumicrobia–chlamydia (PVC) group (Devos and Reynaud 2010). It predicts eukaryotic genes to uncover widespread sequence similarities to planctomycete homologues, a prediction that is so far unfulfilled (McInerney et al. 2011). A third autogenous theory has it that eukaryotes represent the ancestral state of cell organization and that prokaryotes are derived from eukaryotes via a process that was originally called streamlining (Doolittle 1978) and later called thermoreduction (Forterre 1995). It predicts a three-domain topology for genes shared by eukaryotes and prokaryotes (Forterre and Gribaldo 2010). Common to autogenous theories is the assumption that mitochondria had no role in the prokaryote-to-eukaryote transition, a premise that has become increasingly problematic with data accrued over the last 10 years indicating 1) that mitochondria were present in the eukaryote common ancestor (Embley et al. 2003; van der Giezen 2009) and 2) that, for reasons of bioenergetics, mitochondria were strictly required for the origin of the molecular traits that make eukaryotic cells complex in comparison to their prokaryotic counterparts (Lane and Martin 2010).

Symbiogenic hypotheses can be generally divided into two subcategories. The first subcategory invokes an endosymbiosis to derive a mitochondrion-lacking cell that possess a nucleus, whereby the nuclear compartment is usually viewed as deriving from an endosymbiotic prokaryote. Among current formulations that derive the nucleus from endosymbiosis, the assumed symbiotic partners include 1) a *Thermoplasma*-like host and a spirochaete endosymbiont (Margulis et al. 2006), 2) a Gram-negative host and a crenarchaeal endosymbiont (Lake and Rivera 1994; Gupta and Golding 1996), 3) a δ-proteobacterial host and a methanogen-like endosymbiont (Moreira and Lopez-Garcia 1998), 4) a γ-proteobacterial host and a *Pyrococcus*-like endosymbiont (Horiike et al. 2004), and 5) a planctomycete host and a *Crenarchaeum*-like nucleogenic endosymbiont (Forterre 2011). Like autogenous theories, these models assume that mitochondria had no role in the prokaryote-to-eukaryote transition *sensu strictu* because a nucleus-forming endosymbiosis is presumed to have generated the eukaryotic lineage, one member of which then acquires the mitochondrion and the other members of which implicitly become extinct because all eukaryotic lineages possess a mitochondrion or did in their evolutionary past (van der Giezen 2009). As discussed elsewhere, there are many serious fundamental problems with the view that the nucleus was ever a free-living prokaryote (Martin 1999a, 2005; Cavalier-Smith 2002).

The second major subcategory among symbiogenic theories invokes endosymbiosis to derive the mitochondrion directly in a prokaryotic host, without any earlier additional symbiotic cell mergers. Because eukaryotes have an archaebacterial genetic apparatus (Langer et al. 1995; Rivera et al. 1998; Cox et al. 2008; Koonin 2009; Cotton and McInerney 2010) and because mitochondria are clearly derived from an endosymbiotic proteobacterium (Gray et al. 1999; Atteia et al. 2009), these theories posit that the host for the origin of mitochondria was a "garden variety" archaebacterium, either related to *Thermoplasma* (Searcy 1992) or to hydrogen-dependent archaebacteria, with a physiology perhaps similar to methanogens (Martin and Müller 1998; Vellai et al. 1998). In these models, the nucleus arises after the origin of mitochondria and in an autogenous manner that does not require additional endosymbioses (Martin and Koonin 2006).

All of the foregoing theories generate predictions with regard to the branching patterns expected in trees comprising both prokaryotic and eukaryotic genes. Comparatively, few tests of those predictions using alignments for many genes from complete genome data have been reported. Using pairwise sequence similarity matrices, Esser et al. (2004) found that among the 850 yeast genes having homologues among the small prokaryotic sample of 15 archaebacterial and 45 eubacterial genomes, roughly 75% of yeast nuclear-encoded proteins were more similar to eubacterial homologues than to archaebacterial homologues. Using a supertree approach, Pisani et al. (2007) found that when the signal stemming from nuclear genes of cyanobacterial origin in plants is removed from the eukaryote data set, eukaryotes branch among α-proteobacteria, likely reflecting the signal of nuclear genes of mitochondrial origin, and when that signal is removed, eukaryotes then branched with the *Thermoplasma* lineage among euryarchaeotes. Cox et al. (2008) examined the concatenated phylogeny of genes corresponding to the informational class (information storage and processing) and found evidence linking the eukaryote host lineage to crenarchaeotes, in line with the prediction of the eocyte theory (Lake 1988). Yutin et al. (2008) examined individual phylogenies and obtained conflicting results with respect to the euryarchaeal, crenarchaeal, or ancestral archaebacterial origin of eukaryotic informational genes. Kelly et al. (2011) examined genes that link eukaryotes to archaebacteria and found evidence linking the eukaryotes to the *Crenarchaeum/Nitrosopumilus* (thaumarchaeal) lineage of archaebacteria.

Autogenous models and symbiogenic models also differ with respect to the predictions that they make about the ancestor of mitochondria and (in some cases) about the nature of eubacterially related genes in eukaryotic genomes. Several recent studies have addressed the origin of mitochondria, but have focused on sequences residing in mitochondrial DNA (mtDNA) (Thrash et al. 2011; Brindefalk et al. 2011; Georgiades and Raoult 2011). Those studies delivered widely conflicting results because of the small sample of mitochondrion-encoded protein available—about 55 at most that can be used to generate trees (Esser et al. 2004)—and the phylogenetic biases introduced by the rapid evolutionary rate and AT richness of mtDNA, which can cause mtDNA-encoded proteins to artefactually group together with homologues from rapidly evolving and AT-rich bacterial lineages, like Rickettsiales (Thrash et al. 2011). Nuclear-encoded proteins should, in principle, be less subject to AT bias and the elevated substitution rate of mitochondrially encoded proteins. They provide a larger gene sample for investigation of mitochondrial origin or of organelle origins in general (Deusch et al. 2008) but that does not mean that they are fundamentally immune to bias or phylogenetic error.

Investigations of mitochondrial origin using many nuclear genes are still scarce. Trees for pyruvate dehydrogenase subunits pointed to *Rickettsia*-like ancestors (Kurland and Andersson 2000). Trees for Krebs cycle and glyoxylate cycle enzymes (Schnarrenberger and Martin 2002) as well as trees for >200 nuclear-encoded mitochondrial proteins from *Chlamydomonas* point more frequently to origins from generalist, facultatively anaerobic α-proteobacteria (Atteia et al. 2009), than to *Rickettsia*-like ancestors, whereby many proteins indicated a eubacterial, but not a specifically α-proteobacterial ancestry. Recent analysis of 86 yeast nuclear-encoded mitochondrial proteins produced a similar result: some point to *Rickettsia*-like ancestors and some point to facultatively anaerobic *Rhodobacter*-like ancestors (Abhishek et al. 2011). Although many mitochondria function anaerobically (Tielens et al. 2002; Atteia et al. 2006; Mus et al. 2007), nuclear genes for anaerobic mitochondrial energy metabolism cannot implicate *Rickettsia*-like ancestors because Rickettsias are strict aerobes that harbor no genes of anaerobic energy metabolism for comparison. Using an automated pipeline for phylogenetic trees, Gabaldon and Huynen (2003) identified 630 nuclear-encoded protein families that trace to the ancestor of mitochondria, although the α-proteobacteria themselves often failed to form a monophyletic group in that study, pointing to the role of LGT in prokaryote evolution.

Comparative genomics should permit a test of different models for eukaryote origins. Genes suited to such tests are those that are preserved in eukaryotic nuclear genomes that 1) have homologues in prokaryotes and 2) reflect eukaryote monophyly as evidence of their presence in the last eukaryote common ancestor (LECA). Here we have assembled alignments for 712 gene families from 27 eukaryotes, 926 eubacteria, and 68 archaebacteria in order to address the question: given the present (limited) genome sample, how many eukaryotic genes with prokaryotic homologues actually reflect a single origin? Those that trace to the eukaryote common ancestor allow us to furthermore ask: in which prokaryotic lineages are those genes currently found? We then contrast the results with the predictions generated by competing theories for eukaryote origins, but do so in a modern context taking into account the circumstance that free-living prokaryotes have been undergoing LGT during the time since eukaryotes arose, such that the collection of genes that eukaryotes acquired from the ancestor of mitochondria reflects a sample of ancient prokaryote gene diversity, not a collection of genes that we would expect to find in any contemporary free-living prokaryote (Martin 1999b; Esser et al. 2007; Richards and Archibald 2011). In that sense, the concept of prokaryotic lineages is poorly defined when it comes to the phylogeny of individual genes (Doolittle and Bapteste 2007; Bapteste et al. 2009), a circumstance that figures prominently in the interpretation of the results.

Methods

Data

Proteomes of 27 eukaryotes and 994 prokaryotes were retrieved from the public databases. The following proteomes were downloaded from RefSeq database (Pruit et al. 2005): *Hydra magnipapillata*, *Ciona intestinales*, *Caenorhabditis elegans*, *Physcomitrella patens*, all 994 prokaryotes (11/2009 version), *Oryza sativa*, and all fungi and animal sequences (02/2008 version). Additional proteomes were retrieved from the JGI database (http://www.jgi.doe.gov/): *Populus trichocarpa* (v1.1), *Ostreococcus tauri* and *Chlamydomonas reinhardtii* (v3.1). The *Arabidopsis thaliana* proteome was downloaded from TAIR project (Swarbreck et al. 2008). The *Cyanidoschyzon merolae* proteome was downloaded from the genome project Web site (Matsuzaki et al. 2004). The complete list of genomes used is given in supplementary Table S1, Supplementary Material online.

Clusters of Homologous Proteins

For the reconstruction of eukaryotic protein families, a reciprocal best Blast (v2.2.20; Altschul et al. 1997) hit procedure was used (Tatusov et al. 1997). Only BBH having an *e*-value $\leq 1 \times 10^{-10}$ were retained. Pairs of reciprocal BBHs were aligned using the Needleman–Wunsch algorithm by the *needle* program included in the EMBOSS package (Rice et al. 2000). Homologous pairs having ≤40% identical amino acids were excluded from the data set. All remaining eukaryotic homologues were clustered into protein families with MCL (v1.008 Enright et al. 2002) using the default parameters. This analysis yielded 37,101 protein families comprising 165,329 proteins. Excluding protein families comprising less than 4 members in total (18,116) or less than 3 main eukaryotic groups (animals, fungi, algae, plants) resulted in 1,122 protein families retained for further analysis. To find prokaryotic homologues to all clustered eukaryotic proteins, these proteins were BLASTed against 994 prokaryotic proteomes. A total of 367 clusters had no homologues within the prokaryotic genomes. Prokaryotic homologues were added to the clusters using a reciprocal BBH procedure applying an *e*-value threshold of $\leq 1 \times 10^{-10}$ and $\geq$30% identical amino acids. All prokaryotic hits per eukaryotic query protein were added to the respective protein family, and redundant prokaryotic proteins were omitted. In case of multiple prokaryotic homologues in different strains of the same species, only one homologue having the highest sequence similarity was included in the protein family. The analysis resulted in 755 protein families of eukaryotic sequences and their prokaryotic homologues.

The functional classification of all protein families is based on the KOG database (Tatusov et al. 2003). A total of 626 protein families that overlapped with KOG clusters were annotated to have the same function as the matching KOG. The remaining protein families were manually classified by sequence similarity to known KOGs using the KOGnitor tool (http://www.ncbi.nlm.nih.gov/COG/grace/kognitor.html). A total of seven protein families had no homologues in the KOG database and were annotated as unknown function.

Phylogenetic Analysis

The protein families were aligned with MAFFT (v6.832b; Katoh et al. 2002) using Blosum62 substitution matrix (Henikoff and Henikoff 1992). Alignment quality was tested using the HoT procedure (Landan and Graur 2007) and 20 alignments having SPS <70% were excluded from the data set. Phylogenetic trees were calculated from the alignments using maximum-likelihood approach with RAxML (v7.0.4; Stamatakis 2006). Substitution rate per site was estimated from a gamma distribution with four discrete rate categories and the WAG substitution matrix (Whelan and Goldman 2001). The proportion of invariable sites was estimated from the data. Eukaryotic monophyly within the reconstructed trees was tested using an in-house PERL script. A group is considered as monophyletic if there exists a bipartition (branch) in the tree that includes only species from that group. Thus in trees testifying eukaryotic monophyly, there exists a branch that splits between eukaryotes and prokaryotes.

Single eukaryotic sequences branching within the prokaryotic clade were manually tested for possible bacterial contaminations using BLAST at NCBI Web site (http://blast.ncbi.nlm.nih.gov/Blast.cgi) against the *nr* database. We manually excluded sequences that were obvious bacterial contaminations where possible. For example, 39 genes annotated as belonging to the Hydra genome actually belong to the genome of the eubacterial endosymbiont of *Hydra, Curvibacter spec*. (Chapman et al. 2010). Another eight genes in the *Populus* genome are >90% identical at the amino acid level to prokaryotic proteins and were classified as a putative bacterial contamination, as were two sequences from the *Bos taurus* genome. In several cases, the bacterial contamination "*Hydra*" sequence was the only representative from the metazoa, such that in total 24 alignments were excluded from the analysis, leaving 712 alignments for analysis, which are available upon request.

To specify the sister group to the eukaryotes, there are several possibilities. In each tree, the branch connecting the monophyletic eukaryotic clade to the prokaryotes serves to splits the prokaryote clade into two groups, one of which is selected as the sister group. We tested several criteria to decide which is the sister group: the group with the smaller number of sequences, the group with the smaller average distance to the eukaryotes, and the group that does not include the root after midpoint rooting. The sister group frequencies inferred are robust to the three different criteria (supplementary fig. S1, Supplementary Material online). For simplicity, we used the criterion of the smaller clade to specify the sister group, use of other criteria would not alter the results.

We did not initiate exhaustive topology searches or likelihood optimization efforts searches beyond those performed by RAxML in order to find more or fewer cases of eukaryote monophyly in the data. For those genes that did reflect eukaryote monophyly, we were interested in the identity and nature of the genomes harboring the nearest prokaryotic neighbors.

Network and Reference Tree Reconstruction

The prokaryotic clade of the universal reference tree was retrieved from Popa et al. (2011) that reconstructed it from the ribosomal RNA operon sequences within a taxonomic framework. The tree was reconstructed for prokaryotic main taxa by using the consensus sequences of the ribosomal RNA genes for each bacterial group. The groups correspond to the phyla of the bacterial species or the class in the case of Proteobacteria and Firmicutes. Three archaebacterial groups including the Nanoarchaeota, Thaumarchaeota, and Korarchaeota that were missing in the Popa et al. (2011) tree were included according to their phylogenetic position in Makarova et al. (2010). The eukaryotic clade is a consensus tree reconstructed from 12 gene trees that include all eukaryotic species in our analysis (excluding the highly reduced genome of *Encephalitozoon cuniculi*). The network in figure 8 combines 571 sister group specifications as lateral edges connecting vertical edges of the reference tree. The edge weight is the frequency in which species in the given prokaryotic taxon branched within the sister group to the eukaryotic clade. The universal tree with lateral edges signifying prokaryotic contributions was depicted with a MatLab© script.

Results

Eukaryotic Genes Reflect Single Origins

Clustering the eukaryotic and prokaryotic proteins by sequence similarity yielded 712 inter-kingdom families of homologous proteins. All protein families include at least three of the main eukaryotic groups: animals, fungi, algae, and plants. Phylogenetic trees were reconstructed from the protein families by a maximum-likelihood approach and rooted on the branch that maximizes the ratio between eukaryotes and prokaryotes in the resulting clades. The resulting rooted trees were classified into four categories according to the branching pattern of the eukaryotic species within the tree (fig. 1).

Most of the tree topologies (571/712, 80.2%) recovered the eukaryotic genes as a monophyletic clade. The remaining trees fell into three different categories in similar shares. In polyphyletic trees (46, 6.5%), there exists a branch that splits the tree into a eukaryotes-only clade and a prokaryotic clade that includes a few eukaryotes (fig. 1*b*). The frequency of eukaryotes in the prokaryotic clade ranged between 1 and 6 species. In 12 of those polyphyletic trees, the eukaryotes branching within the prokaryotic clade were photosynthetic eukaryotes branching as the nearest neighbors of cyanobacteria, the expected result for genes that were transferred from the ancestor of plastids into the nuclear genome of photosynthetic eukaryotes (Timmis et al. 2004). Genes in this group include the ClpB heat shock protein Kda100 (*C. reinhardtii*, *C. merolae*, and *O. tauri*) and phosphoglycerate kinase (Brinkmann and Martin 1996) from *C. reinhardtii*, *C. merolae*, and *A. thaliana* (supplementary Table 2, Supplementary Material online). Most (34) of the remaining trees in this category included a single eukaryote within the prokaryotic clade, with many proteins being annotated as "predicted protein" or "hypothetical protein." Genes of the red algae *C. merolae* branched frequently within the prokaryotic clade with 18 tree topologies placing this species within the prokaryotic clade next to a noncyanobacterial nearest neighbor (supplementary Table 2, Supplementary Material online).

Paraphyletic trees are the mirror image of polyphyletic trees, as they include a branch that splits the tree into a prokaryotes-only clade and a eukaryotic clade that includes several prokaryotes (fig. 1c). The number of prokaryotes in the latter clade ranged between 1 and 22 (see distribution in supplementary fig. S2, Supplementary Material online). In 29 of the 46 paraphyletic trees, prokaryotes branching within the eukaryotic clade included one or more eukaryote-associated microbes (e.g., human pathogens and plant endosymbionts). These could be prokaryote acquisitions of eukaryotic homologues, as has been previously observed, for example for tubulin (Pilhofer et al. 2007). In six trees, all of the prokaryotes that branched within the eukaryotic clade were cyanobacterial species next to plants or algae. The mixed prokaryotic–eukaryotic branching pattern of the remaining 49 (6.9%) trees did not enable a clear cut rooting and classification of the trees into one of the above categories (fig. 1*d*). In trees with eukaryote monophyly, operational genes and informational genes are present in a ratio of 362:209, in the trees where eukaryotes are nonmonophyletic, operational genes are significantly enriched (129:12, $P < 0.0001$).

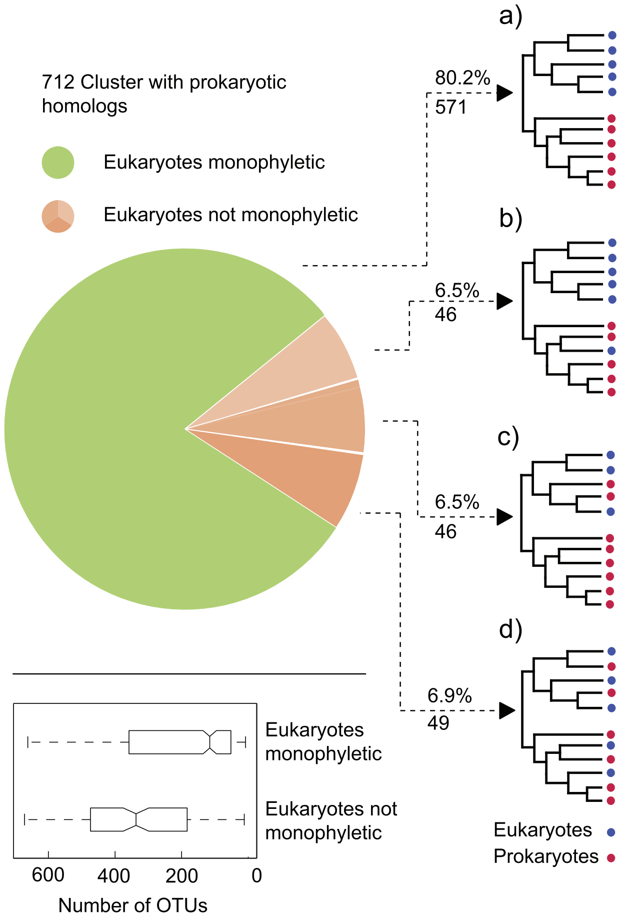


Fig. 1.—A distribution of topologies among 712 inter-kingdom trees. The schematic trees on the right symbolize the branching patterns of eukaryotic and prokaryotic species observed in each category. (*a*) Eukaryotes and prokaryotes form monophyletic clades. (*b*) Eukaryotes are polyphyletic (≤6 eukaryotic species branch within the prokaryotic clade) and prokaryotes are paraphyletic. (*c*) Eukaryotes are paraphyletic (between 1 and 22 prokaryotic species branch within the eukaryotic clade) and prokaryotes are polyphyletic. (*d*) A mixed topology of eukaryotes and prokaryotes. The boxplots in the lower panel show the distribution of the number OTUs in trees where the eukaryotes are 1) monophyletic or 2) not monophyletic.

LECA Genes with Prokaryotic Homologues

For 571 protein families, ML trees indicate eukaryote monophyly. Because we only considered trees that spanned the opisthokont/archaeplastida divide, and because of current views for the placement of the eukaryotic root (Hampl et al. 2009), eukaryote monophyly indicates their presence in LECA. The distribution across prokaryotic genomes of genes that appear as sisters to the eukaryotic nuclear copy is of interest because their phylogenetic affinities can, in principle, help to discriminate between competing theories for eukaryote origins. An overview of the results is shown in figure 2. For simplicity, the topologies can be divided into three general categories with respect to the taxonomic distribution of eukaryote sister genes among prokaryotes. Group 1: The sister genes occur only among genomes of one of the higher prokaryotic taxa shown in figure 2, for example, euryarchaeotes, thaumarchaeotes, α-proteobacteria, firmicutes, or the like. Among the 571 LECA gene families, 375 yield this result. Group 2: The sister genes are not restricted to a particular class or phylum but occur only among members of the archaebacteria or eubacteria, 165 alignments and trees yield this result. Group 3: The sister group to the eukaryotic nuclear gene includes genes that occur among members of both the archaebacteria and the eubacteria, 31 alignments and trees deliver this result.

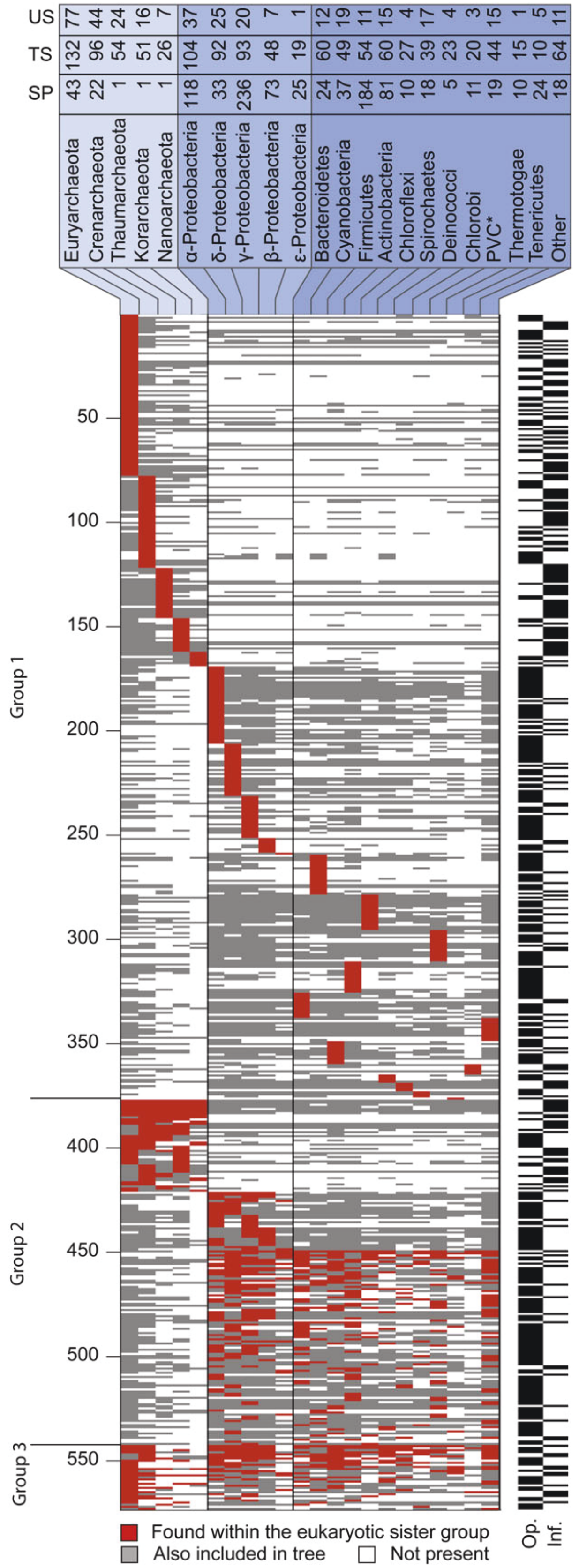
The trees in Group 3 contain the least information and are also the easiest to interpret. Monophyletic eukaryotes nested within and as the sister to clades in which the archaebacteria and the eubacteria are interleaved indicate that the eukaryotic gene reflects a single origin, but that during the time subsequent to the origin of eukaryotes, the prokaryotic gene has undergone so much sequence divergence and/or LGT among prokaryotic groups that it is not possible to generate a strong inference about the source of that gene in LECA through the vantage point of phylogenetic trees. Many phylogenetic artefacts affecting the prokaryote topology might also rest in this category, distinguishing between LGT and phylogeny or alignment artefacts is not straightforward (Roettger et al. 2009). These 31 trees thus are equivocal about gene origins in LECA.

The 165 trees in Group 2 show the eukaryote nuclear genes branching as the sister to groups containing homologues present in several different archaebacterial or several different eubacterial higher taxa. These genes tend to reflect archaebacterial or eubacterial ancestries for the eukaryotic gene, respectively, without implicating a specific higher-level taxon as the donor lineage. Among these genes, 44 reflect an archaebacterial ancestry, whereas 121 reflect a eubacterial ancestry. Of the 44 archaebacterial-derived genes in the LECA, 28 belong to the informational class (involved in information storage and processing), whereas 103 out of the 121 eubacterial-derived LECA genes belong to the operational class (involved in biochemical and biosynthetic processes). Thus, the informational and operational classes of eukaryotic genes well-established in analyses of the yeast genome (Rivera et al. 1998; Cotton and McInerney 2010) as well as the preponderance of eubacterial-derived over archaebacterial-derived genes in eukaryotic genomes (Esser et al. 2004; Pisani et al. 2007) are also evident for these 165 genes present in LECA. However, for these 165 trees, the sister group relationship to the eukaryotic gene appears more or less as a bucket of mixed pickles, but archaebacterial or eubacterial pickles. Although the proteobacteria are clearly the most frequently represented among the 121 trees indicating a eubacterial ancestry of the eukaryote nuclear genes, all eubacterial groups are ultimately represented.

The 375 trees that we classified as Group 1 show one of the higher prokaryotic taxa sampled as harboring the sister gene of the eukaryote common ancestor homologue. That is, the sister of the eukaryotic nuclear gene contained only members of one of the 21 higher prokaryotic taxa (22 including the category "other") shown in figure 2. The most frequent taxon uniquely harboring sister genes to genes present in the eukaryote common ancestor were the euryarchaeotes (77 genes), followed by the crenarchaeotes (44 genes), the α-proteobacteria (37 genes), the δ-proteobacteria (25 genes), the thaumarchaea (24 genes), the γ-proteobacteria (20 genes), the cyanobacteria (19 genes), the spirochaetes (17 genes), the korarchaeote (16 genes), the actinobacteria (15 genes), the PVC group (15 genes), the bacteriodetes (12 genes), etc. In fact, all of the higher taxa sampled harbor a gene with a sister group relationship to the eukaryotic nuclear homologue. Some might conclude from this that all prokaryote lineages sampled donated genes to LECA but that is a much too simplistic inference that entails unrealistic assumptions about the nature of prokaryotic lineages and the affects of LGT over geological timescales as outlined in the Discussion.

Functional Categories

There are only 68 archaebacteria in our genome sample, and it is evident in figure 2 that among those alignments and trees where eukaryotic genes branch with archaebacteria as sisters, the eubacteria are underrepresented. We plotted the frequency distribution of the number of Operational Taxonomic Units (OTUs) in the protein family and for the same distribution the proportion of archaebacterial sequences in each alignment, or tree, as shown in figure 3. Among trees having 68 prokaryotic taxa or fewer, there are 47 that contain only archaebacterial homologues and 61 in which over 90% of the OTUs are archaebacterial. With increasing numbers of prokaryotes in the trees, the proportion of archaebacteria declines quickly, and there is clearly a bimodal distribution with regard to the presence and frequency of archaebacterial sequences in the protein families containing fewer than, or more than, 68 prokaryotic sequences (fig. 3). Because this bimodality is not independent of the informational–operational gene dichotomy, we plotted functional category assignments against sister group relationship for the 571 genes showing



eukaryote monophyly in two bins, that is, trees containing fewer than 68 prokaryotic sequences (fig. 4*a*) or more than 68 prokaryotic sequences (fig. 4*b*).

In figure 4*a*, the archaebacterial nature of the eukaryotic genetic apparatus, ribosome biogenesis in particular, stands out. Figure 4*b* summarizes the eukaryote sisterhood frequencies for the trees with stronger eubacterial representation. In addition to the archaebacterial informational signal, the most notable feature of figure 4*b* is the frequency with which proteobacteria branch as sister to the eukaryotes in operational genes, in particular energy metabolism. We note that the frequencies in figure 4 have not been normalized with respect to the number of species or genes per category. For example, the high frequency of euryarchaeotal sisterhood observed is not completely independent of the heavier taxon sampling for euryarchaeotes, which are twice as frequent in the data (43 genomes) as crenarchaeotes (22 genomes). In the same vein, the appreciable frequency of γ-proteobacterial sisterhood in energy metabolism category or firmicute sisterhood in the category posttranslational modification and chaperones (fig. 4) is not independent of the large number of genomes sampled for these groups in our data, which is, for obvious reasons, strongly skewed toward pathogens: the 236 γ-proteobacterial and 184 firmicute genomes in our data (fig. 2, top). However, normalization is not as easy as it might seem because many of the elements on both matrices (fig. 4*a* and *b*) are empty and because the genomes within the higher taxa indicted are extremely diverse with respect to genome size and frequencies of various functional categories. We plotted sisterhood occurrence for how often a gene from the taxon was found in the sister clade normalized by the frequency of genes in

Fig. 2.—A presence–absence pattern (PAP) of the bacterial taxonomic groups in trees supporting the eukaryotic monophyly. The rows correspond to 571 trees in which the eukaryotes were monophyletic, the columns correspond to 22 bacterial groups. A cell *i*,*j* in the matrix is colored if tree *i* included a homologue from bacterial group *j*. Taxonomic groups harboring a gene that branches as a nearest neighbor to the eukaryotic clade in each tree are marked by a red cell. Taxonomic groups that are also present in the tree are marked by a gray cell. An asterisk indicates that species from the Chlamydiae, Verrumicrobia, and Plancomycetes taxa were combined into one group (PVC) (Wagner and Horn 2006). Group 1 included only trees where exactly one bacterial group was found, Group 2 included trees where 1) only Archaea were found, 2) only proteobacteria were found, and 3) were only eubacteria were found. Group 3 included all other trees. The black and white bars on the right indicate whether the KOG underlying the tree belongs to an informational (Inf.) or to an operational (Op.) class (Rivera et al. 1998). The numbers in the top panel indicate the following. US: The number of times that the given taxon was the only taxon in the sister group to the eukaryotic sequence (unique sisterhood, US). TS: the number of times that the taxon was either included in a unique sister group or in a sister group consisting of mixture of prokaryotic taxa total sisterhood (TS). SP: the number of sequenced strains from that taxon in our genome sample.

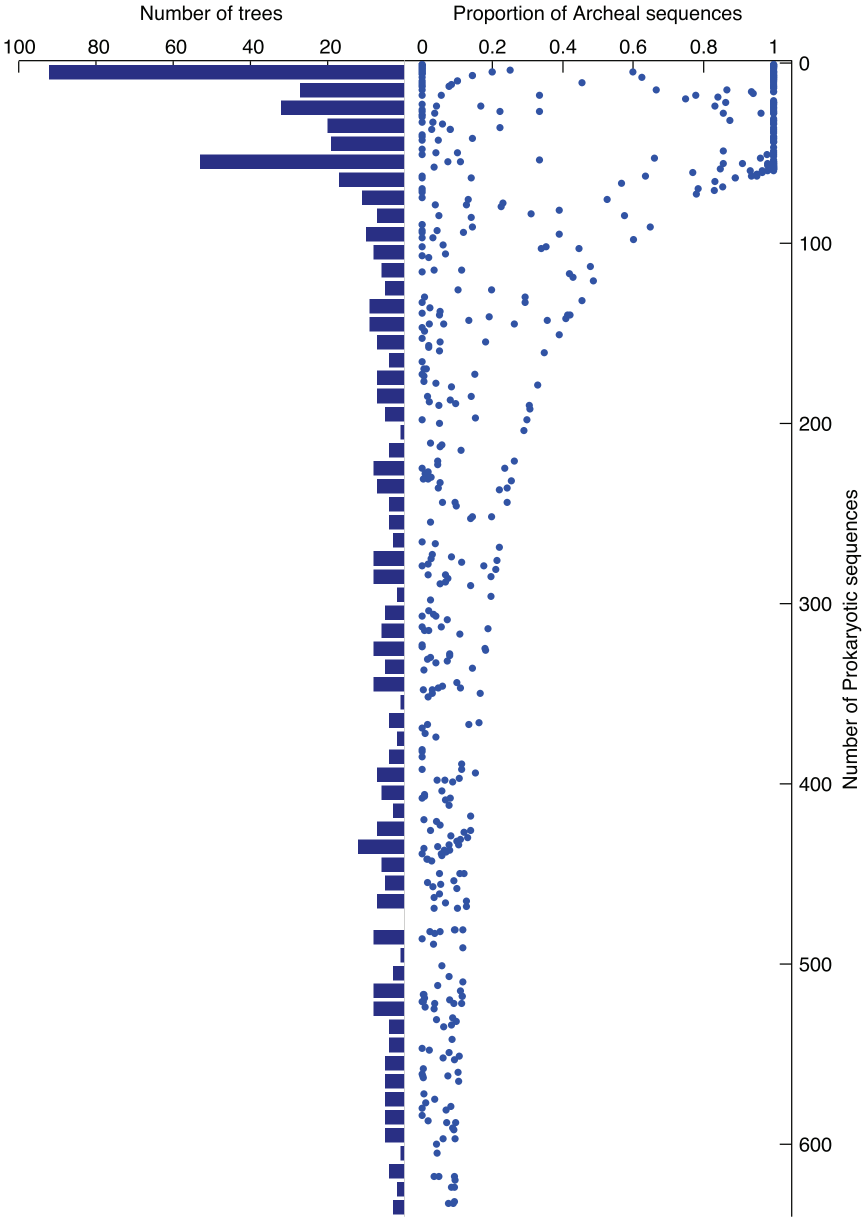


Fig. 3.—Proportion of archaeal sequences per alignment in the data set. The left bar graph shows the distribution of bacterial sequences in all trees where the eukaryotes form a monophyletic clade in bin intervals of 10. The plot on the right indicates the proportion of archaeabacterial sequences in each tree. There were only 68 archaea in the data, hence a skew distribution of trees containing many or mostly archaeabacterial sequences versus eubacterial sequences in alignments with more than 68 OTUs (see also fig. 4).

that taxon that are present in the trees, hence capable of appearing as the eukaryote sister. Although there are sufficient number of observations to normalize at the level of taxa, when normalization is extended to functional categories, spurious results are obtained, even when the empty or nearly empty elements of the matrix are removed (supplementary fig. S3, Supplementary Material online).

The apparent strong contributions of euryarchaeotes and α-proteobacteria are notable and robust. Among the archaebacteria, the crenarchaeotes are, on a per gene basis, more frequently found in the sister group than the euryarchaeotes (fig. 5*a*). Among the α-proteobacteria, there is a positive correlation ($\rho = 0.0437$, $P = 2.8 \times 10^{-6}$) between genome size and eukaryote sisterhood frequency (fig. 5*b*), indicating in the simplest interpretation, that the ancestor of mitochondria had a large genome. If we reduce the result of the functional category analysis to its most basic statement, the data reveal clear evidence for the archaebacterial nature of the eukaryotic genetic apparatus and the eubacterial nature of eukaryotic energy metabolism.

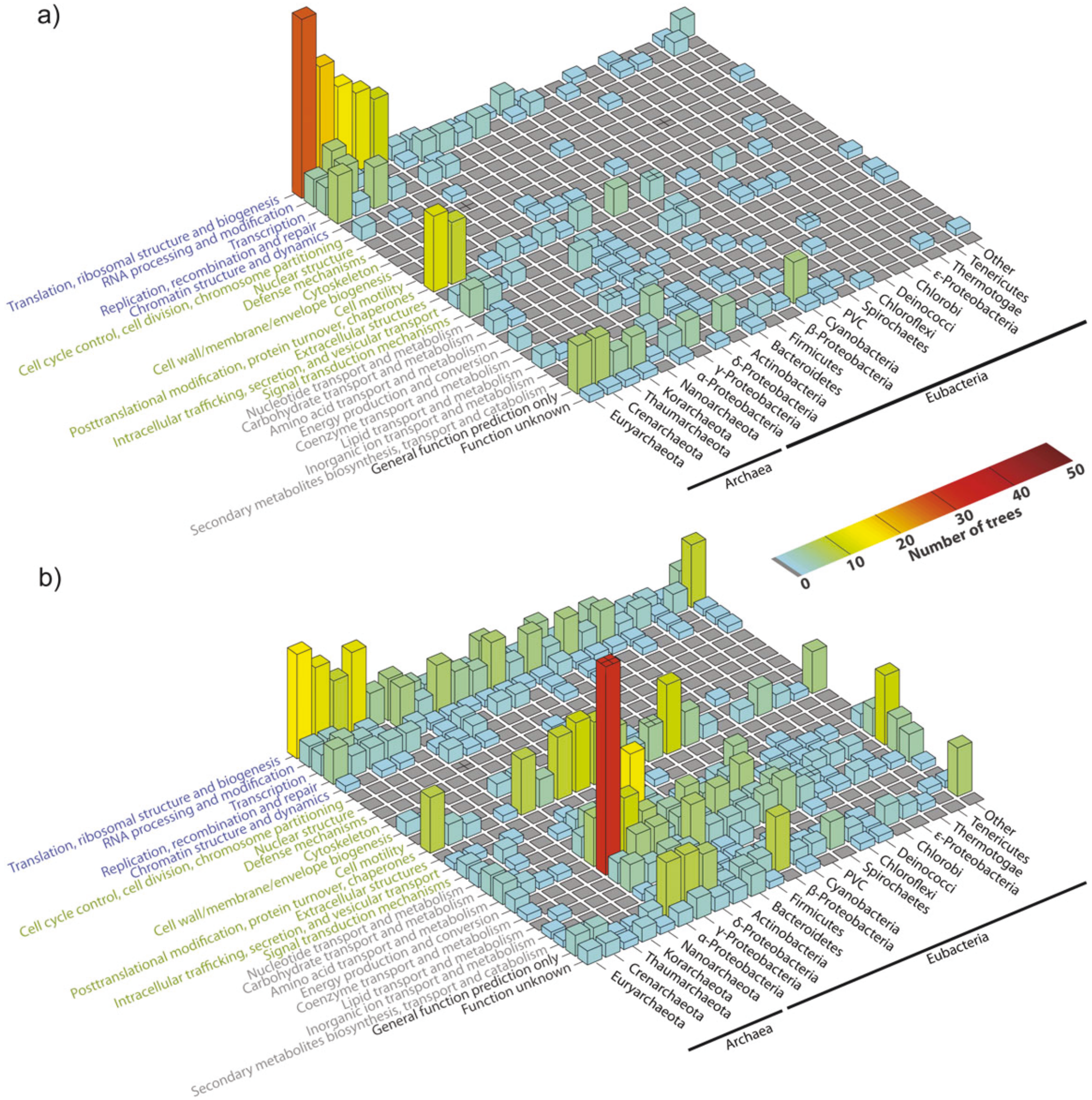


Fig. 4.—Three-dimensional bar graphs of prokaryotic groups found as sister groups to the eukaryotes distributed across functional categories according to KOG groups. The four main groups are information storage and processing (classes colored in blue), cellular processes and signaling (classes colored in green), metabolism (classes colored in gray), and poorly characterized proteins (classes colored in black). (*a*) Including the data from trees with 68 or fewer prokaryotic sequences. (*b*) Including the data from trees with more than 68 bacterial sequences. Bar height and color indicate how often a certain group was found in a tree belonging to a certain category.

Genes Dispersed (Widely) from the Ancestor of Mitochondria

Theories on the origin of eukaryotes differ with respect to the role of mitochondria therein. Some theories view the origin of mitochondria as distinct from and mechanistically irrelevant to the origin of eukaryotes (Kurland et al. 2006; Forterre and Gribaldo 2010). Others view the origin of mitochondria as coinciding with the origin of eukaryotes (Martin and Müller 1998; Embley and Martin 2006), as having precipitated the origin of the nucleus (Martin and Koonin 2006), and as an energetic *conditio sine qua non* for the origin of eukaryote-specific gene families that underpin eukaryotic cell and cell cycle complexity (Lane and Martin 2010). Most studies aiming to identify the sister group to mitochondria have focused on genes encoded in mtDNA. But mtDNA-encoded proteins are often highly divergent or rapidly evolving and phylogenetic problems thus arise with their tendency to branch with proteins from other rapidly evolving lineages such as Rickettsias (Bridefalk et al. 2011; Georgiades and Raoult 2011; Thrash et al. 2011). In phylogenetics, the problem is well-known and called long-branch attraction (Lockhart et al. 1994). The most slowly evolving prokaryotic homologues of eukaryotic nuclear-encoded proteins should be least affected by long-branch attraction,

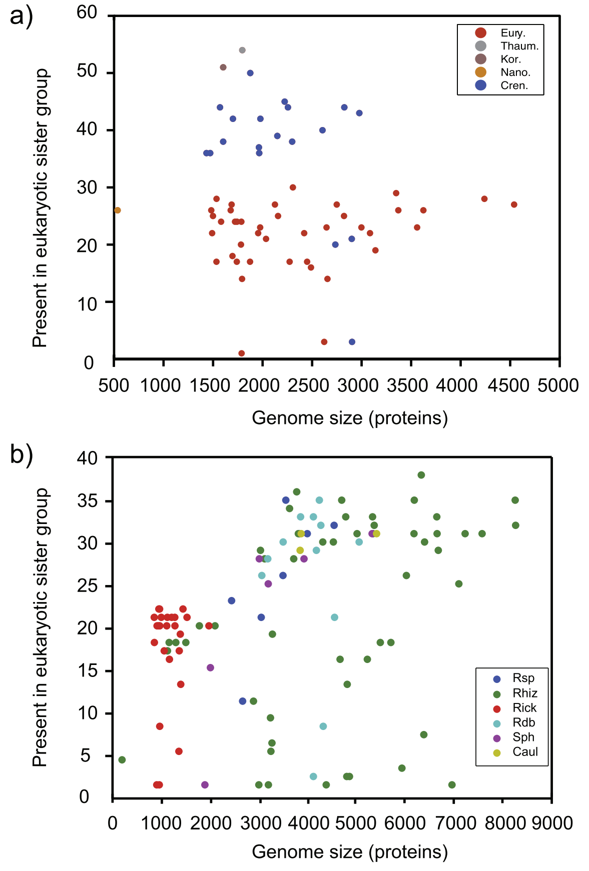


Fig. 5.—Correlation between genome size and strain presence in the eukaryotic sister clade. (*a*) All archaebacterial species that were found as eukaryotic sisters plotted against their genome size. Correlation was measured using the spearman rang correlation, resulted in $\rho = -0.1106$, $P = 0.3882$. (*b*) All α-proteobacterial species that were found as eukaryotic sisters plotted against their genome size. Correlation was measured using the Spearman rang correlation, resulted in $\rho = 0.4371$ and $P = 2.8 \times 10^{-6}$.

and nuclear genome data have not been examined from that standpoint so far. Hence we examined all 571 trees showing eukaryote monophyly to find the prokaryotic homologues that had the least total distance (the shortest path) in the ML tree to the eukaryotic nuclear genes. The result (not shown) was similar to figure 4*a* and *b* in that the highest frequencies of sister group occurrence were observed in the euryarchaeotal category for ribosome biogenessis and the α-proteobacterial category energy metabolism.

The α-proteobacterial genomes harboring the slowly evolving genes were mainly facultative anaerobes from the Rhodospirilliaceae (*Magnetospirillum* and *Rhodospirillum*) with the spectrum of functions represented being metabolic and thereby very distinct from the ribosomal proteins and respiratory chain components that are typically encoded in mitochondrial genomes (fig. 6). Thus, the result that one obtains for the inferred nature of the ancestor of mitochondria depends strongly upon which genes one considers: The fast-evolving genes in mtDNA often point to a fast-evolving mitochondrial ancestor related to Rickettsias (Bridefalk et al. 2011; Thrash et al. 2011; but see also Esser et al. 2004 and Abhishek et al. 2011 for different results), whereas the proteins encoded in nuclear DNA point to facultative anaerobic generalist α-proteobacteria as the mitochondrial ancestor (Atteia et al. 2009)—the most slowly evolving proteins in particular—as seen in figure 6.

There are 106 α-proteobacteria in our sample, about one-fourth of which are intracellular pathogens belonging to the Rickettsiales. Figure 7 plots the frequency of proteins from 106 α-proteobacterial genomes appearing in the sister group to the eukaryotic genes (dark blue fields), how often each genome harbors a protein that does not branch as the eukaryote sister (light blue fields), or whether the gene is missing in the genome altogether (white fields). The α-proteobacterial strains with the highest frequency of occurrence in the sister group were *Rhizobium* NGR 234 (Rhizobiales, 38 times), *Beijerinckia indica* ATCC 9039 (Rhizobiales, 36 times), *Acidiphilium cryptum* JF-5 (Rhodospirillales, 35 times), *Ruegeria pomeroyi* DSS 3 (Rhodobacterales, 35 times), *Sinorhizobium meliloti* (Rhizobiales, 35 times), *Azorhizobium caulinodans* ORS 571 (Rhizobiales, 35 times), and *Methylobacterium nodulans* ORS 2060 (Rhizobiales, 35 times; for the complete list see supplementary Table S3, Supplementary Material online). Clearly, among those genes where an α-proteobacterial homologue resides in the eukaryote sister group, different genes implicate different ancestors of mitochondria within the α-proteobacteria, and each of the genomes is implicated as the sister of a eukaryotic nuclear gene at least once. It has been suggested that such patterns could reflect multiple origins of mitochondria (Georgiades and Raoult 2011). It is more likely however, in our view, that such patterns reflect a single origin of mitochondria followed by subsequent LGT among free-living prokaryotes (Martin 1999b; Richards and Archibald 2011).

A Network Linking LECA to Prokaryotes

Based on the current sample of 994 genomes, 571 trees implicate many different prokaryotes as gene donors to the eukaryote common ancestor (fig. 7). In fact, the trees implicate all 22 prokaryote higher taxa sampled here are gene donors to LECA (figs. 2 and 4). Figure 8 summarizes those results in a network in which the weight of the edges connecting prokaryotes to LECA reflects the relative frequency of gene contribution to LECA by the respective lineage in the current sample. The eubacterial contributions are shaded blue, the archaebacterial contributions are shaded red, and the retention of genes from both sources in diversifying eukaryotic lineages is indicated accordingly. As in earlier studies (Esser et al. 2004; Rivera and Lake 2004; Dagan and Martin 2006; Pisani et al. 2007), the eubacterial

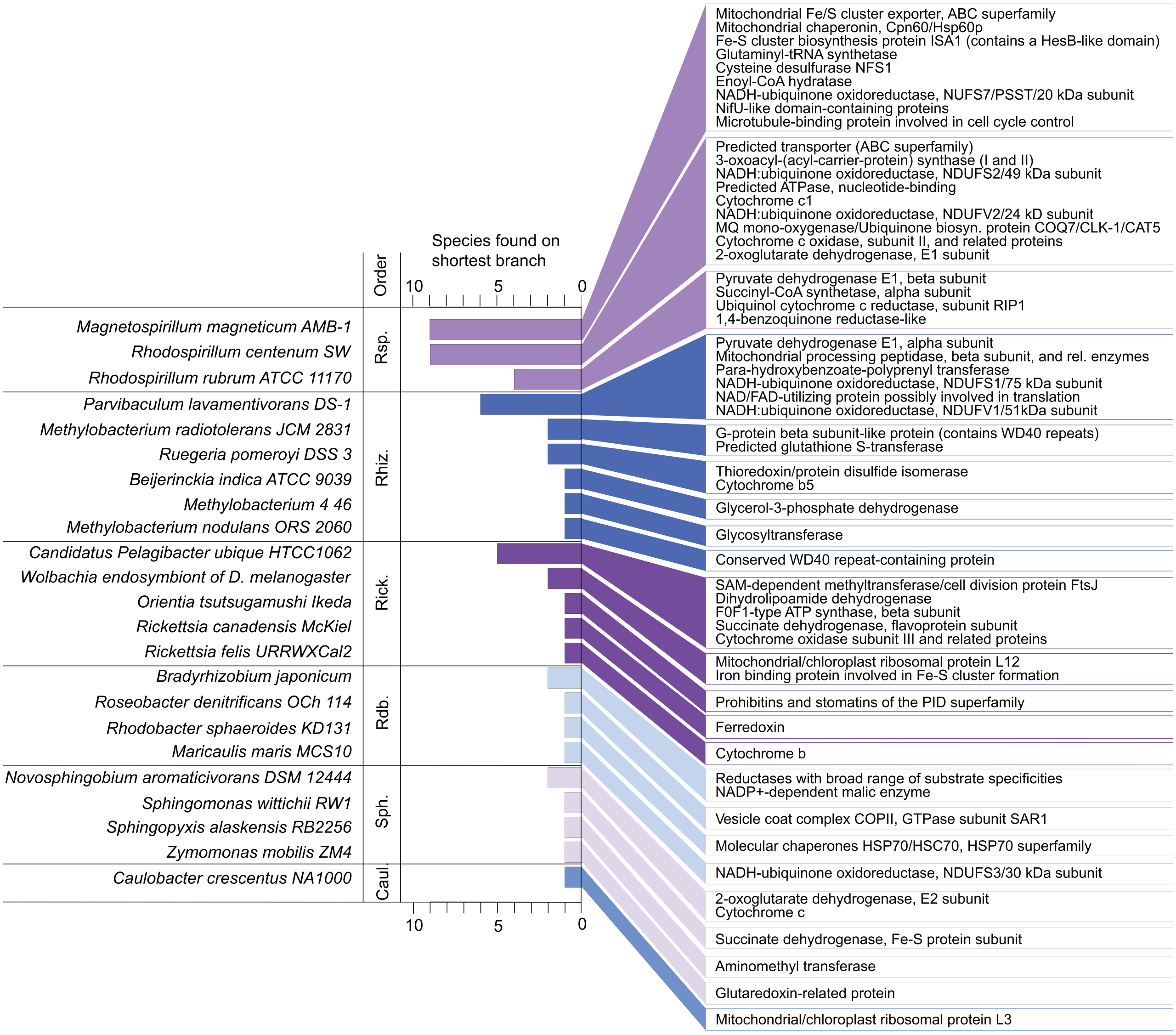


Fig. 6.—Frequency of single α-proteobacterial species found on the shortest path to the eukaryotic clade, by summing up the branch lengths. For each of the respective sequences, the functional annotation is also given. Abbreviations refer to α-proteobacterial families. Rsp: Rhodospirillales, Rhiz: Rhizobiales, Rick: Rickettsiales, Rdb: Rhodobacterales, Sph: Sphingomonadales, and Caul: Caulobacterales.

contribution to eukaryotic genomes is quantitatively predominant, a crucial circumstance that is still too often overlooked (Gribaldo et al. 2010). These distinct contributions from archaebacteria and eubacteria require a network-based framework, rather than a tree-based framework, for addressing eukaryote origins. Under the simplest working hypothesis, the eubacterial and the archaebacterial contributions stem from only one cellular donor each, the eubacterial ancestor of mitochondria and its archaebacterial host, respectively. The (erroneous, in our view) implication of several different donor lineages stems merely from the natural workings of LGTamong free-living prokaryotes subsequent to the origin of eukaryotes, as sketched in figure 9.

In order for that explanation to be tenable, a considerable amount of LGT must have occurred for the genes under study among the ancestors of the groups sampled here. To see if there is evidence for that, we looked to see how often the prokaryotic groups in question were monophyletic across the 571 trees for which the eukaryotes were monophyletic. The result is shown in Table 1. The worst ‘‘LGT offenders’’ were the δ-proteobacteria, the firmicutes, and the γ-proteobacteria, which each were monophyletic groups in less than 10% of the trees studied. Aside from archaebacteria and α-proteobacteria, these three groups were also the largest apparent contributors to the functional classes in figure 4b.

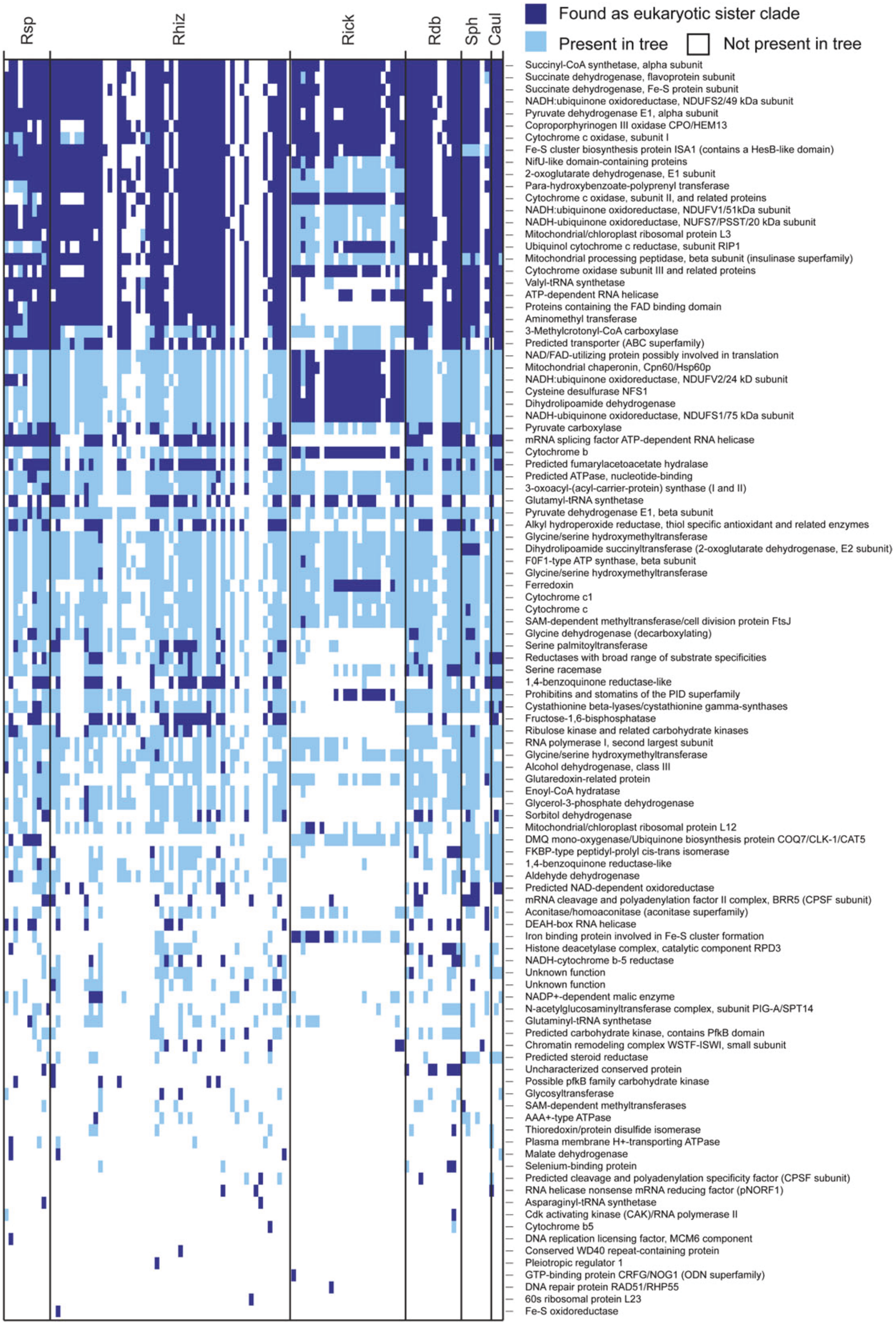


Fig. 7.—Distribution of α-proteobacterial groups found as sister group to the eukaryotic clade. These presence absence matrix gives the functional description of all 104 trees (as rows) where the α-proteobacteria (106 different species, as columns) were found as the eukaryotic sister clade. The color indicates whether a group was found as sister clade (deep blue) or was just present in the tree (light blue). Abbreviations of α-proteobacterial families as given in the legend to figure 6.

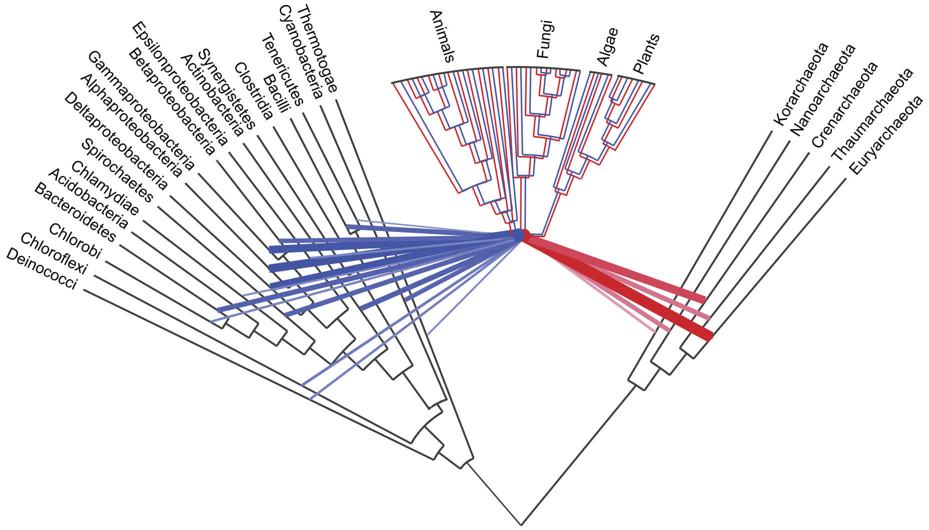


Fig. 8.—Network linking apparent prokaryotic donors to the eukaryote common ancestor according to the present findings. This network based on a traditional phylogenetic tree to which lateral edges were added. Color intensity and width of the lateral edges reflect the frequencies with which these groups appear as sisters to the eukaryotic clade.

Discussion

Single Ancient Acquisition, Not Continuous Influx

We identified and investigated 712 eukaryotic protein families that have prokaryotic homologues. Of those trees, 571 (80%) reflect a single origin for the eukaryotic gene. For the remaining 141 genes whose trees do not directly reflect a single origin, there are causes other than LGT from prokaryotes to eukaryotes that can readily account for the lack of observed eukaryote monophyly, the two most obvious of which are computational (phylogenetic parameters) and biological (differential loss). Examining alignment characteristics that are correlated with the inference of LGT from discordant branching, Roettger et al. (2009) found that the number of OTUs in the alignment (tree) was among the most highly correlated with LGT inference: the more OTUs in the alignment, the more likely the inference of LGT. The median number (± standard error) of OTUs in our 141 trees that did not recover eukaryote monophyly is 337 ± 179, which is significantly higher ($P < 10^{-11}$) than the median value of 115 ± 193 for the 571 trees in which the eukaryotes comprised a monophyletic group. The large number of OTUs is potentially a biased source of alignment and phylogeny artefacts that could disrupt eukaryote monophyly.

Another mechanism that could readily produce the 141 cases of eukaryote non-monophyly is differential loss among paralogous gene families that were inherited as paralogs from the mitochondrial ancestor in LECA. About 50% of the genes in an average contemporary prokaryotic genome have duplicates within the genome (Hooper and Berg 2003). For our present considerations, it is immaterial whether the source of the duplicated prokaryotic gene is from within the chromosome or via lateral acquisition, although genome data argue in favor of the latter (Treangen and Rocha 2011). If the mitochondrial ancestor had a typical genome with about 4,000 genes, it would have then harbored about 2,000 genes existing within paralogous families, like *Escherichia coli* or *Bacillus subtilis* do (D'Antonio and Ciccarelli 2011). Transfer from a single source, for example the host or the mitochondrial ancestor, followed by differential loss within such gene families during eukaryote evolution (Zmasek and Godzik 2011), would produce non-monophyletic eukaryote trees in a manner that does not involve LGT.

At the same time, there are limits as to how many such patterns can be explained with differential loss only. If differential loss (instead of LGT) is invoked to explain the presence/absence patterns of all nonuniversally distributed genes, the Genome of Eden problem (Doolittle et al. 2003) ensues: inferred ancestral genome sizes become orders of magnitude larger than any observed contemporary prokaryotic genome, an untenable proposition (Dagan and Martin 2007). But for a mere 140 genes families the situation is less severe, especially given that the genome of the mitochondrial ancestor probably harbored on the order ~1,000 gene families. Furthermore, at least eight ancient gene duplicate pairs (16 gene families) have been universally, or nearly so, conserved across prokaryotic genomes (Dagan et al. 2010). Thus, it would be unreasonable to assume that there was neither paralogy nor differential loss in the 20% fraction of trees where eukaryotes appeared non-monophyletic, especially given that 50% of a typical prokaryotic genome falls into intragenomic gene families.

Given eukaryote age, there have been ~1.8 billion years (Parfrey et al. 2011) of opportunity for these eukaryote lineages to reacquire these 571 genes from prokarytes via LGT. But that has not happened, indicating that LGT from

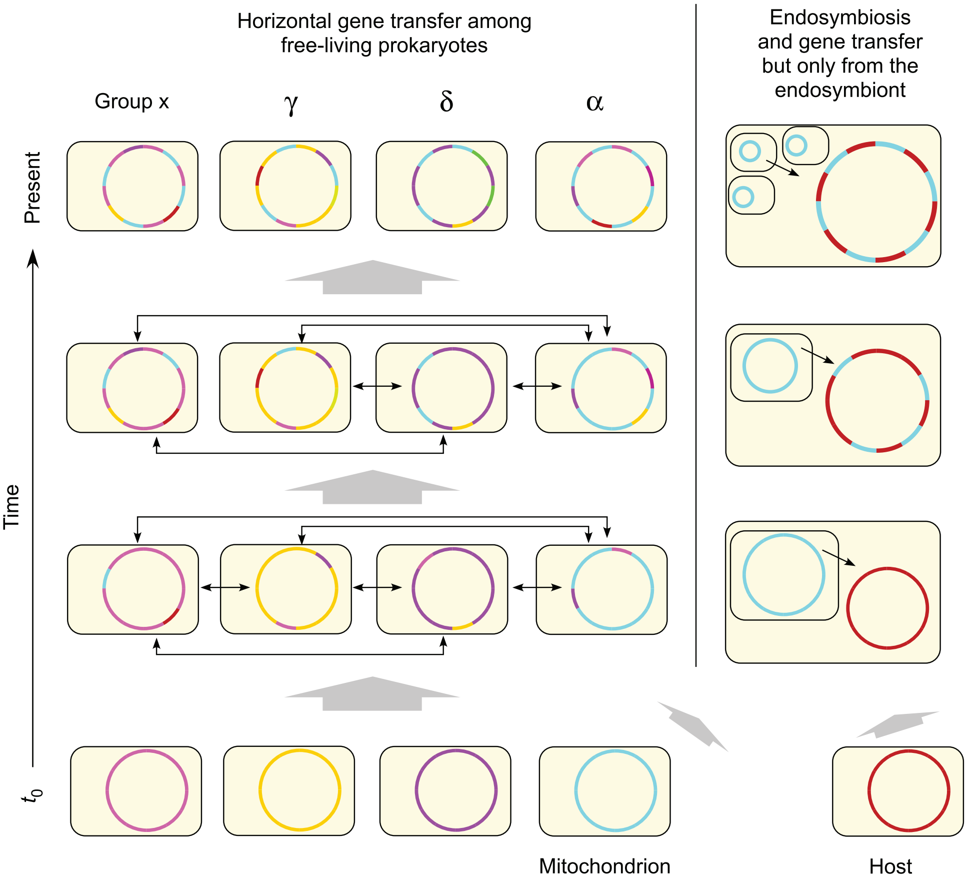

Horizontal gene transfer among free-living prokaryotes
Endosymbiosis and gene transfer but only from the endosymbiont
Group x
γ
δ
α
Present
Time
t_0
Mitochondrion
Host

Fig. 9.—Lateral gene transfer between free-living prokaryotes subsequent to the origin of organelles requires that we think at least twice when interpreting phylogenetic trees for genes that were acquired from mitochondria (or chloroplasts, not shown). Genes that entered the eukaryotic lineage via the genome of the mitochondrial endosymbiont represent a genome-sized sample of prokaryotic gene diversity that existed at the time that mitochondria arose. The uniformly colored chromosomes at t_0 indicate that at the time of mitochondrial origin, there existed for individual prokaryotes specific collections of genes in genomes, much like we see for strains of *Escherichia coli* today. If an *E. coli* cell would become an endosymbiont today, it would not introduce an *E. coli* pangenome's worth of gene diversity (some 18,000 genes) into its host lineage, rather it would introduce some 4,500 genes or so. The free-living relatives of that endosymbiont would go on reassorting genes across chromosomes via gene transfer at the pangenome (species and strain) level, at the genus level, at the family level, and at the level of proteobacteria, the environment, and so forth. After 1.5 billion years, it would be very unreasonable to expect any contemporary prokaryote to harbor exactly the same collection of genes as the original endosymbiont did. Instead, the descendant genes of the endosymbiont (labeled blue in the figure) would be dispersed about myriad chromosomes, and we would eventually find them one at a time through genome sequencing of individuals from different groups. Though not shown here, for reasons of space limitation, exactly the same process also applies, in principle, for the host's genome. Redrawn from Martin (1999b) and from figure 5 of Rujan and Martin (2001).

prokaryotes to eukaryotes—outside the context of endosymbiotic organelle origins—is rare in eukaryote evolution. It is certainly far more rare than LGT among prokaryotes in evolution. This is consistent with the lack of functional gene acquisition by aphids from *Buchnera* endosymbionts (Nikoh et al. 2010), despite more than 100 million years of intracellular coevolution. By contrast, at the origins of chloroplasts and mitochondria, gene transfers from the genomes of the respective endosymbionts and functional integration of those genes into the metabolism of the resulting cell were abundant (Timmis et al. 2004; Lane and Archibald 2008).

Much current thinking on eukaryote origins is still focused on debating the branching orders in alternative trees (Gribaldo et al. 2010): a tree *x* versus tree *y* debate. But a spectrum of alternatives that consider only trees is not broad enough. A considerable amount of evidence indicates that the process of eukaryote origins was not tree-like to begin with. Eukaryote genome evolution entails many non–tree-like processes, and these non–tree-like events (endosymbiosis and gene transfer) could be the decisive events in eukaryote evolution (Lane and Martin 2010; Koonin 2012). Gene origin and evolution in the eukaryotic tree of life has many tree-like components (Bapteste et al. 2009). But when the overall process of eukaryote (genome) evolution is set in the context of a realistic model of prokaryotic genome evolution, with abundant gene transfer among

Table 1

Prokaryote Monophyly in Eukaryote Monophyly Trees

	Degree of Prokaryote Monophyly		
Group	Strict[a]	Outer[b]	Inner[c]
Chlamydiae	0.844	0.856	0.962
Chlorobi	0.672	0.695	0.893
Deinococcus	0.654	0.693	0.882
Thermotogae	0.579	0.635	0.851
ε-Proteobacteria	0.534	0.583	0.785
Cyanobacteria	0.473	0.557	0.760
Crenarchaeota	0.341	0.598	0.660
Chloroflexi	0.286	0.364	0.665
Spirochaetes	0.249	0.298	0.641
β-Proteobacteria	0.237	0.415	0.501
Bacteroidetes	0.214	0.334	0.642
Euryarchaeota	0.200	0.476	0.505
α-Proteobacteria	0.194	0.398	0.499
Actinobacteria	0.170	0.359	0.534
Archaea, other[d]	0.092	0.159	0.486
γ-Proteobacteria	0.090	0.382	0.325
Firmicutes	0.080	0.310	0.345
δ-Proteobacteria	0.056	0.152	0.361
Bacteria, other	0.038	0.099	0.292

[a] The proportion of trees in which the group is monophyletic.

[b] The proportion of the members of the given group that are present in the tree and contained within the smallest clade containing all members of the group (and members of other groups); n_{group}/n_{clade}, where n_{group} is the number of members of the group in the clade and n_{clade} is the number of OTUs in that clade. Value shown is the mean across all trees.

[c] The proportion of the members of the given group that are present in the tree and contained within the group's largest monophyletic clade; $n_{group\{clade\}}/n_{group\{tree\}}$, where $n_{group\{clade\}}$ is the number of members of the group in the clade and $n_{group\{tree\}}$ is the number of group members in the tree. Value shown is the mean across all trees.

[d] Designates a grouping of Nanoarchaea, Thaumarchaea, and Korarchaeota lumped together, the individual samples of which are either one or too small to consider monophyly.

prokaryotes and occasional major influxes into eukaryote genomes via endosymbiosis and gene transfers from organelles, the non-treelike evolutionary events (Lane and Archibald 2008; McInerney et al. 2008) stand out—they are components of eukaryote genome evolution that do not fit on a tree. They require network approaches.

Many Theories and Many Trees

Among the 571 trees that recovered eukaryote monophyly, the higher prokaryotic taxa harboring a gene with a sister group relationship to the eukaryotic nuclear homologue are shown in figure 2. These genes could provide evidence to discriminate between different current theories for eukaryote origin. We start with theories that received the least support.

One theory has it that the eukaryotic lineage is of equal age as the two prokaryotic domains (Kurland et al. 2006); it predicts that we should mainly obtain a topology of three monophyletic domains among our trees. The three-domain tree was however observed in only three cases out of 571 (0.5% of all trees): the 60s ribosomal protein L2/L8 (KOG2309), the 40S ribosomal protein S16 (KOG1753), and the large subunit of RNA polymerase III (KOG0261). Gribaldo et al. (2010) argued that the three-domain tree is correct, but reported no new analyses to test that view and considered only a specific subset of genes—those that specifically link eukaryotes and archaebacteria and hence fit the metaphor of a tree. In doing so, they disregared the eubacterial majority of genes in eukaryotic genomes. Among theories considered here, the three-domain tree received the least support. The next lowest rung on the ladder of support is occupied by the theory that the eukaryotic nucleus arose within an endospore forming Gram-positive bacterium (Gould and Dring 1979), in which case eukaryotes should branch with firmicutes, which was observed in only 11 trees (fig. 2).

That is followed by theories that entail the planctomycetes (the PVC group) as intermediate steps in the prokaryote-to-eukaryote transition (Devos and Reynaud 2010) or as the host for an endosymbiotic origin of the nucleus (Forterre 2011). PVC sisterhood is observed in 15 trees (2.6%). That is the same level of support as current versions of the neomuran theory receive (Cavalier-Smith 2002) because actinobacteria branched as eukaryote sisters in 15 trees. The theory of the late Lynn Margulis that spirochaetes were crucial to eukaryote origin via the simultaneous origin of eukaryotic flagella and the nucleus (Margulis et al. 2006) fared incrementally better, with 17 trees pegging spirochaetes as eukaryotic sisters (fig. 2). Better still fared the original version of the neomuran theory (Cavalier-Smith 1975) with eukaryotes viewed as direct descendants of cyanobacteria (19 trees).

Some theories have it that the nucleus arose as an archaebacterial endosymbiont in a Gram-negative host (Gupta and Golding 1996), in some formulations a γ-proteobacterial host (Horiike et al. 2004). The corresponding γ-proteobacterial sisterhood is observed in 20 trees. Related theories have it that the host for an endosymbiotic origin of the nucleus was a δ-proteobacterium (Moreira and Lopez-Garcia 1998), a topology that is observed for 25 genes (fig. 2). Although an endosymbiotic theory for the origin of the nucleus belongs to the very first formulations of endosymbiotic theory (Mereschkowsky 1905, 1910), there are a number of fundamental and serious problems with the view that the nucleus was ever free-living prokaryote (Martin 1999a; Cavalier-Smith 2002).

Several modern formulations of endosymbiotic theory that posit only two cells at eukaryote origin, an archaebacterial host and a mitochondrial endosymbiont (Searcy 1992; Martin and Müller 1998; Vellai et al. 1998). One formulation of endosymbiotic theory entailing a prokaryotic host posits mass transfer of genes from the genome of the mitochondrial endosymbiont to the chromosomes of the host, while directly accounting for the common ancestry of mitochondria and hydrogenosomes (Martin and Müller 1998), and an autogenous origin of the nucleus in a mitochondrion-bearing cell, mechanistically precipitated via the invasion of Group II introns from the symbiont into the host's chromosomes and their transition there to spliceosomal introns (Martin and Koonin 2006). This is supported by the most frequent class of eubacterial sisterhood observed was for α-proteobacteria (37 genes; fig. 2).

The archaebacterial genomes sampled revealed some cases of eukaryotic sisterhood for Nanoarchaea (Huber et al. 2002) and Korarchaeota (Elkins et al. 2008), which have so far not been implicated in eukaryote origins, as well as more frequent sisterhood for mesophilic crenarchaeotes currently called thaumarchaeotes (24 trees), crenarchaeotes (44 trees), and euryarchaeotes (77 trees), which have (Embley and Martin 2006; Cox et al. 2008; Kelly et al. 2011). Because of imbalanced lineage sampling, the data do not speak unambiguously in favor on any particular theory. Nevertheless, the distribution of signals in figure 4 is more in line with the prediction of an "archaebacterial nature of the eukaryotic genetic apparatus and a eubacterial ancestry of eukaryotic energy metabolism" (Martin and Müller 1998) than with the predictions of other theories.

Some might take the sisterhood frequency of δ-proteobacterial genes to eukaryotic homologues as evidence in favor of a participation of δ-proteobacteria at eukaryote origins, but the same logic would then have to be applied to γ-proteobacterial genes, actinobacterial genes, cyanobacterial genes, spirochaete genes, and so forth; the various theories for the origin of eukaryotes that generate those predictions cannot all be simultaneously correct. The simplest interpretation in our view is that shown in figure 9. Particularly with regard to δ-proteobacterial sisterhood, we point out that in gene sharing networks of proteobacteria, the frequency of lateral gene sharing between δ-proteobacteria and α-proteobacteria is higher than within α-proteobacteria themselves (Kloesges et al. 2011), such that gene transfer among prokaryotes prior to, and subsequent to, the origin of mitochondria could readily account for the observation. In that sense, the mitochondrion remains a plausible alternative as the sole biological source of oddly branching eubacterial genes in the genome of the eukaryotic ancestor, one requiring little in the way of corollary assumptions—all we have to assume is the gene transfer among prokaryotes has always been more or less like it is today, and we do not have to assume additional cellular partners whose ribosomes disappear. Eukaryotes (that lack plastids) possess only two kinds of ribosomes: archaebacetrial ribosomes in the cytosol and eubacterial ribosomes in the mitochondrion. Theories that posit cellular partners other than the mitochondrion and its host have to account for the disappearance of the additional genomes and ribsomes, and why the data should tend to support one partner (a δ-proteobacterium for example) over another (a γ-proteobacterium or spirochaete) even though the competing alternative signals are more or less equally strong.

Too Many Inferred Cells and Donor Lineages

In the literature on endosymbiosis and gene transfer from organelles to the nucleus, it is commonplace to speak about "eukaryotic genes of α-proteobacterial origin." But the taxonomic or lineage designation "α-proteobacterial" is in fact very problematic, and perhaps even more arbitrary that problematic. In the context of eukaryote gene origins, most readers will associate "α-proteobacterial" with "mitochondrial," and attribution of a gene origin to a cellular partner is uncontroversial; the existence of a donor cell is inferred from an observation in a phylogenetic tree. The implicit reasoning is: per donor lineage identified, add one cellular partner. That is seemingly unproblematic for α-proteobacteria (or cyanobacteria for plastids), but by that measure, we would infer 22 different prokaryotic cells (including a cell from the phylum "other") at the origin of eukaryotes on the basis of the present data. The participation of 22 different cells to construct LECA and then no subsequent additions for the next ~1.8 billion years for the genes and lineages sampled here are not likely in our view, though not prohibitively complex as an idea, if we think openly. But the more subtle problem lies elsewhere: the arbitrary level at which we define a lineage from which the cell is to be inferred.

Namely, if we alter the level of taxonomic specificity with which we describe in figure 2, we will infer fewer or more numerous cells participating at eukaryote origin as endosymbionts and hosts. For example, if we were to increase the taxonomic resolution for our designation/definition of a "donor lineage" to the level of prokaryotic families, then we would infer 148 different donor lineages (i.e., how many families there are in our sample whose genes populate eukaryote sister groups in our trees) and hence 148 different prokaryotic cells in symbiotic association at eukaryote origin, given the present sample. Or we can take things one step further: at the level of genera and species, the numbers increase to 349 cells and 768 cells participating in the origin of the eukaryote common ancestor, respectively. And as the sample of sequenced genomes grows over time, so will the number of inferred donors to LECA.

Thus, that avenue of interpretation (one cell per lineage) is clearly problematic and leads to chaos because of the arbitrariness of choosing or defining the taxonomic level at which to seek or find a donor lineage. One solution is to simply zoom out in terms of taxonomic resolution, and conceptually operate at the level of domains, in which case we would conveniently have one eubacterium (the mitochondrial endosymbiont) and one archaebacterium (the host) implicated at eukaryote origins. That would solve the "one cell inferred per lineage identified" conundrum, but it is only half of the problem. The other half concerns the concept of a prokaryotic "lineage" in the context of the amount of geological time (about 1.8 billion years) that the fossil record implores us to keep in mind when considering eukaryote origins.

In terms of how genes behave in chromosomes over time, there are two ways to think about prokaryote lineages: they have static chromosomes that are immune to LGT and differences in gene content across members of a lineage are generated only by gene loss or they have fluid chromosomes with genes coming into exiting genomes of members of the "lineage" over time. The latter fluid chromosome view has recently been presented in more generalized form as the public goods hypothesis for prokaryotic genes (McInerney, Pisani, et al. 2011). Readers familiar with prokaryote chromosome evolution will immediately complain that the static chromosome model is unrealistic and outdated, but it is very real and manifest—but usually implicit—in literature concerning eukaryote gene origins. Clearly, what we consider to be a "donor lineage" at eukaryote origin depends on the level of taxonomic resolution chosen to represent a "lineage" in our prokaryotic survey.

For example, if we go to the level of species or strains, and adhere to the concept of a static prokaryote chromosome model (that is if we neglected LGT among prokaryotes over geological time, which we do not), we would conclude that the most potent donor of genes to the common ancestor of the eukaryotic lineage was the creanarchaeon (thaumarchaeote) *Nitrosopumilus maritimus* strain SCM1, which appears in the clade adjacent to the eukaryote gene 54 times, 3 time more than the next most potent apparent donor, the korarchaeon *Candidatus korarchaeum cryptofilum* strain OPF8. But the origin of eukaryotes occurred some ~1.8 billion years ago (Parfrey et al. 2011). If we were assuming a static chromosome model and thus claiming (which we are not) that specific strains of prokaryotes served as donors of eukaryotic genes ~1.8 billion years ago, then we would be assuming (which we do not) that *Nitrosopumilus maritimus* SCM1, defined by the specific collection of 1,795 genes in its genome, existed ~1.8 billion years ago. Implicitly, we would then also be assuming (which we do not) that all the other species and strains in the present study also existed in their modern form ~1.8 billion years ago.

The cogent reader would immediately, and rightly, protest that no contemporary prokaryotic strain with its current specific collection of genes could have existed ~1.8 billion years ago. This is all the more evident in light of gene content differences among *E. coli* strains, a well-studied case. In 61 sequenced *E. coli* genomes, there are about 18,000 genes, only about 4,500 of which occur in any individual strain (Lukjancenko et al. 2010). Each new sequenced strain uncovers a new combination of genes present in the *E. coli* pan-genome and about 200 ORFs new to the species pan-genome. The *E. coli* core genome (present in all 61 sequenced strains) is currently ≤1,000 genes, and no *E. coli* genome harbors more than about 25% of all 18,000 genes found in the species (Lukjancenko et al. 2010). If we move up the taxonomic scale to the level of γ-proteobacteria, the problem gets worse: A sample of 157 γ–proteobacteria (subclass) was found to harbor 40,327 different gene families (Kloesges et al. 2011), and a sample of 329 proteobacteria (class or phylum) was found to harbor about 75,000 different genes (Kloesges et al. 2011). An average individual proteobacterium only has about 3,000–4,000 genes and none has more than 9,703, the number found in *Sorangium cellulosum* So ce 56. The difference between ~75,000 genes present in proteobacteria and ~3,000 present in a given strain is not attributable to differential loss from a proteobacterial ancestor that had 75,000 genes, but it is readily attributable to gene flow (LGT) among individuals and individual strains of proteobacteria and among proteobacteria with other prokaryotes (Doolittle 1999; Doolittle et al. 2003; Dagan et al. 2008).

Because of the foregoing considerations, reasoning along the lines of "one cell inferred per donor lineage identified" is misleading, mainly because if we are talking about genes in genomes (which we are), the word "lineage" does not have a well-specified meaning in a context where we try to relate collections of genes present in individual prokaryote genomes that existed ~1.8 billion years ago to genes present in individual modern sequenced genomes. Nonetheless, there is a current trend in the literature of inferring cells where a few genes give unexpected trees, the invisible "chlamydial" sidekick at plastid origins being perhaps the most prominent example (Huang and Gogarten 2007; Becker et al. 2008; Moustafa et al. 2008). But that line of argumentation will necessarily subside at some point, for two reasons. First, it will not lead to any form of convergence as more genomes become sampled. On the contrary, with increased sampling of prokaryotic genomes for reference comparisons, the phylogenetic "identity" of donor lineages to the eukaryotic common ancestor will continue to change and will furthermore continue to spread out across more prokaryotic genomes. Second, and more severely, if one looks for a few genes that suggest a chamydial partner one will find them, but if one looks for a few genes that suggest a spirochaete partner, one will find them too, and a clostridial partner to boot, and so forth. Because eukaryotes have so many genes, if we look for a particular phylogenetic pattern, the chances are that we will find it at a low frequency in eukaryote genomes by chance alone (Stiller 2011), but we need to look at all the genes in eukaryote genomes, not just ad hoc gene samples that support a particular story. In phylogenomics, we need to keep random phylogenetic error in mind (Stiller 2011), and the interpretations of the data need to encompass gene transfer among prokarytes because it is a very real component of microbial evolution.

All things being equal, if our present considerations are approximately on target, as sampling improves the end result might tend to asymptotically approach one apparent prokaryotic donor chromosome per eukaryotic gene, even if—as we maintain—only two cells, a mitochondrial endosymbiont and its archaebacterial host, each with a discrete and specific collection of genes, participated at eukaryote origin. As sampling improves, a more realistic, temporally dynamic prokaryote lineage concept with fluid genomes and genes as public goods will figure more widely into thinking on eukaryote gene origins. Because genes that contributed to the eukaryote common ancestor lineage have different individual histories, networks, rather that trees alone, are integral to the study of eukaryote origins, and the explanatory context needs to recognize the importance of endosymbiosis and gene transfer in evolution.

Supplementary Material

Supplementary tables and figures are available at Genome Biology Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank the European Research Council for financial support (NETWORKORIGINS grant number 232975) and James O. McInerney for helpful comments on the text.

Literature Cited

Abhishek A, Bavishi A, Choudhary M. 2011. Bacterial genome chimaerism and the origin of mitochondria. Can J Microbiol. 57:49–61.

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

Atteia A, et al. 2006. Pyruvate formate-lyase and a novel route of eukaryotic ATP-synthesis in anaerobic *Chlamydomonas* mitochondria. J Biol Chem. 281:9909–9918.

Atteia A, et al. 2009. A proteomic survey of *Chlamydomonas reinhardtii* mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the α-proteobacterial mitochondrial ancestor. Mol Biol Evol. 29:1533–1548.

Bapteste E, et al. 2009. Prokaryotic evolution and the tree of life are two different things. Biol Direct. 4:34.

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

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

Brinkmann H, Martin W. 1996. Higher-plant chloroplast and cytosolic 3-phosphoglycerate kinases: a case of endosymbiotic gene replacement. Plant Mol Biol. 30:65–75.

Brown JR, Doolittle WF. 1997. Archaea and the prokaryote-to-eukaryote transition. Microbiol Mol Biol Rev. 61:456–502.

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

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

Chapman JA, et al. 2010. The dynamic genome of *Hydra*. Nature 464:592–596.

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

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

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

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

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

Dagan T, Roettger M, Bryant D, Martin W. 2010. Genome networks root the tree of life between prokaryotic domains. Genome Biol Evol. 2:379–392.

D'Antonio M, Ciccarelli DF. 2011. Modification of gene duplicability during the evolution of protein interaction network. PloS Comp Biol. 7:e1002029.

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

Devos DP, Reynaud EG. 2010. Evolution. Intermediate steps. Science 330:1187–1188.

Doolittle WF. 1978. Genes in pieces: were they ever together? Nature 272:581–582.

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

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the tree of life hypothesis. Proc Natl Acad Sci U S A. 104:2043–2049.

Doolittle WF, et al. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philos Trans R Soc Lond B Biol Sci. 358:39–58.

Elkins JG, et al. 2008. A korarchaeal genome reveals insights into the evolution of the Archaea. Proc Natl Acad Sci U S A. 105:8102–8107.

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

Embley TM, et al. 2003. Hydrogenosomes, mitochondria and early eukaryotic evolution. IUBMB Life. 55:387–395.

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

Esser C, Martin W, Dagan T. 2007. The origin of mitochondria in light of a fluid prokaryotic chromosome model. Biol Lett. 3:180–184.

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

Forterre P. 1995. Thermoreduction, a hypothesis for the origin of prokaryotes. C R Acad Sci III. 318:415–422.

Forterre P. 2011. A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong. Res Microbiol. 162:77–91.

Forterre P, Gribaldo S. 2010. Bacteria with a eukaryotic touch: a glimpse of ancient evolution? Proc Natl Acad Sci U S A. 107:12739–12740.

Gabaldon T, Huynen MA. 2003. Reconstruction of the proto-mitochondrial metabolism. Science 301:609–609.

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

Gould GW, Dring GJ. 1979. Possible relationship between bacterial endospore formation and the origin of eukaryotic cells. J Theoret Biol. 81:47–53.

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

Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C. 2010. The origin of eukaryotes and their relationship with the Archaea: are we a phylogenomic impasse? Nat Rev Microbiol. 8:743–752.

Gupta RS, Golding GB. 1996. The origin of the eukaryotic cell. Trends Biochem Sci. 21:166–171.

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

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 89:10915–101919.

Hooper SD, Berg OG. 2003. On the nature of gene innovation: duplication patterns in microbial genomes. Mol Biol Evol. 20:945–954.

Horiike T, Hamada K, Miyata D, Shinozawa T. 2004. The origin of eukaryotes is suggested as the symbiosis of Pyrococcus into γ-proteobacteria by phylogenetic tree based on gene content. J Mol Evol. 59:606–619.

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

Huber H, et al. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 417:63–67.

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

Kelly S, Wickstead B, Gull K. 2011. Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. Proc R Soc Lond B Biol Sci. 278:1009–1018.

Kloesges T, Martin W, Dagan T. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol Biol Evol. 28:1057–1074.

Koonin EV. 2009. Darwinian evolution in the light of genomics. Nucleic Acids Res. 37:1011–1034.

Koonin EV. 2012. The logic of chance: the nature and origin of biological evolution. Upper Saddle River (NJ): FT Press.

Kurland CG, Andersson SG. 2000. Origin and evolution of the mitochondrial proteome. Microbiol Mol Biol Rev. 64:786–820.

Kurland CG, Collins LJ, Penny D. 2006. Genomics and the irreducible nature of eukaryotic cells. Science 312:1011–1014.

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

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

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

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

Lane N. 2009. Life ascending: the ten greatest inventions of evolution. London: Profile Books. 344 p.

Lane N. 2011. Energetics and genetics across the prokaryote-eukaryote divide. Biol Direct. 6:e35.

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

Langer D, Hain J, Thuriaux P, Zillig W. 1995. Transcription in Archaea: similarity to that in Eukarya. Proc Natl Acad Sci U S A. 92:5768–5772.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic evolutionary model. Mol Biol Evol. 11:605–612.

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

Makarova KS, Yutin N, Bell SD, Koonin EV. 2010. Evolution of diverse cell division and vesicle formation systems in Archaea. Nat Rev Microbiol. 8:731–741.

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

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

Martin W. 1999a. A briefly argued case that mitochondria and plastids are descendants of endosymbionts, but that the nuclear compartment is not. Proc R Soc Lond B Biol Sci. 266:1387–1395.

Martin W. 1999b. Mosaic bacterial chromosomes—a challenge en route to a tree of genomes. BioEssays 21:99–104.

Martin W. 2005. Archaebacteria (Archaea) and the origin of the eukaryotic nucleus. Curr Opin Microbiol. 8:630–637.

Martin W, Koonin EV. 2006. Introns and the origin of nucleus–cytosol compartmentalization. Nature 440:41–45.

Matsuzaki M, et al. 2004. Genome sequence of the ultrasmall unicellular red algae *Cyanidioschizon merolae* 10D. Nature 428:653–657.

Maynard-Smith J, Szathmáry E. 1995. The major transitions in evolution. Oxford: Oxford University Press. 346 p.

McInerney JO, Cotton JA, Pisani D. 2008. The prokaryotic tree of life: past, present, and future? Trends Ecol Evol. 23:276–281.

McInerney JO, Pisani D, Bapteste E, O'Connell MJ. 2011. The public goods hypothesis for the evolution of life on Earth. Biol Direct. 6:41.

McInerney JO, et al. 2011. Planctomycetes and eukaryotes: a case of analogy not homology. BioEssays 33:810–817.

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

Mereschkowsky C. 1910. Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis, einer neuen Lehre von der Entstehung der Organismen. Biol Centralbl. 30: 278–288, 289–303, 321–347, 353–367.

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

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

Mus F, Dubini A, Seibert M, Posewitz MC, Grossman AR. 2007. Anaerobic acclimation in *Chlamydomonas reinhardtii*—anoxic gene expression, hydrogenase induction, and metabolic pathways. J Biol Chem. 282:25475–25486.

Nikoh N, et al. 2010. Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. PLoS Genet. 6:e1000827.

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

Pilhofer M, Rosati G, Ludwig W, Schleifer KH. 2007. Coexistence of tubulin and ftsZ in different Prosthecobacter species. Mol Biol Evol. 24:1439–1442.

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

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. 2011. Directed networks reveal barriers and bypasses to lateral gene traffic among sequenced prokaryote genomes. Genome Res. 21:599–609.

Pruit KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33:D501–D504.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.

Richards TA, Archibald JM. 2011. Gene transfer agents and the origin of mitochondria. Curr Biol. 21:R112–R114.

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

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

Roettger M, Martin W, Dagan T. 2009. A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. Mol Biol Evol. 26:1931–1939.

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

Schnarrenberger C, Martin W. 2002. Evolution of the enzymes of the citric acid cycle and the glyoxylate cycle of higher plants: a case study of endosymbiotic gene transfer. Eur J Biochem. 269: 868–883.

Searcy DG. 1992. Origins of mitochondria and chloroplasts from sulphur-based symbioses. In: Hartman H, Matsuno K, editors. The origin and evolution of the cell. Singapore: World Scientific. p. 47–78.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

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

Swarbreck D, et al. 2008. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res. 36:D1009–D1014.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic respective on protein families. Science 24:631–637.

Tatusov RL, et al. 2003. The COG database an updated version includes eukaryotes. BMC Bioinformatics. 4:41.

Thrash JC, et al. 2011. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Sci Rep. 1:e13.

Tielens AG, Rotte MC, van Hellemond JJ, Martin W. 2002. Mitochondria as we don't know them. Trends Biochem Sci. 27:564–572.

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

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

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

Vellai T, Takacs K, Vida G. 1998. A new aspect to the origin and evolution of eukaryotes. J Mol Evol. 46:499–507.

Wagner M, Horn M. 2006. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr Opin Biotech. 17: 41–49.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.

Yutin N, Makarova KS, Mekhedov SL, Wolf YI, Koonin EV. 2008. The deep archaeal roots of eukaryotes. Mol Biol Evol. 25:1619–1630.

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

Associate editor: John Archibald

5.2 Concatenated alignments and the case of the disappearing tree

Thorsten Thiergart[1], Giddy Landan[2], William F. Martin[1]

[1] Institut für molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Deutschland

[2] Gruppe für genomische Mikrobiologie, Institut für Mikrobiologie, Christian-Albrechts-Universität Kiel, Deutschland

Dieser Artikel wurde im August 2014 bei BMC Evolutionary Biology eingereicht

Beitrag von Thorsten Thiergart:

Ich habe sämtliche Daten, die für Analysen benötigt wurden, erstellt (Einteiliung der Proteine in funktionelle Gruppen, Zusammenstellung der homologen Gruppen, Erstellung der phylogenetischen Bäume). Sämtliche Analysen wurden von mir durchgeführt, teilweise mit dem Bereitstellen von technischer Unterstützung von Dr. Giddy Landan. Eine Rohfassung des kompletten Manuskriptes wurde von mir verfasst, und später zusammen mit Prof. Dr. William Martin und Dr. Giddy Landan überarbeitet.

Concatenated alignments and the case of the disappearing tree

Thorsten Thiergart[1*], Giddy Landan[2], William F. Martin[1]

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

[2] Genomic Microbiology Group, Institute of Microbiology, Christian-Albrechts-Universität Kiel, Kiel, Germany.

* Corresponding author

Email addresses:

TT : thorsten.thiergart@hhu.de
GL : glandan@ifam.uni-kiel.de
WFM : W.Martin@uni-duesseldorf.de

Abstract

Background:

Analyzed individually, gene trees for a given taxon set tend to harbour incongruent or conflicting signals. One popular approach to deal with this circumstance is to use concatenated data. But especially in prokaryotes, where lateral gene transfer (LGT) is a natural mechanism of generating genetic diversity, there are open questions as to whether concatenation amplifies or averages phylogenetic signals residing in individual genes. Here we investigate concatenations of prokaryotic and eukaryotic datasets to investigate possible sources of incongruence in phylogenetic trees and to examine the level of overlap between individual and concatenated alignments.

Results:

We analyzed prokaryotic datasets comprising 248 invidual gene trees from 315 genomes at three taxonomic depths spanning gammaproteobacteria, proteobacteria, and prokaryotes (bacteria plus archaea), and eukaryotic datasets comprising 279 invidual gene trees from 85 genomes at two taxonomic depths: across plants-animals-fungi and within fungi. Consistent with findings from previous studies, the branches in trees made from concatenated alignments are, in general, not supported by any of their underlying individual gene trees, even though the concatenation trees tend to possess very high bootstrap proportions values. For the prokaryote data, this observation is independent of both phylogenetic depth and sequence conservation. The eukaryotic data show better agreement between concatenated and single gene trees. Sequence length in individual alignments, but not sequence divergence, was found to correlate with the generation of branches that correspond to the concatenated tree. Simulated data reproduced this result, but the overall congruence in the simulated data was much higher than in real data.

Conclusions:

The weak correspondence of concatenation trees with single gene trees gives rise to the question where the phylogenetic signal in concatenated trees is coming from, if it is not coming from the individual alignments. In general, the eukaryote data reveals a better correspondence between individual and concatenation trees than the prokaryote data. But the question of whether the lack of correspondence between individual genes and the concatenation tree in the prokaryotic data is due to LGT is remains unanswered, because were LGT the cause of incongruence between concatenation and individual trees, we would have

expected to see greater degrees of incongruence for more divergent prokaryotic data sets, which was however not observed.

Keywords: phylogeny, concatenation, conflicting signals, bootstrapping

Background

Constructing trees out of concatenated alignments is now common practice in phylogenetics [1, 2]. A problem encountered in some of the first concatenation studies is that the concatenation tree is fully supported via bootstrapping at many or all branches but trees for the individual genes do not support the concatenation result, or conflict with it [3, 4]. In investigations of prokaryotic gene trees, the topological differences between individual trees underlying a concatenation are usually ascribed to lateral gene transfer (LGT) [5], which is not unreasonable, because prokaryotes really do undergo LGT frequently and have several biochemically and genetically well-characterized mechanisms to spread their genes within and across taxonomic boundaries: conjugation, transformation, transduction and gene transfer agents [6].

However there are other potential sources of phylogenetic conflict between gene trees and concatenated alignment trees. One of them is uncertain orthology or hidden praralogy. For example, Rinke et al. [7] examined a tree of concatenated alignments comprising newly characterized archaeal lineages, the concatenated result recovered the familiar three domains tree, with eukaryotes branching as sisters to archaebacteria. Williams and Embley [8] reinspected that data and found that the sequence collection procedure used by Rinke et al. [7] had included several nuclear genes of mitochondrial and plastid origin among the eukaryotic sequences; when those were removed and replaced by eukaryotic nuclear genes that had not been acquired from mitochondria or plastids, the two-domain tree was obtained [8], in which eukaryotes branch within the archaea [9]. Another source of conflict is phylogenetic error due to unknown factors that are often subsumed into the term model misspecification. For sequences from 10 plastid genomes, where neither paralogy nor orthologous replacement of sequences via LGT are known to occur, the species tree was fully resolved by the concatenation of 42 protein coding plastid gene families, but only 11 of the 42 gene trees recovered the concatenated topology, the remainder supported different trees [4]. The reason for the differing results are best explained by the circumstance that different proteins undergo amino acid substitution in different ways over evolutionary time, and according to different processes, models for which can be approximated mathematically [10].

One of the more controversial applications of alignment concatenation concerns its use to construct phylogenies for prokaryotes. At the center of the debate is the question whether there is a meaningful phylogeny of prokaryotes or not [11] and if so, does it extend back to the depths of evolutionary time [12], or does a tree only exist for the tips of prokaryotic trees [13]. In genomes, there exists a set of about 33 genes that are universally conserved among prokaryotes and that can readily be identified using standard ("manual") sequence comparison procedures [14, 15]. The existence of that universal set has been confirmed using semi-automated procedures [16]. Concatenation of those alignments produces a tree [14, 15, 16], but individually, the proteins in question do not tend to support any particular branching order, especially for the deeper branches or prokaryote phylogeny [17, 18].

Why do concatenation trees that are strongly supported, in terms of bootstrap proportions, fail to be supported by the individual gene trees constructed from the same underlying data? We reasoned that if LGT is the cause of conflict between individual gene trees, then its effect should be greater in prokaryotic than in eukaryotic data sets of similar sequence divergence, because LGT is far more prevalent among prokaryotes than it is among eukaryotes [19]. If model misspecification is the cause, then prokaryotic and eukaryotic data sets of similar sequence divergence should show similar levels of conflict. In prokaryote genomes, analyses of more closely related prokaryotic sequences should uncover greater congruence than for more distantly related prokaryotic sequences, because accurate phylogenetic inference becomes more problematic as sequence divergence increases [9] and because both LGT and sequence divergence accumulate over time [20]. In an effort to discriminate these possible causes, we undertook investigations of real data analyzed as individual and concatenated alignments.

Methods

Data

Proteome datasets were downloaded from RefSeq database [21]. These were: 1606 prokaryotic proteomes (v03.2012), 81 fungal proteomes (v03.2012), 86 animal proteomes (v03.2013) and 22 plant proteomes (v03.2013).

Gene families

Prokaryotic gene families were retrieved from the clusters of orthologous groups database (COGs, [22]). To avoid bias of the sampling and ensure an even taxonomic representation of the major taxonomic phyla of both prokaryotic domains, 50 archaeabacterial and 50 eubacterial genomes were chosen for further analysis. We avoided highly reduced genomes in our sample and were thus able to identify 48 genes that were present in a sample of 100 prokaryotic genomes containing 50 bacteria and 50 archaea (Supp. File 1). Homologues gene sequences of these 48 gene families were collected for two additional datasets, one containing 100 proteobacteria (Supp. File 2) and one containing 100 gammaproteobacteria (Supp. File 3). Additionally, a search was performed within all gammaproteobacteria species, yielding a dataset comprising 100 gammaproteobacteria species (Supp. File 6) and 200 universal gene families. For comparions between eukaryotic datasets and gammaproteobacteria data, this dataset was pruned to 50 taxa (Supp. File 6) .

Two datasets were generated for the eukaryotic analysis: one comprising only fungal species, and one containing plant, fungal and animal sequences. Universal protein families were reconstructed by an initial search for similar proteins with BLAST [23]. BLAST hits above 35% identity, an e-value $\leq 10^{-10}$ and an alignment length ≥ 75 were retained. Sequence pairs with ≥30% global identity using the needle algorithm (EMBOSS-package, [24]) were used as input for clustering with MCL [25]. Protein families were then sorted according to their universality. The first 200 families were chosen for the fungal set (50 species, Supp. File 5), 79 universal protein families were retrieved from the mixed eukaryote set (50 species, supp. File 4). A taxonomic flittering procedure was applied on both datasets to reduce oversampling. To filter for possible paralogous sequences in all datasets, the subset of all possible paralogs/orthologs that have the smallest reciprocal distance and that included all species having multiple copies was chosen. Clusters in which the subset did not include all species were not considered further.

Alignments and phylogenetic methods

Sequences were aligned with MAFFT (multiple alignment using fast Fourier transformation, v6.832b) using the “grinsi” parameters [26]. Trees were constructed with RAxML v7.0.4 [27]. The substitution rate per site was estimated from a gamma distribution with four discrete rate categories and the WAG substitution matrix [28]. The proportion of invariable sites was estimated from the data. Concatenated alignment trees were generated from the original alignments for the different datasets. Alignments were weighted, so that they have similar length within the concatenated alignment to avoid length bias towards longer alignments. Phylogenetic trees from prokaryotic datasets were rooted i) between archaea and bacteria, ii) using epsilonbacteria as the outgroup for the proteobacterial dataset or iii) using *Francisella* sp. as outgroup for the gammaproteobacteria. Phylogenetic trees from the eukaryotic datasets were rooted between plants and fungi/animals or between ascomycetes and basidiomycetes in the case of the fungi dataset. Full species names and additional taxonomic information are given in supplemental tables 1-5.

Simulations

Simulated alignments were created using a modified DAWG [29] version that is able to simulate evolution of amino acid sequences. The input tree was obtained from the weighted concatenated alignment of the γ-proteobacteria dataset, consisting of the 48 conserved genes. Datasets with an alignment length of 200 and 1000 positions were simulated, using the following DAWG parameters: Tree Scale=1, SubsModel=WAG, IndelModel=zipf, Indel Param 1.6, 100, Indel Rate 0.0011. This specific indel rate was used to match the one obtained for the alignments that originated the input tree.

Statistical analysis

All incompatible splits that were present in a given set of gene family trees were referred as the split pool. Pairs of splits were classified as compatible, when they can occur simultaneously in a binary tree, and classified as incompatible otherwise [30]. For each node in the concatenation trees, the amount of identical nodes within gene family trees were counted. This value is termed the node score. All three splits that are connected with a respective node were considered for this. If all three splits for each node, were compatible with the splits from the second node, both were accepted as similar. The topological distance from the root to the tip of a tree was calculated as the average number of branches separating a node from its descendant leafs. All statistical tests were performed using Matlab. Correlation measurements were done using the Pearson's linear correlation coefficient. To test if the difference between node score values for different datasets is significant, we used the MATLAB *multcompare* function (based on a one-way analysis of variance, alpha 0.05).

Results

The disappearing tree phenomenon

Concordance between the branches in individual gene trees and their concatenated incarnation is weak, as suggested by earlier studies [5, 31]. For the present data, this is shown in Figure 1, using a dataset of 48 genes present in three samples of 100 prokaryotic genomes spanning three phylogenetic depths: 50 archeabacteria and 50 eubacteria (Fig. 1a), 100 proteobacteria (Fig. 1b), and 100 gammaproteobacteria (Fig. 1c). In each 100 genome, 48 gene sample, the frequency of branches in 48 individual gene trees were compared to the set of branches in the concatenation tree. For each internal node within the concatenation tree, the node score was specified as the number of times that the corresponding node was observed among the 48 individual gene trees.

For the most divergent data set (Fig. 1a), deeper internal nodes of the concatenated tree have almost no congruence with the nodes in single gene trees, except the branch separating archaebacteria and eubacteria. At the tips of the tree, much greater congruence between the individual genes and the concatenation tree is observed. Surprisingly, the same "tree of tips" [13] or "disappearing tree" [31] phenomenon was observed for the proteobacterial sample (Fig. 1b) and for the gammaproteobacterial sample (Fig. 1c). For all three samples of phylogenetic/taxonomic depth, congruence between the deeper internal branches and branches recovered in individual trees disappears, yet the bootstrap proportions (BP) for virtually all branches in the concatenation trees were very high: for all three concatenation trees combined, only 15 internal branches had a BP *below* 80 (nodes marked with a red dot in Fig.1) and the average BP was 90, 99, and 96 for Fig1a-c respectively.

For the deep prokaryote sample, there are mainly two areas of low BPs within the tree: one within the euryarchaeota, and one spanning firmicutes, actinobacteria and tenericutes. The corresponding node support values relative to individual trees are low as well. But it is clearly visible that the rest of the internal branches have low node scores (congruence among individual trees) and high bootstrap support (site pattern sampling in the concatenation tree). The total number of splits present within the 48 single gene trees (split pool), is a simple measure to reflect the observed incongruence within the concatenation tree. On the range between total congruence with a split pool of 97 splits and total incongruence with 4,656 possible splits, 1,830 different splits were observed for the set of 48 trees summarized in Figure 1a. For trees in Fig1 b and c, 1,905 and 1,804 splits were observed, respectively. In other words, each internal branch of the species tree generates more than 18 conflicting splits on average. Especially for deeper phylogenetic relationships, the topology in the concatenation tree is not present in any of the family gene trees, despite the corresponding branches of the concatenation tree showing high BP values.

Influence of LGT – Comparing eukaryotic and prokaryotic data

The main effect of LGT on prokaryote genome evolution is to alter the number and kinds of genes that are found in a prokaryotic genome, not to promote orthologous replacement [32]. Thus, if LGT is the main reason why the present set of prokaryotic "core" genes analyzed individually tend to obtain different phylogenetic results, then this tendency should be more pronounced in prokaryotes than in eukaryotes. This is because eukaryotes counteract Muller's ratchet using meiosis and sex (process that generate reciprocal recombination), while prokaryotes rely on mechanisms of LGT — transformation, conjugation and transduction — processes that spread genes unidirectionally from donors to recipients. In order to see whether the congruence between concatenation trees and individual phylogenies is greater in prokaryotes or eukaryotes, we compared two additional datasets: one comprising of 50 fungal genomes, and one comprising of 50 eukaryotes, spanning plants, animals and fungi (PAF). Both datasets were composed of 50 genes with comparable length and different average pairwise identities (61% in fungi, 49% in the mixed set).

The results, summarised in Figure 2 a and b, show that both eukaryotic concatenation trees tend to have weaker node scores in the deeper branches than at the tips, like the prokaryote concatenation trees, but the overall agreement between concatenation trees and individual gene trees is far better for the eukaryotic data than for the prokaryotic data. As in the prokaryotic example, the eukaryotic concatenation trees show high BPs, averaging 96 and 97, respectively. The PAF tree shows a clear correspondence between low bootstrap support and low node score in the clade spanning the higher plants. But, as in the case of the prokaryotic trees, sampling at increasing phylogenetic depth does not reduce the congruence between individual gene trees and concatenated trees, as the average node score, 25%±14, for the fungal data set (Fig. 2b) is slightly higher than the value for the plant-animal-fungi dataset, 19%±11 (Fig. 2a)($P = 0.026$). Out of possible 2,350 splits we observed 350 different splits within the PAF dataset and 390 splits within the fungi dataset.

Factors affecting node scores

We investigated different factors that might affect node scores, which are a proxy for the tendency of individual trees to recover branches found in the concatenated tree. For this, we plotted, for each node in the concatenation tree, the frequency with which it was recovered in different data samples in order of increasing frequency (abscissas in Fig 3).

First we looked at phylogenetic depth (Fig 3a) because distantly related groups have distantly related sequences, which are notoriously hard to align, and their phylogenetic analysis can be further hampered by substitution levels that can approach saturation or algorithmic biases such as long branch attraction. The prokaryotic datasets shown in Figure 1 were compared. All three samples — prokaryotes, proteobacteria, gammaproteobacteria — encompass the same 48 genes, but because of their different phylogenetic depth, they span diferent levels of sequence divergence, the average pairwise identity being 32%, 48% and 67% respectively. Perhaps surprisingly, there is no significant difference (P = 0.67, P = 0.40, P = 0.70) between the node score distributions of the three samples (Fig3 A), despite the samples spanning a twofold decrease in average pairwise sequence identity. Thus, for these samples, phylogenetic depth is not a cause of low node scores.

In Fig 3b we plotted node score distributions for the eukaryotic data sets shown in Fig. 2. The comparison of the plant-animal-fungi vs. the fungal samples also revealed no significant difference, such that, like the prokaryotic samples, increasing sequence divergence stemming from greater phylogenetic depth (52% average pairwise identity PAF vs 58% fungi) had no detectable effect on node scores. To see if differences between prokaryote and eukaryote samples could be detected, we constructed a gammaproteobacterial sample with the same number of sequences and taxa (50) as the eukaryotic samples and consisting of genes with similar lengths (avg. 441 gammaproteobacteria, avg 438 PAF, avg. 441 fungi) and similar sequence conservation (avg. 58% for the gammaproteobacteria). Despite having very similar sequence atttributes as the eukaryotic samples, the node score distribution for the 50-genome gammaproteobacterial sample is strongly shifted towards lower values and is significantly different from that for the eukaryotes (P = 0.0007 , $P = 1.96 \times 10^{-5}$) (Fig. 3b). This would be consistent with an effect of LGT in the gammaproteobacterial sample, but if so, it remains puzzling why we do not see a decrease in the prokaryotic node score with increasing phylogenetic depth (Fig. 1, Fig 3a). Notably here the two eukaryote sets show no significant difference in their node score distribution (P = 0.2377).

Figure 3c shows the node score distributions for gammaproteobacterial gene samples of 50 genes each that were separated into categories of differing in sequence divergence (average pairwise sequence identity 44%, 61% and 70% respectively). No significant difference between the node score distributions was observed (P = 0.59*,* P = 0.86*,* P = 0.46). This suggests that sequence divergence at similar phylogenetic depth is not a factor affecting node score.

Another possible factor affecting generation of the same branches in individual and concatenated analyses is sequence length, or small site sample size. To check this, we assembled three more samples from the gammaproteobacterial data, each consisting of 50 genes for the same 100 species. The three samples consist of sequences with different average sequence length (124aa, 305aa, 692aa). The distributions of the node scores for the individual genes vs. the respective concatenation tree in two of the three samples are significantly different (P =3.8 x 10^{-5} , P =0.0172), with the longer sequences providing higher values than the shorter sequences (Fig 3d).

To investigate this effect further, we assembled fungal and proteobacterial datasets consisting of 200 genes each for 50 genomes and binned the individual alignments by their sequence length. As a measure of tree similarity within each bin, we simply counted the number of different splits observed for the five trees in each bin. In the case of five identical topologies, we would observe 47 splits, in the case that no common branches were observed across all five trees in a bin, we would observe 235 splits. The numbers of splits observed in each bin are plotted against sequence length in Figure 4a. A very strong correlation is observed both for the gammaproteobacterial (r = –0.87, P = 2.2 x 10^{-13}) and for the fungal bins (r = –0.8, P = 3.0 x 10^{-10}). The corresponding analysis for alignment length, rather than sequence length, takes the influence of gaps into account, and very similar distributions to those obtained for sequence length were obtained (suppl fig. 1: gammaproteobacterial sample r = –0.78, P = 2.7 x 10^{-9}, fungal sample r = –0.73, P = 5.6 x 10^{-8}). Although these fungal and gammaproteobacterial samples have comparable sequence lengths (in terms of having comparable counterparts) and similar average pairwise identity distributions (fungi: 56%±5; gamma: 59%±6), the fungal data tends much more strongly to recover the same tree than the gammaproteobacterial sample does, this again might point to a greater role for LGT in the gammaproteobacterial genes than in the fungal genes sampled.

Because bootstrapping provides information about the number of sites with similar distributions of site patterns needed to obtain the same tree in every pseudosample [33], it is perhaps not surprising that the average BP for each tree in the 200-gene gammaproteobacterial and fungal samples is strongly correlated with sequence length (Fig 4c). In the case of the gammaproteobacteria data (200 trees, all having the same 50 species), there is a strong positive correlation between bootstrap support in a gene family tree and sequence length (fig 3B, r = 0.76, P = 5.72 x 10^{-40}). A similar strong positive correlation between BS and sequence length is observed in the fungi dataset (fig 3B, r = 0.73, P = 1.7 x 10^{-35}). Other parameters do not show this strong correlation with BPs, or show no correlation at all. The alignment length has a slightly lower correlation with BP, than sequence length (gammaproteobacteria: r = 0.69, P = 2.23 x 10^{-30}, fungi: r = 0.68, P = 1.06 x 10^{-28}). The pairwise identity of genes within one gene family tree appears to correlate with BP (gammaproteobacteria: r = –0.31, P = 4.81 x 10^{-6}, fun: r = –0.44, P = 3.5 10^{-11}), but much less strongly than sequence length. Moreover, sequence length and pairwise identity are themselves only weakly or not correlated (fungi: r = –0.16, P = 0.017, gammaproteobacteria: r = –0.03, P = 0.598).

Simulations to investigate the influence of sequence length

To see if the sequence length effect is repeatable with perfect alignments, we simulated alignments along a known evolutionary history. As an input for these alignments we used the concatenation tree made of the 48 conserved genes from gammaproteobacteria (Fig. 1). Two datasets consisting of 50 alignments were generated, one with an initial alignment length of 1,000 positions, one with 200 positions. The dataset based on 1,000-position long alignments yield a nearly perfect distribution of splits. Nearly all of the 50 trees supporting the same splits, meaning all the trees are almost identical. The 200-position dataset trees have twice the amount of splits in their split pool than the longer ones (107 vs. 225 splits). To check whether the alignment process itself makes a difference, two additional datasets were made by recovering the sequences from the simulated alignment and aligning them using the same procedure as for the biological sequences. Again, no effect was detected. Increasing the tree length (sequence divergence) by a factor of three for the shorter 200 position alignments increases the number of individual splits to 350, which is still much less than observed in real data.

Comparing the distribution of incompatible splits between simulated data and real data make the differences more obvious (Fig 5). In real data, in this case the gammaproteobacteria, most of the observed splits appear only in one of the trees. This is true for data made from short sequences as well as for long sequence data (Fig 5 A, B). Whereas in the long sequence data, the number of splits observed in single trees is strongly reduced. Within the simulated data most of the splits are present in all trees (Fig 4 C, D). In the short simulated data, some splits are only present in single trees.

Discussion

Reconstructing a single phylogenetic tree from a collection of individual genes by using concatenated alignments has long been common practice in phylogenetic analyses. Although concatenation is widely implemented, most investigations of its underlying properties are, like the present study, empirical rather than theoretical in nature [34, 35, 36, 37, 38]. The result is that observations and correlations can be gleaned regarding the behaviour of the data in concatenation, but the responsible causalities remain obscure.

Concatenation entails the *a priori* assumption that the individual genes in question evolved along a common phylogeny. This is often difficult to demonstrate for real data, especially for data from prokaryotes [5]. Thus, inferences that are based on concatenation trees assume — explicitly or implicitly — that the concatenated genes were not subject to processes such as recombination, gene conversion, lateral gene transfer and the like, processes that are not fundamentally tree-like in nature. Yet even when all genes follow the same phylogeny, their trees might still differ owing to variety of aspects, such as evolutionary rates, selective, structural and functional constrains, and the level of stochastic noise introduced by neutral substitutions. Such evolutionary mechanisms can lead to model misspecification even in the analysis of a single gene family. In the context of alignment concatenation, however, the problem becomes acute, since no single model can subsume all genes simultaneously, and model misspecification is more or less guaranteed. Some current methods of phylogenetic inference can deal with such factors better than others [9].

The reliability of phylogenetic trees reconstructed from concatenated alignments can be assessed from two opposing perspectives. Bootstrap analysis, which originally was proposed as a methodology appropriate for single gene trees [33], can be applied to any alignment-like data, such as the concatenated alignment of several genes. This approach ignores the fact that different parts of the concatenated alignment originate from different genes, and focuses on the robustness of the estimated topology given the totality of the sequence data. An alternative approach views concatenated alignment trees as consensus-like, and focuses on the congruence between such trees and the underlying gene trees [34, 35, 37]. In the presence of long alignments, bootstrap analysis typically assigns high support to almost every branch of the concatenation tree while comparison to the individual gene trees indicates that congruence is observed only at the tips of the tree, and that deeper internal branches are typically highly incongruent among gene trees and between gene trees and the concatenation tree. The high bootstrap support observed here for concatenated alignments may be artificial, resulting from the large sample size and possibly biased by signals generated by a few genes. It is well-known that bootstrap and similar support values increase with the increasing number of sites sampled [34] such that a high BPs for a concatenated phylogeny does not simultaneously mean that the tree is thus likely to be correct [35, 37, 39]. For very large data sets, phylogenetic results become increasingly dependent upon the model, rather than the number of sites sampled [9, 34]. Congruence analysis, on the other hand, reveals the variety of evolutionary signals in the underlying collection of genes, and thus provides a more conservative interpretation of the phylogenomic signals, thereby informing data collation strategies.

Galtier [36] showed higher levels of congruence for eukaryotic than for prokaryotic data, similar to our present findings, and furthermore that in bacteria the congruence is slightly positively correlated to the sequence length of the chosen genes, an effect that we observed in a more pronounced manner in the present data. In a study encompassing 21 fungal species and 246 single copy genes [38], gene size was also shown to be a proxy for the phylogenetic performance of individual genes, an effect detected in all gene samples examined here. Although we suspect that LGT in prokaryotes might underlie the finding that congruence between individual trees and the concatenation tree is higher for data from eukaryotic genomes than it is for prokaryotic genomes, a causal relationship could not be established for this effect.

Conclusions

In general, for the prokaryotic data we observe, like others before us [13, 17], a tree of tips, where the terminal branches seem well supported but the deeper branches are not recovered by any of the individual genes studied. Unexpectedly for us, this was observed recurrently for three data sets spanning very different phylogenetic depths among prokaryotes, almost in a fractal-like manner. The lack of congruence among individual genes for deeper branches, which show high BPs in the concatenated analyses, we call the "disappearing tree" effect. Its cause remains obscure, but it provides a source of many caveats when it comes to attempting to infer evolutionary events from branches with high BPs in prokaryotic genome phylogenies. If an ancient evolutionary signal is real, for example the bacteria-archaea split [40], then it should be supported by individual genes, which we observe in the present study. Concatenation is an important aspect of modern phylogenomics and is not likely to go away any time soon, it is therefore all the more important to understand the properties of concatenation and its relationship to the individual underlying trees.

Competing Interests: The authors declare that they have no competing interests.

Authors contributions: TT carried out the computational analysis and drafted the Manuscript. GL designed analyses and contributed to the computational analysis. WFM designed the study. All authors interpreted results, drafted the manuscript, read and approved the final version.

Acknowledgments: We thank Filipa L. Sousa for helping to prepare the manuscript. Computational support and infrastructure was provided by the "Center for Information and Media Technology" (ZIM) at the Heinrich Heine University of Duesseldorf (Germany). G.L. is supported by an ERC grant (281357 to Tal Dagan).

References

1. Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF: **A kingdom-level phylogeny of eukaryotes based on combinded protein data**. *Science* 2000, **290**:972-977

2. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ: **Universal trees based on large combined protein sequence data sets**. *Nat Genet* 2001, **28**:281-285.

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

4. 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.

5. Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF: **Do orthologous gene phylogenies really support tree-thinking?** *BMC Evol Biol* 2005, **5**:33.

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

7. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Wouke T: **Insights into the phylogeny and coding potential of microbial dark matter**. *Nature* 2013, **499**:431-437.

8. Williams TA, Embley TM: **Archaeal “dark matter” and the origin of eukaryotes.** *Genome Biol Evol* 2014, **6**:474-481.

9. 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.

10. Lockhart PJ, Steel MA, Hendy MD, Penny D: **Recovering evolutionary trees under a more realistic evolutionary model.** *Mol Biol Evol* 1994, **11**:605-612.

11. Doolittle WF, Bapteste E: **Pattern pluralism and the tree of life hypothesis**. *Proc Natl Acad Sci USA* 2007, **104**:2043–2049.

12. Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C: **The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse?** *Nature Rev Microbiol* 2010, **8**:743-752.

13. Creevey CJ, Fitzpatrick DA, Philip GK, Kinsella RJ, O'Connell MJ, Pentony MM, Travers SA, Wilkinson M, McInerney JO: **Does a tree-like phylogeny only exist at the tips in the prokaryotes?** *Proc Roy Soc Lond B* 2004, **271**:2551–2558.

14. Hansmann S, Martin W: **Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes**. *Int J Syst Evol Microbiol* 2000, **50**:1655–1663.

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

16. 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.

17. Bapteste E, Sukso E, Leigh J, Ruiz-Trillo I, Bucknam J, Doolittle WF: **Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny**. *Mol Biol Evol* 2008, **25**:83-91.

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

19. Bapteste E, O'Malley M, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe F-J, Dupré J, Dagan T, Boucher Y, Martin W: **Prokaryotic evolution and the tree of life are two different things**. *Biol Direct* 2009*,* **4**:34.

20. Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T: **Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene traffic among prokaryotes**. *Genome Res* 2011*,* **21**:599–609.

21. Pruit KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2005, **33**:D501–504.

22. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution**. *Nucleic acids res* 2000, **28**:33-36.

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

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

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

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

27. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics* 2006, **22**:2688–2690.

28. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach**. *Mol Biol Evol* 2001, **18**:691–699.

29. Cartwright RA: **DNA assembly with gaps (Dwag): simulating sequence evolution**. *Bioinformatics* 2005, **21**:iii31-iii38.

30. Buneman P: **The Recovery of Trees from Measures of Dissimilarity**. In *Mathematics in the Archaeological and Historical Sciences*. Edited by Kendall DG and Tăutu P. Edinburgh University Press: 1971:387-395.

31. Sousa FL, Thiergart T, Landan G, Nelson-Sathi S, Pereira IAC, Allen JF, Lane N, Martin WF: **Early bioenergetic evolution**. *Philosophical transaction of the royal society B* 2013, 368:20130088.

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

33. Felsenstein J: **Confidence limits on phylogenies: An approach using the bootstrap**. *Evolution* 1985, **39**:783-791.

34. Phillips MJ, Delsuc F, Penny D: **Genome-scale phylogeny and the detection of systematic biases**. *Mol Biol Evol* 2004, **21**:1455-1458.

35. 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.

36. Galtier N: **A model of horizontal gene transfer and the bacterial phylogeny problem**. *Syst Biol* 2007, **56**: 633–642.

37. Nishihara H, Okada N, Hasegawa M: **Rooting the eutherian tree: the power and pitfalls of phylogenomics**. *Genome Biol* 2007, **8**:R199.

38. Aguileta G, Marthey S, Chiapello H, Lebrun MH, Rodolphe F, Fournier E, Gendrault-aquemard A, Giraud T: **Assessing the performance of single-copy genes for reovering robust phylogenies.** *Syst Biol* 2008, **57**:613–627.

39. 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.

40. 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.

Figure Legends:

Figure 1: Single gene tree support projected on three concatenated prokaruyotic trees of different taxonomic depth levels. All trees based on the concatenation of 48 universal genes. Nodes in concatenated trees were compared with nodes present in the underlying single gene trees. Each node and their outgoing branches were colored according to presence of this node

within single gene trees, from 0 to all 48 single gene trees. The trees include **A)** a prokaryotic dataset including 100 archaebacteria and eubacteria, **B)** 100 proteobacteria, **C)** 100 gammaproteobacteria species. Exact species names are given in supplementary file 1-3.

Figure 2: Single gene tree support projected on two concatenated eukaryotic trees of different taxonomic depth levels. All trees based on the concatenation of 50 universal genes, respectively. Nodes in concatenated trees were compared with nodes present in the underlying single gene trees. Each node and their outgoing branches were colored according to presence of this node within single gene trees, from 0 to all 48 single gene trees. The trees include **A)** 50 fungi, plant and animal species, **B)** 50 fungi. Exact species names are given in supplementary file 4-5.

Figure 3: Parameters influencing node score in concatenated trees and single gene trees. Nodes from concatenated trees were plotted according to their support level compared to single gene trees. All datasets based on 50 single gene trees, except in A), there are only 48 genes. **A)** Comparisons of the support level at different phylogenetic depths: prokaryotes, proteobacteria, gammaproteobacteria. **B)** Comparison of the support level for two eukaryotic and one prokaryotic dataset (all datasets, prokaryotic and eukaryotic, consists of 50 taxa), where the underlying single gene trees have similar average sequence length: Eukaryotes mixed: 438±96 aa, Fungi: 441±45 aa, gammaproteobacteria: 441±57aa **C)** Comparison of the support level at different pairwise identity levels. Values are average percent identities in all pairwise sequence comparisons. **D)** Comparison of the node score for different average sequence lengths. Values are the average length of all protein sequences in each set.

Figure 4: Correlation between sequence length, tree incongruence and bootstrap proportions. **A)** Number of different splits observed in bins of 5 trees is plotted against the average sequence length. **B)** Average bootstrap proportions within one tree plotted against the average sequence length. Both datasets, prokaryotic and eukaryotic, consists of 50 taxa. The polynomial regression is indicated as a colored dotted line. The black dotted line indicate the expected number of splits if all trees had identical topologies.

Figure 5: Splits distribution in single gene data sets for real and simulated data. For each data set all splits within single gene trees were plotted according to their topological distance to the tips of the tree and the number of trees where they are present. **A)** Dataset of 50 short

gammaproteobacteria genes. **B)** Dataset of 50 long gammaproteobacteria genes. **C)** Simulated dataset with initial sequence length of 200 positions. **D)** Simulated dataset with initial sequence length of 1000 positions.

Supp Fig1: Correlation between alignment length, tree incongruence and bootstrap proportions. **A)** Number of different splits observed in bins of 5 trees is plotted against the average alignment length. **B)** Average bootstrap proportions within one tree plotted against the average alignment length. The polynomial regression is indicated as a dotted line.

Fig.1

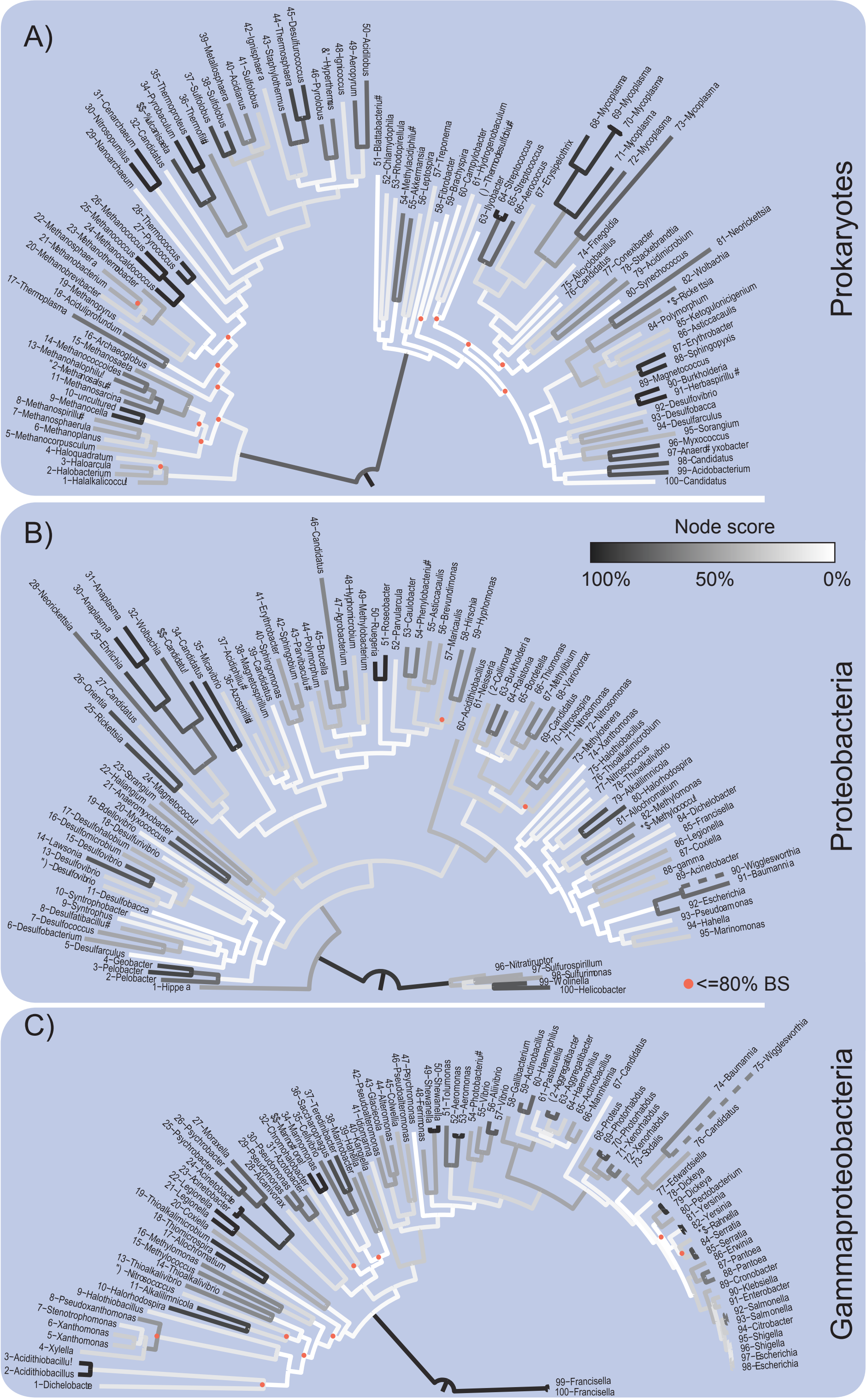

A)
1−Halalkalicoccu!
2−Halobacterium
3−Haloarcula
4−Haloquadratum
5−Methanocorpusculum
6−Methanoplanus
7−Methanosphaerula
8−Methanospirillu#
9−Methanocella
10−uncultured
11−Methanosarcina
"2−Methanosalsu#
13−Methanohalophilu!
14−Methanococcoides
15−Methanosaeta
16−Archaeoglobus
17−Thermoplasma
18−Aciduliprofundum
19−Methanopyrus
20−Methanobrevibacter
21−Methanobacterium
22−Methanosphaera
23−Methanothermobacter
24−Methanocaldococcus
25−Methanococcus
26−Methanococcus
27−Pyrococcus
28−Thermococcus
29−Nanoarchaeum
30−Nitrosopumilus
31−Cenarchaeum
32−Candidatus
$$−%ulcanisaeta
34−Pyrobaculum
35−Thermoproteus
36−Thermofil#
37−Sulfolobus
38−Sulfolobus
39−Metallosphaera
40−Acidianus
41−Sulfolobus
42−Ignisphaera
43−Staphylothermus
44−Thermosphaera
45−Desulfurococcus
46−Pyrolobus
&'−Hyperthermus
48−Ignicoccus
49−Aeropyrum
50−Acidilobus
51−Blattabacteriu#
52−Chlamydophila
53−Rhodopirellula
54−Methylacidiphilu#
55−Akkermansia
56−Leptospira
57−Treponema
58−Fibrobacter
59−Brachyspira
60−Campylobacter
61−Hydrogenobaculum
( )−Thermodesulfobiu#
63−Ilyobacter
64−Streptococcus
65−Streptococcus
66−Aerococcus
67−Erysipelothrix
68−Mycoplasma
69−Mycoplasma
70−Mycoplasma
71−Mycoplasma
72−Mycoplasma
73−Mycoplasma
74−Finegoldia
75−Alicyclobacillus
76−Candidatus
77−Conexibacter
78−Stackebrandtia
79−Acidimicrobium
80−Synechococcus
81−Neorickettsia
82−Wolbachia
*$−Ricke ttsia
84−Polymorphum
85−Ketogulonicigenium
86−Asticcacaulis
87−Erythrobacter
88−Sphingopyxis
89−Magnetococcus
90−Burkholderia
91−Herbaspirillu#
92−Desulfovibrio
93−Desulfobacca
94−Desulfarculus
95−Sorangium
96−Myxococcus
97−Anaero# yxobacter
98−Candidatus
99−Acidobacterium
100−Candidatus
Prokaryotes
B)
Node score
100%
50%
0%
1−Hippe a
2−Pelobacter
3−Pelobacter
4−Geobacter
5−Desulfarculus
6−Desulfobacterium
7−Desulfococcus
8−Desulfatibacillu#
9−Syntrophus
10−Syntrophobacter
11−Desulfobacca
") −Desulfovibrio
13−Desulfovibrio
14−Lawsonia
15−Desulfovibrio
16−Desulfomicrobium
17−Desulfohalobium
18−Desulfurivibrio
19−Bdellovibrio
20−Myxococcus
21−Anaeromyxobacter
22−Haliangium
23−Sorangium
24−Magnetococcu!
25−Rickettsia
26−Orientia
27−Candidatus
28−Neorickettsia
29−Ehrlichia
30−Anaplasma
31−Anaplasma
32−Wolbachia
$$−Candidatu!
34−Candidatus
35−Micavibrio
36−Azospirill#
37−Acidiphiliu#
38−Magnetospirillum
39−Candidatus
40−Sphingomonas
41−Erythrobacter
42−Sphingobium
43−Parvibaculu#
44−Polymorphum
45−Brucella
46−Candidatus
47−Agrobacterium
48−Hyphomicrobium
49−Methylobacterium
50−Ruegeria
51−Roseobacter
52−Parvularcula
53−Caulobacter
54−Phenylobacteriu#
55−Asticcacaulis
56−Brevundimonas
57−Maricaulis
58−Hirschia
59−Hyphomonas
60−Acidithiobacillus
61−Neisseria
(2−Collimona!
63−Burkholderia
64−Ralstonia
65−Bordetella
66−Thiomonas
67−Methylibium
68−Variovorax
69−Candidatus
70−Nitrosospira
71−Nitrosomonas
72−Nitrosomonas
73−Methylotenera
74−Xanthomonas
75−Halothiobacillus
76−Thioalkalimicrobium
77−Nitrosococcus
78−Thioalkalivibrio
79−Alkalilimnicola
80−Halorhodospira
81−Allochromatium
82−Methylomonas
*$−Methylococcu!
84−Dichelobacter
85−Francisella
86−Legionella
87−Coxiella
88−gamma
89−Acinetobacter
90−Wigglesworthia
91−Baumannia
92−Escherichia
93−Pseudomonas
94−Hahella
95−Marinomonas
96−Nitratiruptor
97−Sulfurospirillum
98−Sulfurimonas
99−Wolinella
100−Helicobacter
<=80% BS
Proteobacteria
C)
1−Dichelobacte
2−Acidithiobacillus
3−Acidithiobacillu!
4−Xylella
5−Xanthomonas
6−Xanthomonas
7−Stenotrophomonas
8−Pseudoxanthomonas
9−Halothiobacillus
10−Halorhodospira
11−Alkalilimnicola
") −Nitrosococcus
13−Thioalkalivibrio
14−Thioalkalivibrio
15−Methylococcus
16−Methylomonas
17−Allochromatium
18−Thiomicrospira
19−Thioalkalimicrobium
20−Coxiella
21−Legionella
22−Legionella
23−Acinetobacter
24−Acinetobacte
25−Psychrobacter
26−Psychrobacter
27−Moraxella
28−Alcanivorax
29−Pseudomonas
30−Pseudomonas
31−Azotobacter
32−Chromohalobacter
$$−Marino# ona!
34−Marinomonas
35−Cellvibrio
36−Saccharophagus
37−Teredinibacter
38−Marinobacter
39−Hahella
40−Kangiella
41−Idiomarina
42−Pseudoalteromonas
43−Glaciecola
44−Alteromonas
45−Colwellia
46−Pseudoalteromonas
47−Psychromonas
48−Ferrimonas
49−Shewanella
50−Shewanella
51−Tolumonas
52−Aeromonas
53−Aeromonas
54−Photobacteriu#
55−Vibrio
56−Aliivibrio
57−Vibrio
58−Gallibacterium
59−Actinobacillus
60−Haemophilus
61−Pasteurella
(2−Aggregatibacter
63−Aggregatibacter
64−Haemophilus
65−Actinobacillus
66−Mannheimia
67−Candidatus
68−Proteus
69−Photorhabdus
70−Photorhabdus
71−Xenorhabdus
72−Xenorhabdus
73−Sodalis
74−Baumannia
75−Wigglesworthia
76−Candidatus
77−Edwardsiella
78−Dickeya
79−Dickeya
80−Pectobacterium
81−Yersinia
82−Yersinia
*$−Rahnella
84−Serratia
85−Serratia
86−Erwinia
87−Pantoea
88−Pantoea
89−Cronobacter
90−Klebsiella
91−Enterobacter
92−Salmonella
93−Salmonella
94−Citrobacter
95−Shigella
96−Shigella
97−Escherichia
98−Escherichia
99−Francisella
100−Francisella
Gammaproteobacteria

Fig. 2

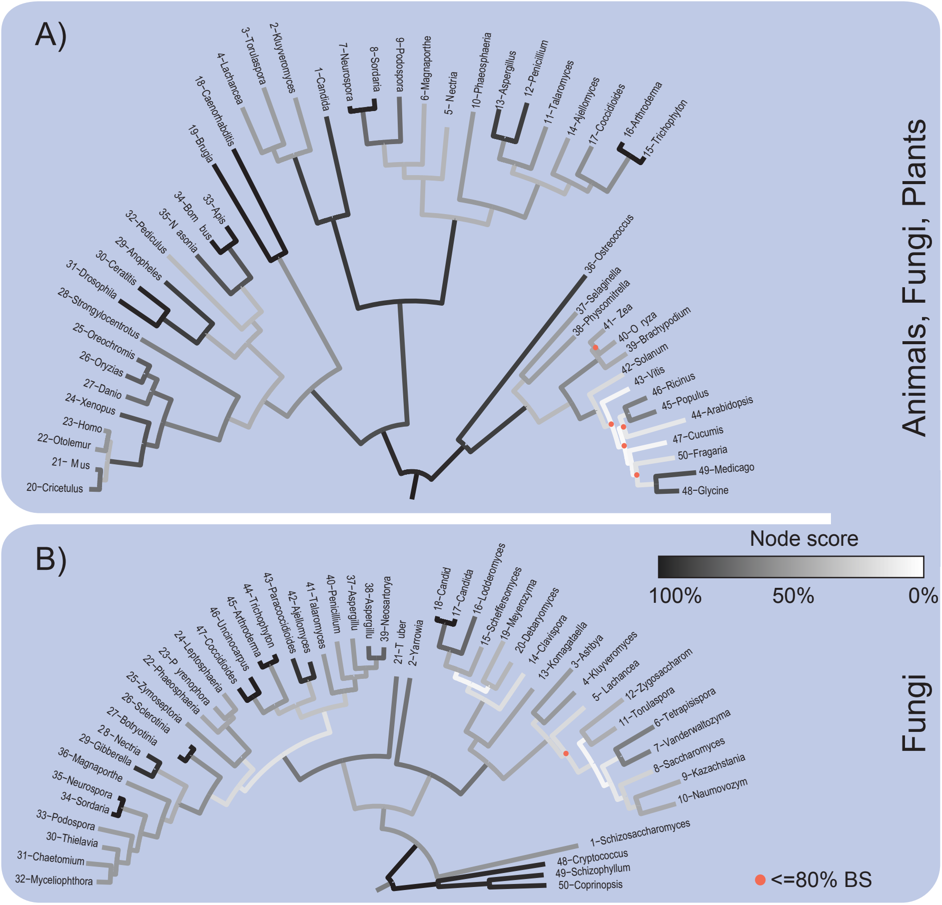

A)
Animals, Fungi, Plants
1–Candida
2–Kluyveromyces
3–Torulaspora
4–Lachancea
5– Nectria
6–Magnaporthe
7–Neurospora
8–Sordaria
9–Podospora
10–Phaeosphaeria
11–Talaromyces
12–Penicillium
13–Aspergillus
14–Ajellomyces
15–Trichophyton
16–Arthroderma
17–Coccidioides
18–Caenorhabditis
19–Brugia
20–Cricetulus
21– Mus
22–Otolemur
23–Homo
24–Xenopus
25–Oreochromis
26–Oryzias
27–Danio
28–Strongylocentrotus
29–Anopheles
30–Ceratitis
31–Drosophila
32–Pediculus
33–Apis
34–Bom bus
35–N asonia
36–Ostreococcus
37–Selaginella
38–Physcomitrella
39–Brachypodium
40–O ryza
41– Zea
42–Solanum
43–Vitis
44–Arabidopsis
45–Populus
46–Ricinus
47–Cucumis
48–Glycine
49–Medicago
50–Fragaria
B)
Fungi
1–Schizosaccharomyces
2–Yarrowia
3–Ashbya
4–Kluyveromyces
5– Lachancea
6–Tetrapisispora
7–Vanderwaltozyma
8–Saccharomyces
9–Kazachstania
10–Naumovozym
11–Torulaspora
12–Zygosaccharom
13–Komagataella
14–Clavispora
15–Scheffersomyces
16–Lodderomyces
17–Candida
18–Candid
19–Meyerozyma
20–Debaryomyces
21–T uber
22–Phaeosphaeria
23–P yrenophora
24–Leptosphaeria
25–Zymoseptoria
26–Sclerotinia
27–Botryotinia
28–Nectria
29–Gibberella
30–Thielavia
31–Chaetomium
32–Myceliophthora
33–Podospora
34–Sordaria
35–Neurospora
36–Magnaporthe
37–Aspergillu
38–Aspergillu
39–Neosartorya
40–Penicillium
41–Talaromyces
42–Ajellomyces
43–Paracoccidioides
44–Trichophyton
45–Arthroderma
46–Uncinocarpus
47–Coccidioides
48–Cryptococcus
49–Schizophyllum
50–Coprinopsis
Node score
100%
50%
0%
● <=80% BS

Fig. 3

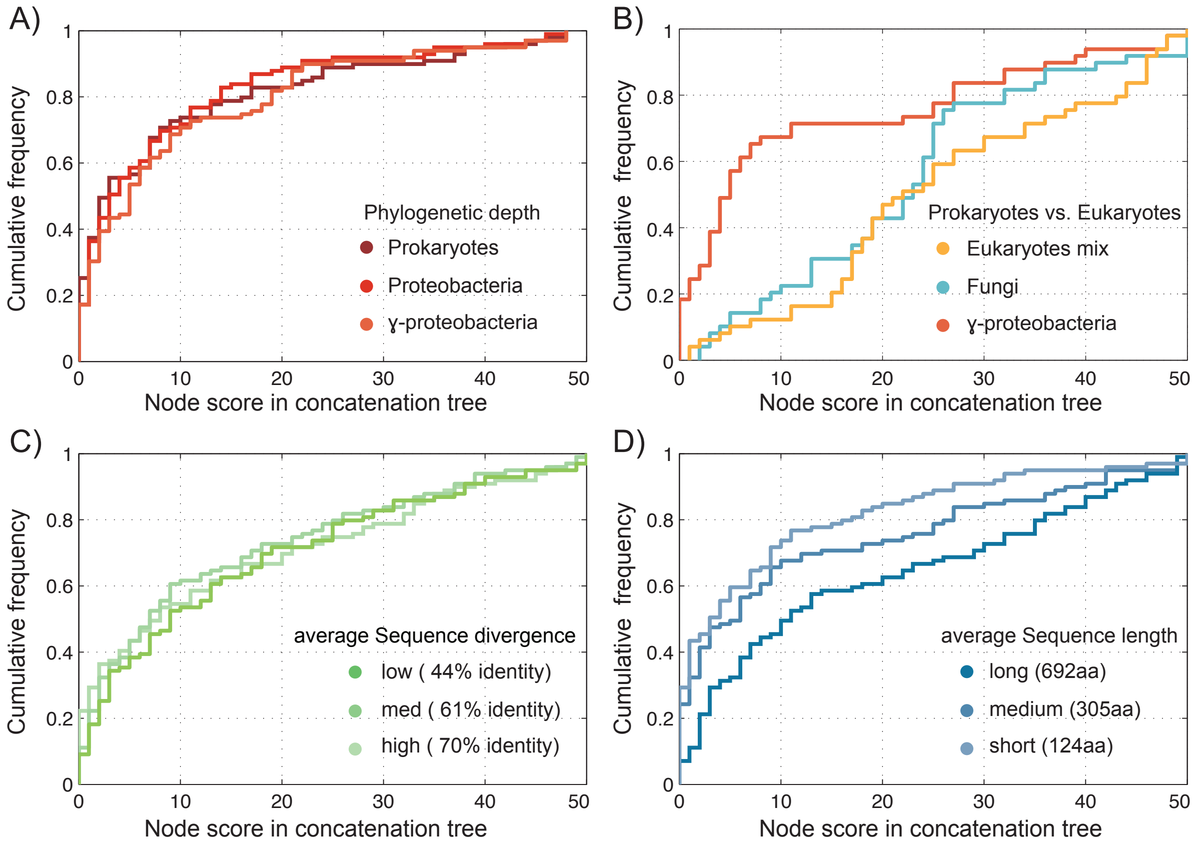

A)
Cumulative frequency
Node score in concatenation tree
Phylogenetic depth
● Prokaryotes
● Proteobacteria
● γ-proteobacteria
B)
Cumulative frequency
Node score in concatenation tree
Prokaryotes vs. Eukaryotes
● Eukaryotes mix
● Fungi
● γ-proteobacteria
C)
Cumulative frequency
Node score in concatenation tree
average Sequence divergence
● low (44% identity)
● med (61% identity)
● high (70% identity)
D)
Cumulative frequency
Node score in concatenation tree
average Sequence length
● long (692aa)
● medium (305aa)
● short (124aa)
0
0.2
0.4
0.6
0.8
1
0
10
20
30
40
50

Fig. 4

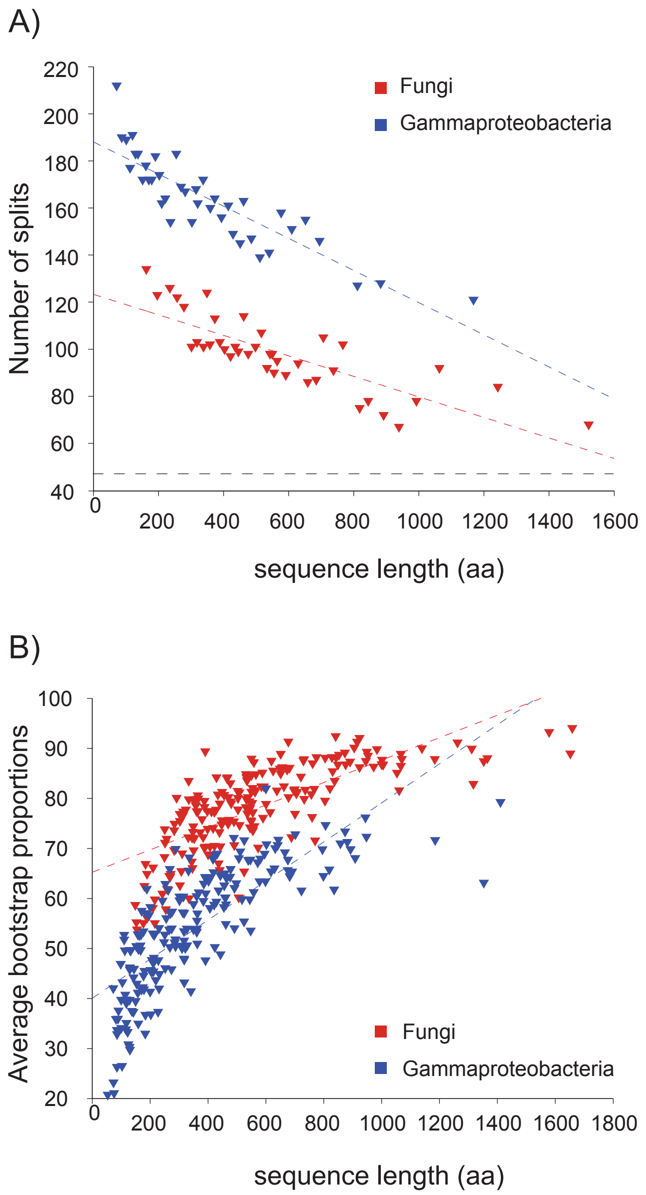
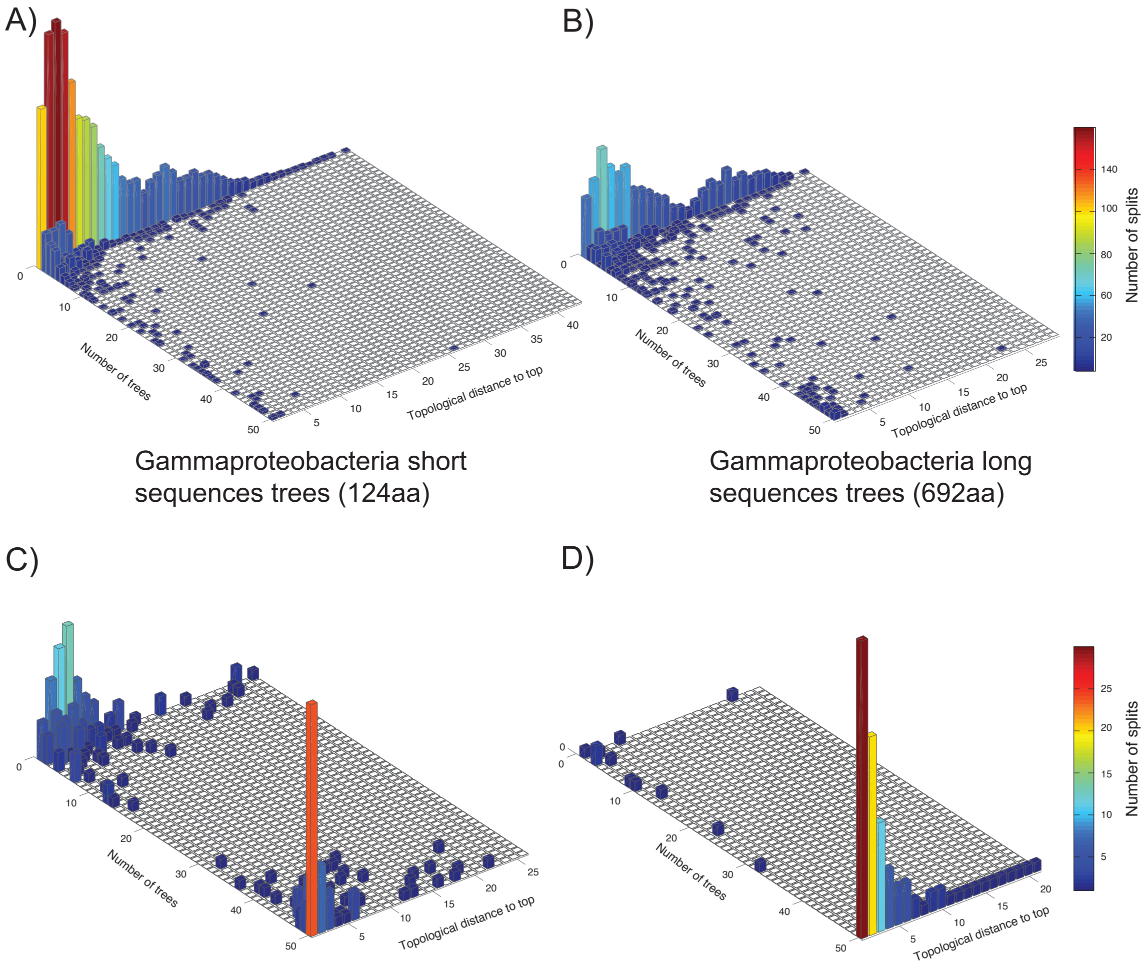

A)
Fungi
Gammaproteobacteria
Number of splits
220
200
180
160
140
120
100
80
60
40
0
200
400
600
800
1000
1200
1400
1600
sequence length (aa)
B)
Average bootstrap proportions
100
90
80
70
60
50
40
30
20
Fungi
Gammaproteobacteria
0
200
400
600
800
1000
1200
1400
1600
1800
sequence length (aa)

Fig. 5



Gammaproteobacteria short sequences trees (124aa)

Gammaproteobacteria long sequences trees (692aa)

Simulated short sequences trees(200aa)

Simulated long sequences trees (1000aa)

5.3 Early bioenergetic evolution

Filipa L. Sousa[1], Thorsten Thiergart[1], Giddy Landan[2], Shijulal Nelson-Sathi[1], Inês A. C. Pereira[3], John F. Allen[4,5], Nick Lane[5], William F. Martin[1]

[1] Institut für molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Deutschland

[2] Institut für genomische Mikrobiologie, Christian-albrechts-Universität Kiel, Kiel, Deutschland

[3] Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal

[4] School of Biological and Chemical Sciences, Queen Mary, University of London, London, UK

[5] Research Department of Genetics, Evolution and Environment, University College, London, London, UK

Dieser Artikel wurde 2013 in Philosophical Transactions of the Royal Society B veröffentlich Ausgabe 368

Beitrag von Thorsten Thiergart:

Ich habe die Analyse zur Bestimmung der Vorkommen von Eisen-Schwefel-Cluster formenden Proteinen in Prokaryoten durchgeführt (Abbildung 5), welche die Grundlage für Kapitel 7 im Manuskript bildet. Die Analyse der 48 universellen prokaryotischen Gene und ihrer phylogenetischen Bäume (Abbildung 6) wurde von mir durchgeführt, teilweise mit technischer Unterstützung von Dr. Giddy Landan. Diese Analyse war die Grundlage für Kapitel 11.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Review

Cite this article: Sousa FL, Thiergart T, Landan G, Nelson-Sathi S, Pereira IAC, Allen JF, Lane N, Martin WF. 2013 Early bioenergetic evolution. Phil Trans R Soc B 368: 20130088. http://dx.doi.org/10.1098/rstb.2013.0088

One contribution of 14 to a Discussion Meeting Issue 'Energy transduction and genome function: an evolutionary synthesis'.

Subject Areas:
evolution

Keywords:
transition metals, acetogens, methanogens, sulfate reducers, origin of life, hydrothermal vents

Author for correspondence:
William F. Martin
e-mail: bill@hhu.de

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2013.0088 or via http://rstb.royalsocietypublishing.org.

Early bioenergetic evolution

Filipa L. Sousa[1], Thorsten Thiergart[1], Giddy Landan[2], Shijulal Nelson-Sathi[1], Inês A. C. Pereira[3], John F. Allen[4,5], Nick Lane[5] and William F. Martin[1]

[1]Institute of Molecular Evolution, and [2]Institute of Genomic Microbiology, University of Düsseldorf, 40225 Düsseldorf, Germany
[3]Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal
[4]School of Biological and Chemical Sciences, Queen Mary, University of London, London, UK
[5]Research Department of Genetics, Evolution and Environment, University College London, Gower Street, London, UK

Life is the harnessing of chemical energy in such a way that the energy-harnessing device makes a copy of itself. This paper outlines an energetically feasible path from a particular inorganic setting for the origin of life to the first free-living cells. The sources of energy available to early organic synthesis, early evolving systems and early cells stand in the foreground, as do the possible mechanisms of their conversion into harnessable chemical energy for synthetic reactions. With regard to the possible temporal sequence of events, we focus on: (i) alkaline hydrothermal vents as the far-from-equilibrium setting, (ii) the Wood–Ljungdahl (acetyl-CoA) pathway as the route that could have underpinned carbon assimilation for these processes, (iii) biochemical divergence, within the naturally formed inorganic compartments at a hydrothermal mound, of geochemically confined replicating entities with a complexity below that of free-living prokaryotes, and (iv) acetogenesis and methanogenesis as the ancestral forms of carbon and energy metabolism in the first free-living ancestors of the eubacteria and archaebacteria, respectively. In terms of the main evolutionary transitions in early bioenergetic evolution, we focus on: (i) thioester-dependent substrate-level phosphorylations, (ii) harnessing of naturally existing proton gradients at the vent–ocean interface via the ATP synthase, (iii) harnessing of Na^+ gradients generated by H^+/Na^+ antiporters, (iv) flavin-based bifurcation-dependent gradient generation, and finally (v) quinone-based (and Q-cycle-dependent) proton gradient generation. Of those five transitions, the first four are posited to have taken place at the vent. Ultimately, all of these bioenergetic processes depend, even today, upon CO_2 reduction with low-potential ferredoxin (Fd), generated either chemosynthetically or photosynthetically, suggesting a reaction of the type 'reduced iron → reduced carbon' at the beginning of bioenergetic evolution.

1. Introduction

Life is a net exergonic chemical reaction, it releases energy to go forward. Many settings have been proposed as the site for the chemical synthesis for life's building blocks [1] and the ignition of the continuous chemical reaction that includes us as its descendants. Such proposed settings include oceans or ponds of organic soup stocked either by ultraviolet light-dependent organic synthesis [2] or by organics delivered from space [3]; borate evaporites [4]; terrestrial zinc-rich hydrothermal settings [5]; ice [6] or pumice [7]—to name but a few. However, since their discovery, submarine hydrothermal vents stand out among the possible environments for life's origin, holding particular promise for understanding the transition from geochemistry to biochemistry [8–12]. Submarine hydrothermal vents harbour highly reactive chemical environments with far-from-equilibrium conditions, being rich in gradients of redox, pH and temperature. Furthermore, hydrothermal vents generate spontaneously, forming systems of inorganic microcompartments [9,13] that, in an origin-of-life context, could readily serve to concentrate the chemical reaction products and substrates that

Royal Society Publishing
Informing the science of the future

form there, at their site of their synthesis. The far-from-equilibrium condition is important because it harbours the potential for exergonic chemical reactions [14]. Concentration is important as a prerequisite for autocatalysis and for self-regulation, defining characteristics of living systems [15].

2. Alkaline hydrothermal vents

A number of submarine hydrothermal vents have been studied. They differ in the specific temperature and the chemical compositions of their effluent fluid [10,16]. Most are located at the seafloor spreading zone, that is, directly above magma chambers. As such, their effluent comes directly into contact with magma and emerges at the vent–ocean interface with a temperature often exceeding 300°C. Such ‘black smokers’ are not only very hot, but they are also short-lived, lasting in the order of decades. In contrast, off-ridge vents of the kind exemplified by Lost City [13,17] are situated several kilometres away from the spreading zone; hence the water circulating through them makes no contact with magma and emerges at a temperature of around 70–90°C. Unlike vent systems that reside directly on the spreading zone, off-ridge vents can be stably active over geological time; Lost City has been active for about 120 000 years [18]. The chemical disequilibrium at vents arises from the contact between their reducing, H_2-containing effluent with CO_2-containing ocean water, which is more oxidized. Hydrothermal vents harbour up to 50 mM H_2 [10]. At Lost City, the H_2 concentration of the alkaline effluent (approx. pH 9–10) is about 10 mM [19].

Hydrogen is a reducing agent: an electron donor. Given a suitable acceptor, of which there are many [20], hydrogen is a source of energy. The H_2 in hydrothermal vents comes from a geochemical process called serpentinization [21–25]. Serpentinization takes place because most of the suboceanic crust consists of iron–magnesium silicates like olivine, in which the redox state of the iron is Fe^{2+}. Seawater penetrates the ocean floor crust, reaching depths on the order of 3–5 km below the surface. There at high pressure and temperatures of around 150°C, water oxidizes the Fe^{2+} to Fe^{3+}, the water being reduced to H_2 with the oxygen in water being retained in the rock as iron oxides. Bach *et al.* [21] write the serpentinization reaction as in equation (2.1) whereas Sleep *et al.* [22] represent the main redox reaction as in equation (2.2).

$Mg_{2.85}Fe_{0.15}Si_2O_5(OH)_4$ is serpentinite, the mineral from which the process is named, and $4Fe_3O_4$ is magnetite, the mineral containing the oxidized iron. During serpentinization, a cubic metre of olivine eventually yields about 500 mol of H_2 [24]. The heated water convects back up to the ocean floor, carrying H_2 and other reaction products with it. The production of H_2 via serpentinization has probably been taking place since there was liquid water on the Earth [22], and it continues today because, relative to the amount of water on the Earth, the amount of Fe^{2+} in the crust is inexhaustible. The $Mg(OH)_2$ in equation (2.1) produces alkaline hydrothermal fluid in agreement with the pH observed for Lost City effluent [17].

Lost City effluent also contains about 1 mM methane of abiogenic origin [19,26], along with about 100 μM formate of abiogenic origin [27] and traces of longer hydrocarbons up to pentane [19]. This indicates that serpentinization in the crust at Lost City reduces CO_2 to an organic product spectrum. The magma-hosted black smoker systems are less conducive to organic synthesis in the crust, because, above approximately 250°C, carbon in contact with water becomes CO_2 [28,29], and magma has a temperature of around 1200°C. If Lost City harbours organic synthesis today, then organic synthesis in similar serpentinizing systems should also have been possible 4 billion years ago. While the composition of ocean water has changed since then [30], the composition of the seafloor crust, where serpentinization reactions take place, has remained roughly constant, as indicated by the zircon compositions of Early (Hadean) mantle-derived melts [31].

The chemical disequilibrium at off-ridge hydrothermal vents on the early Earth was, in all likelihood, much greater than that today. This is because there was much more CO_2 in the atmosphere, perhaps up to 1000 times more [32,33], meaning that there would have been much higher CO_2 concentrations in the ocean. Hence when H_2 in the convective currents of serpentinizing systems made contact with marine CO_2 (stemming *inter alia* from volcanos) there was disequilibrium and potential for organic synthesis.

How much potential do vents harbour for the synthesis of what kind of organic products? Shock and co-workers have studied the question of organic synthesis at hydrothermal vents from the thermodynamic standpoint, and what they find is encouraging from an origin-of-life perspective (reviewed in [14]). They find that CO_2 reduction and organic synthesis is thermodynamically favoured—exergonic—when reduced vent fluid mixes with more oxidized ocean water at the vent–ocean interface, and hence can take place under the conditions presented by hydrothermal vents. This is true for the synthesis of carboxylic acids, alcohols and ketones [28,29], amino acids and proteins [34] and total microbial cell mass [35,36].

The nature and proportion of organic products that are thermodynamically most favoured depend upon the specific chemical conditions, for example, H_2 availability, temperature and the redox state of the environment [14,37,38]. Under the conditions found at Lost City, for example, the overall synthesis of microbial cell mass from H_2, CO_2 and NH_3 is exergonic in the temperature range 50–125°C [36].

$$\left.\begin{array}{c} 2Mg_{1.8}Fe_{0.2}SiO_4 + 3H_2O \rightarrow Mg_{2.85}Fe_{0.15}Si_2O_5(OH)_4 + Mg_{0.75}Fe_{0.25}(OH)_2 \\ \text{and} \\ 57Mg_{0.75}Fe_{0.25}(OH)_2 + 30SiO_2(aq) \rightarrow 15Mg_{2.85}Fe_{0.15}Si_2O_5(OH)_4 + 23H_2O + 4Fe_3O_4 + 4H_2, \end{array}\right\} \tag{2.1}$$

$$2Fe_3Si_2O_5(OH)_4 + 6Mg(OH)_2 \rightarrow 2Mg_3Si_2O_5(OH)_4 + 2Fe_3O_4 + 4H_2O + 2H_2. \tag{2.2}$$

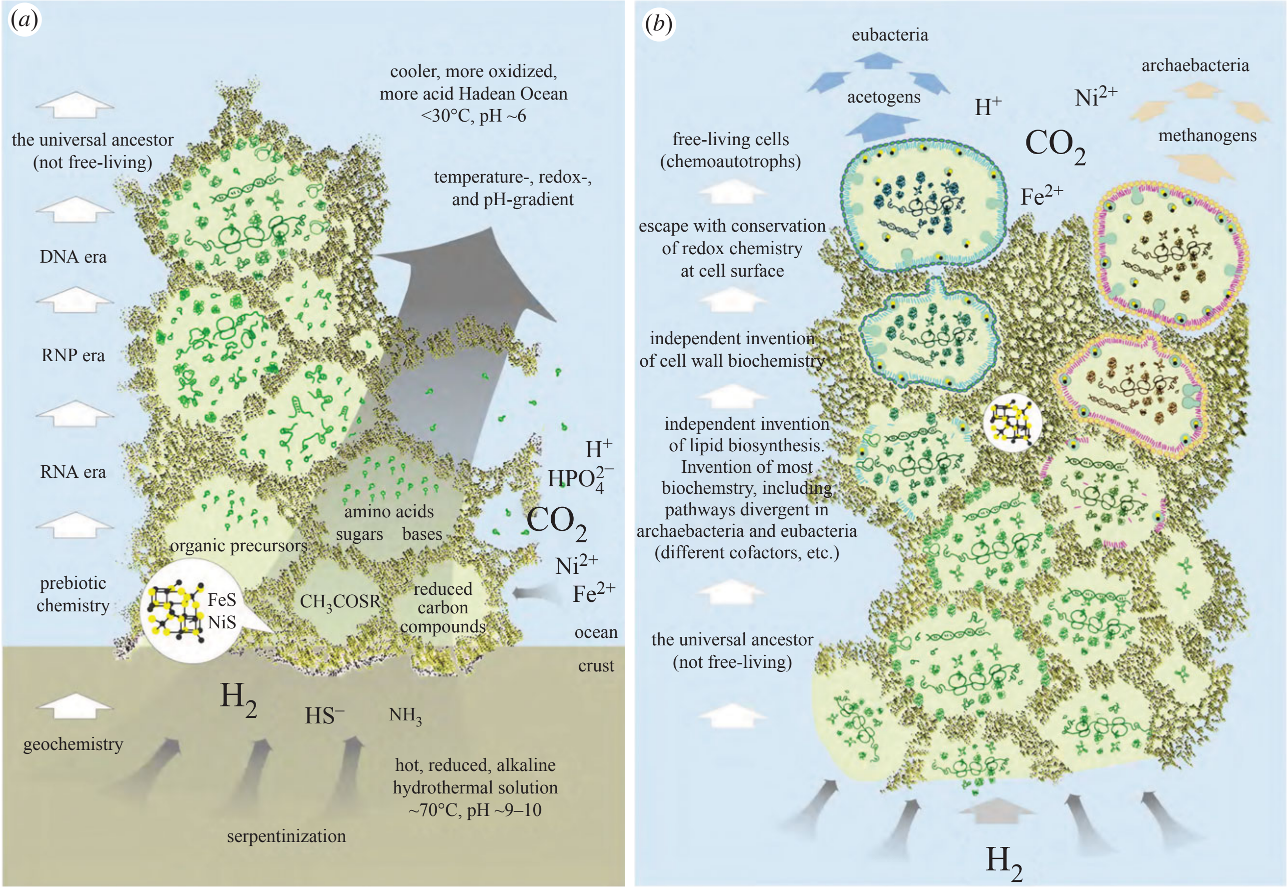

(a)
cooler, more oxidized, more acid Hadean Ocean <30°C, pH ~6
the universal ancestor (not free-living)
temperature-, redox-, and pH-gradient
DNA era
RNP era
RNA era
H+
HPO$_4^{2-}$
amino acids
sugars bases
CO$_2$
organic precursors
prebiotic chemistry
FeS
NiS
CH$_3$COSR
reduced carbon compounds
Ni2+
Fe2+
ocean
crust
H$_2$
HS−
NH$_3$
geochemistry
hot, reduced, alkaline hydrothermal solution ~70°C, pH ~9–10
serpentinization
(b)
eubacteria
acetogens
archaebacteria
H+
Ni2+
methanogens
free-living cells (chemoautotrophs)
CO$_2$
Fe2+
escape with conservation of redox chemistry at cell surface
independent invention of cell wall biochemistry
independent invention of lipid biosynthesis. Invention of most biochemstry, including pathways divergent in archaebacteria and eubacteria (different cofactors, etc.)
the universal ancestor (not free-living)
H$_2$

Figure 1. A scheme for the origin of cells [42,46].

Such findings are broadly consistent with the findings of Thauer *et al.* [39] that under conditions relevant for microbes, in the reaction of H_2 with CO_2, the equilibrium lies on the side of reduced carbon compounds, which is why acetogens can grow from the reaction

$$4H_2 + 2HCO_3^- + H^+ \rightarrow CH_3COO^- + 2H_2O \qquad (2.3)$$

with $\Delta G^{\circ\prime} = -104.6\ \text{kJ mol}^{-1}$ [39] and methanogens can grow from the reaction

$$4H_2 + CO_2 \rightarrow CH_4 + 2H_2O \qquad (2.4)$$

with $\Delta G^{\circ\prime} = -131\ \text{kJ mol}^{-1}$ [40] as their sole source of energy, respectively. They harness (conserve) chemical energy from those reactions to push the life process forward. For a prokaryote, the main energetic cost in the life process is amino acid and protein synthesis, which consumes about 75 per cent of the cell's ATP budget, with RNA monomer and polymer synthesis consuming only about 12 per cent [41].

Because hydrothermal vents present conditions where the synthesis of proteins from H_2, CO_2 and NH_3 would be an exergonic process [14,34], they truly appear to be special among the various settings that have been considered for the origin of life. Hydrothermal vents are particularly rich in chemical and thermodynamic similarities to the core energy releasing reactions of modern acetogens and methanogens [42,43], lineages that over 40 years ago—before the discovery of either hydrothermal vents or archaebacteria—were proposed to be the most ancient microbes, because they are anaerobic chemoautotrophs that live from the H_2–CO_2 redox couple [44].

The far-from-equilibrium conditions and favourable thermodynamic setting of hydrothermal vents do not indicate which chemical syntheses will occur, merely which are energetically possible [14]. The nature of catalysts present can also influence the kinds of products that are formed [45], depending on whether the reaction is thermodynamically controlled (the most stable products accumulate) or kinetically controlled (the most rapidly synthesized products accumulate). Here, we revisit a model for the origin of life (figure 1) as set forth previously in these pages while incorporating newer findings.

3. Getting from rocks and water to cells

One can assume that off-ridge vents such as Lost City were more prevalent on the early Earth than today [11,22,23]. But the chemistry at a Hadean vent–ocean interface would have been slightly different, with abundant Fe^{2+} [30] and far more CO_2 [32,33] in the ocean. Given that, and given the presence of sufficient sulfide in the hydrothermal fluid to steadily precipitate transition metal sulfides (FeS, NiS and the like) upon mixing of effluent and ocean water at the vent interface [9], the model posits that hydrothermal flux deposited a mound at the vent, consisting crucially of transition metal sulfides, among other possible minerals. The premise that a Hadean mound could have existed is underpinned by the report of fossilized metal sulfide hydrothermal mounds 360 Myr of age—the rocks that gave rise to this model in the first place [47]—found in Tynach, Ireland by Russell and co-workers [48]. These mounds reveal elaborate systems of naturally forming, inorganically walled microcompartments (on the order of 1 µm–1 mm in diameter). The walls of these compartments unite two very important properties for early chemical evolution: catalysis and compartmentation.

The catalysis was provided by the transition metals and transition metal sulfides themselves of the hydrothermal mound. That is, the microcompartments have inorganic walls made of essentially the same catalysts that are commonly found in redox-active enzymes of modern microbes [9,49–51] and commonly used in organometallic catalysis in the chemical industry [52,53]. This is an important similarity to modern anaerobic autotrophs, whose enzymes of core carbon and energy pathways are replete with FeS, FeNiS and MoS (WS) centres, FeS and NiS usually being incorporated into proteins via cysteine residues [54–58], Mo and W usually being complexed by thiols of the molybdopterin cofactor MoCo [59]. Importantly, such FeS and FeNiS centres are particularly common in H_2-oxidizing and CO_2-reducing proteins of anaerobic prokaryotes [60]. Because the nature of the catalysts at the interface of gases (H_2, CO_2, N_2, H_2S, SO_2) and organic matter that (i) we posit to be present at the mound and (ii) occur in acetogens and methanogens is very similar, some congruence between the kinds of products obtained in the two systems is not an unreasonable proposition [45].

Interfacing with the CO_2-rich seawater at the vent was highly reduced, H_2-containing, vent effluent that probably contained some reduced carbon compounds itself, as Lost City does today [19,26,27], stemming from serpentinization reactions deep within the crust. How reducing might the effluent in an ancient serpentinizing system have been? A clue is provided by the mineral awaruite (Ni_3Fe), an alloy of metallic iron and nickel, $Fe^{\pm 0}$ and $Ni^{\pm 0}$, that is quite commonly found in serpentinizing systems and that is produced from Fe^{2+} and Ni^{2+} minerals via reduction to the raw metals *in situ* by H_2 [25,61]. Awaruite forms only when sulfide concentrations are very low, but that is not the point, nor is the point that $Fe^{\pm 0}$ readily reduces CO_2 to acetate and formate under hydrothermal conditions [62]. The point is that in the temperature range 150–300°C, activities of H_2 exceeding 200 mM are needed for awaruite to form [25]. To stress the point, 200 mM H_2 represents a very large amount of hydrogen and a very reducing environment. The presence of awaruite in serpentinizing systems thus clearly indicates that at some point during their lifespan, serpentinizing systems regularly generate extremely reducing conditions on the order of greater than or equal to 200 mM H_2, which would be conducive to organic synthesis, as Klein & Bach [25] point out. So while conditions at a Hadean-serpentinizing system cannot be pinpointed, we can say that serpentinizing systems do, in principle, have the reducing power at hand that would be needed for making organic compounds out of CO_2 (or even elemental iron and nickel out of their salts).

For the building blocks of life to form, nitrogen is also essential. Where does reduced nitrogen come from? Some modern vent systems contain up to 1 mM NH_3 [10], although it is not known that serpentinization is responsible for N_2 reduction. However, simulated hydrothermal conditions of 300°C and high pressure readily convert nitrite into NH_3 [63], and under very mild conditions, FeS minerals can reduce N_2 to NH_3 at yields approaching 0.1 per cent [64], whereby native $Fe^{\pm 0}$ is also an effective catalyst of N_2 reduction under hydrothermal conditions [65], recalling once again awaruite as an indicator of serpentinization [25]. So it should be safe to assume that reduced nitrogen would have been available in a Hadean-serpentinizing hydrothermal system. Phosphate is more soluble at pH 6 than at pH 8 and was not likely to have been globally abundant in the early oceans [11,30], although there was surely considerable local phosphate input from volcanos [66,67]. Phosphate was possibly concentrated at hydrothermal vents, however, via the mineral brucite—$Mg(OH)_2$—driven precipitation processes that Holm *et al.* [68] suggest to have involved sepentinization.

The foregoing would provide a mixture of H_2, CO_2 and NH_3 similar to that assumed by Amend *et al.* [14] for synthesis of chemical constituents of cells, in an environment laden with transition metal catalysts, with abundant hydrothermal HS^-. That is a starting point for the synthesis of compounds with high-energy bonds. Under mild laboratory conditions, Heinen & Lauwers [69] reduced CO_2 to a spectrum of sulfur containing organic products, including methylsulfide (CH_3SH), using FeS as the catalyst, and Huber & Wächtershäuser [70] reacted CH_3SH and CO under mild conditions (ambient pressure, 100°C), using FeS, NiS or Ni^{2+} as catalysts, obtaining up to 40 mol% of the thioester methylthioacetate (or acetyl methylsufide, CH_3COSCH_3). Thioesters are energy-rich compounds because the thioester bond has a large free energy of hydrolysis, $\Delta G^{\circ\prime} = -32\ \mathrm{kJ\ mol^{-1}}$ [71]. Starting from such simple chemicals, a high spontaneous yield of a compound with an energy-rich thioester bond might seem surprising, but the result is expected from thermodynamics because even to the level of the energy-rich thioester, the reaction

$$2CO_2 + 4H_2 + CoASH \rightarrow CH_3COSCoA + 3H_2O \qquad (3.1)$$

is highly exergonic with $\Delta G^{\circ\prime} = -59\ \mathrm{kJ\ mol^{-1}}$ [72], whereby the pathway becomes endergonic at low H_2 partial pressures, with $\Delta G^{\circ\prime} = +29\ \mathrm{kJ\ mol^{-1}}$ at approximately 10 Pa H_2 [73]. Equation (3.1) entails the thiol group of coenzyme A (CoASH) rather than CH_3SH, and it is the line reaction of the acetyl-CoA or Wood–Ljungdahl [74,75] pathway of microbial CO_2 fixation. Such favourable energetics led Shock *et al.* [29, p. 73] to conclude that organisms using the acetyl-CoA pathway (acetogens, methanogens and many sulfate reducers) get 'a free lunch that they are paid to eat'. Provided that transition metals in a hydrothermal vent could catalyse similar reactions as in the laboratory [51], then one would have, in principle, a route for sustained geochemical synthesis of acetyl thioesters (among many other products). That would constitute a critical/crucial link between geochemistry and biochemistry, because acetyl thioesters (acetyl-CoA) are the most central compounds in all of metabolism, as biochemical pathway maps will attest; and apart from the Calvin cycle, which is obviously a late evolutionary invention [76], acetyl-CoA is the direct product of all known CO_2 fixation pathways [73]. The centrality of acetyl thioesters in modern metabolism is, we contend, a relic of their role in the chemical events that gave rise to metabolism.

Acetyl thioesters are furthermore the most probable entry point of phosphate into metabolism [42,77,78]. In a plethora of microbes, acetyl-CoA is enzymatically cleaved via phosphorolysis to yield acetyl phosphate [79,80], which, with its high free energy of hydrolysis ($\Delta G^{\circ\prime}$) of $-43\ \mathrm{kJ\ mol^{-1}}$, can phosphorylate any number of substrates, including ADP to generate ATP, with a free energy of hydrolysis ($\Delta G^{\circ\prime}$) of $-31\ \mathrm{kJ\ mol^{-1}}$. Weber [81] reported the non-enzymatic synthesis of acyl phosphates from a thioester in the presence of phosphate during the synthesis of pyrophosphate from thioesters, although the acyl phosphate itself was not isolated. A high free energy of hydrolysis of the acyl phosphate bond and simple phosphorylytic synthesis from thoiesters make acyl phosphates good candidates for the first universal currency of high-energy bonds [42], the precursors of ATP.

4. Antiquity of the acetyl-CoA pathway

Provided there was a sustained geochemical source of chemically accessible methyl groups (e.g. methyl sufide), the first reactions relevant to the origin of carbon and energy metabolism might have been very similar to those catalysed by the bifunctional enzyme carbon monoxide dehydrogenase/acetyl-CoA synthase (CODH/ACS; figure 2*b*). The catalysis in CODH/ACS is performed by transition metals (Fe and Ni) coordinated as sulfide clusters [50,56,73]. In the enzymatic mechanism [56], electrons from a low-potential Fd are used by the enzyme to reduce one molecule of CO_2 to CO at an FeNiS cluster called the C cluster; that CO migrates to a second FeNiS cluster in the enzyme (the A cluster), where it binds one of two Ni atoms at the Ni_p site to form a Ni-bound carbonyl group, where the two-electron donating property of Ni is important in the catalytic mechanism [56]. A methyl group, donated by the corrinoid iron–sulfur protein CoFeSP, binds the same Ni atom [56] in an unusual metal-to-metal methyl transfer reaction [88], and the carbonyl group inserts the Ni–CH_3 bond to generate a Ni-bound acetyl group, which is then removed from the enzyme via thiolysis by CoASH to generate the thioester. Given the methyl group, this route of thioester synthesis involves only metals as catalysts and electron donors, no phosphate is involved, and given CO, the reaction works *in vitro*, without enzymes, using either Fe^{2+} or Ni^{2+} as catalysts [70].

CODH/ACS thus has hallmarks of a very ancient reaction, one that could well have proceeded readily before the advent of genes and enzymes. Across the divide that separates acetogens and methanogens, this primordial chemistry at CODH/ACS is homologous and conserved, as is the CoFeS protein [83], but the synthesis of the methyl group is fundamentally different [73], involving distinct, unrelated enzymes and different cofactors—tetrahydrofolate (H_4F) and tetrahydromethanopterin (H_4MPT)—both of which are, however, pterins (figure 2*d*), and chemically more similar than the different names suggest [85].

In acetogens, methyl synthesis entails reduction of CO_2 to formate via an NAD(P)H-dependent formate dehydrogenase, an ATP-consuming step at the 10-formyl-H_4F synthetase reaction, water elimination via 5,10-methenyl-H_4F cyclohydrolase, reduction with NADH by 5,10-methylene-H_4F dehydrogenase, and Fd-dependent reduction via 5,10-methylene-H_4F reductase to yield 5-methyl-H_4F which donates the methyl group to corrinoid iron–sulfur protein (CoFeSP) [73]. Of energetic consequence, the formation of the formyl pterin formyl-H_4F is endergonic, requiring ATP hydrolysis, while the 5,10-methylene-H_4F reductase step is highly exergonic (-57 kJ mol^{-1}) [73] and is suspected to be a site of energy conservation in some acetogen species [89].

In methanogens, the synthesis of the methyl group involves different enzymes and to some extent different cofactors [83]. The first step is the formation of the carbamate formyl-methanofuran (formyl-MF) from CO_2 and the primary amine of methanofuran. Notably, this reaction is spontaneous, requiring only the cofactor and the substrate, with no help from enzymes [90]. Fd-dependent formyl-MF dehydrogenase and a formyl tranferase yield the formyl pterin formyl-H_4MPT. A cyclohydrolase, F_{420}-dependent 5,10-methylene-H_4MPT dehydrogenase and $F_{420}H_2$-dependent 5,10-methylene-H_4MPT reductase yield methyl-H_4MPT which donates the methyl group to CoFeSP in carbon metabolism, involving CODH/ACS. In energy metabolism, methyl-H_4MPT donates its methyl group to coenzyme-M (CoM).

Thus, the reactions at CODH/ACS are catalysed by homologous enzymes, and the methyl-carrying corrinoid protein CoFeSP is homologous in acetogens and methanogens. But in the methyl synthesis branch of the two groups, the only homologous enzymes are at the CO_2-reduction step. The MoCo-binding subunits of Mo-dependent formyl-methanofuran dehydrogenase, FmdB, and of the W-dependent enzyme, FwdB, in methanogens are related to formate dehydrogenase in acetogens [91], which also exist as Mo- or W-dependent enzymes [92]. The other enzymes are not related, use different cofactors, and in some cases catalyse different reactions. Thus, while it is generally agreed that the acetyl-CoA pathway is the most ancient among modern CO_2-fixing pathways, the methyl synthesis branches of the Wood–Ljungdahl pathway in acetogens and methanogens are unrelated, even though the methyl groups serve a homologous function at a homologous enzyme, CODH/ACS, en route to thioester synthesis. This suggests three things: (i) the gene for CODH/ACS was present and functioning in their non-free-living common ancestor, (ii) the genes and proteins for the methyl synthesis pathways of acetogens and methanogens arose independently and subsequently to divergence between the archaebacterial and eubacterial stem lineages within the hydrothermal mound, and (iii) that a continuous and abundant supply of accessible methyl groups stemming from geochemical processes—either serpentinization or geochemical redox processes occurring directly at the vent–ocean interface [43]—served as the source of biochemical methyl moieties prior to the advent of the genes and proteins that underpin the methyl synthesis branch of the Wood–Ljungdahl pathway in acetogens and methanogens. This is another way of saying that when the enzymes of methyl synthesis arose, they did not invent methyl synthesis, they just helped chemical reactions that tended to occur anyway, to occur more rapidly.

In this view, the thermodynamically favoured reaction of H_2 and CO_2 to methyl groups and methane, similar to that observed in the laboratory [93] or at Lost City [19], was taking place spontaneously (figure 2*a*), and the genes and proteins (figure 2*b*) just helped speed it along and add specificity to a highly selectable function (accelerated self-synthesis, given the necessity of methyl groups for self-synthesis). However, those genes arose in independent stem lineages, from which it follows that methyl groups, whether formed by serpentinization or in the walls of the mound, were abundant at the mound over geological periods of time.

If the reader will grant us, for the purpose of argument, a sustained source of acetyl thioesters at a hydrothermal vent setting, the spontaneous and continued accumulation of organic compounds, including peptides, would be thermodynamically possible. A reaction of this type is what Morowitz [45,94] has, rightly, always demanded. Thioesters could lead to some simple carbon chemistry as outlined in fig. 6 of Fuchs [73], some transition metal catalysed reductive aminations of 2-oxo acids as outlined by Huber & Wächtershäuser [95], and some peptide synthesis aided by transition metals and simple reactive compounds such as carbonyl sulfide [96,97].

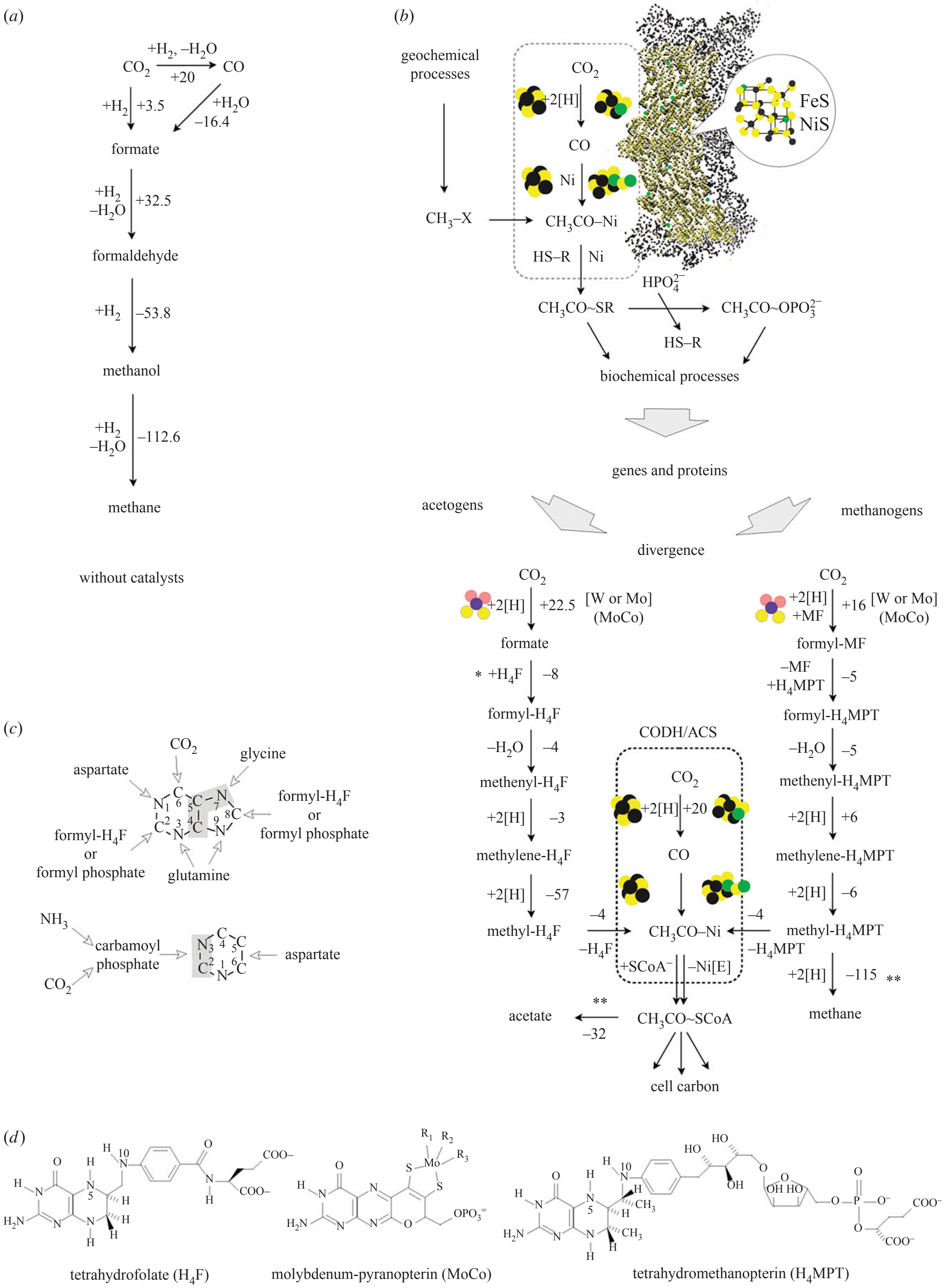


Figure 2. (*Caption opposite.*)

Keeping in mind that the pyruvate-synthesizing enzyme of acetogens and methanogens, pyruvate synthase, is replete with FeS centres and uses acetyl-CoA, reduced Fd (Fd_{red}) and CO_2 as substrates [60], and also that reducing conditions of hydrothermal vents make energetics work in favour of amino acid and peptide synthesis [14], an alkaline hydrothermal vent would, in principle, be working in the right direction thermodynamically, because cells consist mainly

Figure 2. (*Opposite.*) Illustration of some concepts relevant to this paper. (*a*) Abiotic methane production. Summary representation of the H_2-dependent conversions of CO_2 to methane without catalysts, adapted from [82]. Numbers next to arrows indicate the approximate change in free energy, $\Delta G^{\circ\prime}$, for the step indicated at physiological conditions (25°C and pH 7) in kJ mol^{-1}, conditions which do not generally exist in geochemical environments. Note, however, that Lost City does produce methane of abiotic origin [20]. The thermodynamic values are taken from Maden [83] and Rother & Metcalf [84]. Regarding the equilibria between the different C_1 species in the absence of catalysts, see Seewald *et al.* [82]. (*b*) Scheme suggesting homology between the acetyl-CoA pathway in modern acetogens and methanogens to geochemical processes in the hydrothermal vent of a serpentinizing system in the crust of the ancient Earth. See text. Numbers next to arrows indicate the approximate change in free energy, $\Delta G^{\circ\prime}$, for the step indicated at physiological conditions (25°C and pH 7) in kJ mol^{-1} as reported in [73]. An asterisk indicates ATP investment, a double asterisk indicates ATP return. Note that net ATP return in acetogens and methanogens requires chemisosmotic coupling. FeS and FeNiS clusters are symbolized. Fuchs [73] gives the free energy change for the CODH/ACS reaction as $\Delta G^{\circ\prime} = 0$ kJ mol^{-1}. The thermodynamic value for the methane-producing step is from [85]. H_4F, tetrahydrofolate; MF, methanofuran; H_4MPT, tetrahydromethanopterin; Ni(E), an Fe–Ni–S cluster in CODH/ACS; HSCoA, coenzyme A. For the acetyl-CoA pathway, see also Bender *et al.* [60], Ragsdale [56] and Fuchs [73]. The formate to formyl-H_4F conversion in acetogens entails ATP hydrolysis (not shown), which lowers $\Delta G^{\circ\prime}$ for the reaction to -10 kJ mol^{-1}. Though all reactions shown are reversible, arrows are shown in one direction only for convenience. (*c*) The source of carbon and nitrogen atoms in the purine and pyrimidine backbone. Data from Stryer [86] and from Ownby *et al.* [87]. (*d*) Pterin cofactors involved in the WL-pathway: MoCo [59] folate and methanopterin [85].

(approx. 60% dry weight) of protein and to a lesser extent (approx. 25% dry weight) of nucleic acids.

5. RNA and the code arose, but that is not our focus

Our focus here is not the primordial synthesis of bases, nor how bases or nucleosides might react with one another, nor what an RNA world might have looked similar in detail, nor the origin of the genetic code. We are interested in the chemical environment and the energetic processes that might have made those important evolutionary steps possible, as exergonic reactions or, more likely, as side products of a central exergonic reaction involving H_2–CO_2 disequilibrium (similar to the case in modern microbial metabolism). Three points can be noted.

First, DNA is synthesized from RNA in metabolism, which has long served as an argument that RNA came before DNA in evolution. By the same reasoning, we infer that amino acids came before RNA in evolution, because in metabolism RNA bases are synthesized from amino acids: glycine, aspartate, glutamine, CO_2 and some C_1 units in the case of purines; aspartate, ammonia and CO_2 via carbamoyl phosphate in the case of pyrimidines (figure 2*c*). Amino acids are thermodynamically favoured over nucleotides and thus more likely to accumulate in larger amounts in a hydrothermal setting anyway [35,36]; similarly in metabolism, some might spill over into RNA-like bases. If an RNA world-like system arose in the same environment where the acetyl-CoA pathway arose—a logical consequence of the idea that the acetyl-CoA pathway is the most ancient pathway of carbon fixation [73]—then the RNA world arose in a chemical environment that was very rich in CO_2, reactive C_1 moieties and reactive methyl groups. Two of the five carbon atoms in the purine backbone come from reactive C_1 moieties: 10-formyl-H_4F or formyl phosphate [87], and one of the five comes from CO_2, while one of the four carbon atoms of the pyrimidine backbone comes from CO_2 via carbamoylphosphate synthase reaction (figure 2*c*).

This might be a biosynthetic fossil, an imprint of the chemical environment in which the genes arose that catalysed the first, genetically specified, enzymatic base biosynthesis. That is, the substrates of the enzymes that catalyse base synthesis were probably in existence before the origin of those genes, and the four carbon atoms that stem from CO_2 or formyl moieties in all modern purine and pyrimidine bases (figure 2*c*) might be relics of life's autotrophic origin. There are limits as to how far one can go with such reasoning, but it is something to consider.

Second, the RNA world as it is currently construed in the laboratory, operates with pure mixtures of usually four bases [98]. But, in the beginning there is just no way that base synthesis could have been highly specific. Any RNA world that might have really existed must have consisted of an ill-defined mixture of bases synthesized spontaneously without the help of either genes or pure chemicals. The chemical environment central to our considerations here is replete with reactive C_1 moieties and methyl groups (on their way to becoming methane, acetyl groups and acetate); modern ribosome–tRNA interactions at the heart of the genetic code require myriad base modifications in tRNA and rRNA [99], the vast majority of which are methylations [100]. This possible significance of tRNA modifications [101] becomes all the more evident when those common to all prokaryotes that allow the code to function at the wobble base [102] are considered. Or, we can just look at the nature of the chemical modifications that are present in the 28 modified bases common to archaebacteria and eubacteria [100]. They include 19 methyl-, one acetyl-, five thio-, two methylthio-, one seleno-, two methylaminomethyl- and four carbamoyl-moieties [103]. Those modifications might be a chemical imprint—a molecular fossil—of the environment in which tRNA–rRNA interactions (genes and proteins) arose that is preserved in the chemical structure of modified tRNA and rRNA bases. Assuming that any RNA world-like reaction nexus ever really existed in early evolution (a reasonable premise), it would by necessity have consisted of spontaneously and unspecifically synthesized mixtures of bases. The substantial number of enzymes that life forms have carried around for almost 4 billion years to perform these cumbersome and specific modifications, suggests that the modifications are essential, otherwise they would have been discarded. That those enzymes recreate the ancestral state of the RNA world is something to consider.

Third and finally, there is a rather severe and very general problem for all theories on the origin of life. Namely, before there was any kind of genetic feedback via replicating molecules, whether self-replicating or replicated by auxiliary catalysts, we currently have no option but to accept the premise that some replicating chemical entity, catalytic and able to specify synthesis of self, arose spontaneously, which is theoretically possible in the context of autocatalytic networks [104–106]. Morowitz has observed that the chemistry of life is deterministic in many ways, but that does not mean that its emergence is inevitable, because a number of contingent factors, such as the nature of catalysts present, are involved

> The origin of life is a deterministic event, the result of the operation of the laws of nature of a physical chemical system of a certain type. This system evolves in time, is governed by physical principles and eventually gives rise to living forms. The details need not be totally deterministic in every respect, but the overall behaviour follows in a particular way. [94, p. 3]

In a world of gene and proteins, evolution via variation and selection can readily supplant chemical determinism. But until we get to things that evolve via variation and selection, all we have are spontaneous reactions under thermodynamic and kinetic control. In this way, thermodynamics is very much chemistry's deterministic version of natural selection [107]. Getting from rocks and water to genuinely evolving systems remains a big problem. But energetics and thermodynamics can help a lot, because they place rather narrow contraints on which paths chemistry can traverse en route in evolving systems that ultimately spawn free-living cells.

That bottom-up perspective has a top-down corollary [108]. If we search among modern microbes and their metabolism for ancient physiological phenotypes, the acetogens and methanogens stand out [44,109]. These obtain their energy from making organic compounds out of CO_2. In fact, they are often diazotrophic and as such they just live from gases: CO_2, CO, H_2, N_2 and H_2S. At the level of overall chemical similarity, core carbon and energy metabolism in methanogens and acetogens has quite a lot in common with geochemical processes at alkaline hydrothermal vents [43], and that is probably not coincidence.

Again, we have little to contribute to the nature of an RNA-like world [98], other than the insights that it probably arose in the same environment where the acetyl-CoA pathway did, a setting full of amino acids, reactive C_1 moieties and CO_2, and that molecular fossils of that environment might be preserved in RNA base modifications today. Nor do we have anything to contribute to the origin of the genetic code and translation, other than agreeing with our colleagues that it is a very significant and difficult problem [110–112], that it did occur, that the code increased in complexity during evolution [113,114], that the simplicity of the code in methanogens is intriguing [115] and that the origin of the code could have occurred at alkaline hydrothermal vents [116], supported by a continuous flux of harnessable carbon and harnessable energy.

(a) Chemiosmotic coupling

It should be explicitly stated that, for the purposes of this paper, we are assuming that the origin of genes and proteins could have been fuelled with acyl phosphates (acetyl phosphate) as the central currency of 'high-energy bonds', that these were generated by substrate-level phosphorylation via reactions as outlined earlier (figure 2*b*), and that this requires a steady input of spontaneously synthesized methyl groups of geochemical origin [109], either from serpentinization or from redox processes at the vent–ocean interface [43]. The standard free energy of hydrolysis ($\Delta G^{\circ\prime}$) of acetyl phosphate is $-43\ \mathrm{kJ\ mol^{-1}}$, meaning it can phosphorylate ADP to ATP ($\Delta G^{\circ\prime}\ -31\ \mathrm{kJ\ mol^{-1}}$), and by the same token it can drive phosphorylations in intermediary metabolism as well as ATP does. However, the actual $\Delta G^{\circ\prime}$ provided by a reaction depends on the level of disequilibrium between rates of formation and breakdown—specifically the steady-state ratio of acetyl thioesters to acetyl phosphate to acetate, which is ultimately driven far from equilibrium by the chemical disequilibrium generated by serpentinization.

Once increasing chemical complexity transforms into conventional evolution of genetically encoded proteins, the invention of new functions potentially becomes a very fast process. The jump we just took, to genes and proteins, is a big one, but all models for the origin of life entail such a step, and ours does too. Recalling that its mechanisms are, as stated earlier, not our focus, we proceed. At some early point, the advent of ATP as the universal energy currency was an important step in bioenergetic evolution, displacing (we posit) acetyl phosphate. However, while ATP is universal across lineages, it is not the sole energy currency within the metabolism of individual cells by any means [80]. This is an interesting and possibly significant point, suggesting that much went on in early biochemistry before ATP became a common currency. The simplest explanation for ATP's rise to prominence is that it was a consequence of the substrate specificity of the rotor–stator-type ATPase, a protein that is as universal among cells as the code, and that is unquestionably an invention of the world of genes and proteins [42]. Given genes and proteins, the origin of molecular machines such as the ATPase is an admittedly impressive, but not conceptually challenging, evolutionary step; it is a far less problematic increment than the hurdles that had to be surmounted at the origin of the ribosome and the code [110,117].

In an alkaline hydrothermal environment, the ATPase has immediate beneficial function. The naturally preexisting proton gradient [9] at the interface of approximately pH 10 alkaline effluent (alkaline from serpentinization) with approximately pH 6 ocean water (slightly acidic by virtue of high CO_2 content) provides a geochemically generated chemiosmotic potential that 'just' needs to be tapped, that is, converted into chemically harnessable energy in the form of high-energy bonds, exactly what rotor–stator-type ATPases do. That would have put a very high-energy charge on the contents of the inorganic compartments within which we presume that this was all taking place, accelerating biochemical innovations, energetically financing gene inventions, and the selective pressure on evolving proteins to adapt to a new energy currency is evident. ATP-binding domains are so prevalent in genomes [118], not because ATP is a constituent of RNA, but because it became the most popular energy currency. The ATPase transduced a geochemically generated ion gradient into usable chemical energy, and since the energy was free, the means to harness it as ATP 'just' required a suitable protein for the job, a complicated protein [119], but a protein.

One might object that the multi-subunit rotor–stator-type ATPase is too complicated as a starting point, and that it is hence preferable to assume that some simpler form of energy transducing protein, for example, a pyrophosphatase, preceeded it [120]. But the observation from modern microbes to be accounted for, in the evolutionary sense, is that all prokaryotes harbour the rotor–stator-type ATPase. Similar to a handful of only 30 or so other proteins, it is as universal as the ribosome; hence it was probably present in the common ancestor of all cells, which in this model was not free-living but rather was confined to the system of inorganically formed compartments within which it arose. Consistent with that, prokaryotic ATPases show a clear dichotomy into two types: the eubacterial (or F-type) ATPase and the archaebacterial (or A-type) ATPase [121]. That dichotomy is also mirrored by a number of other cellular, molecular and gene homology attributes, indicating that the deepest, most ancient split in the prokaryotic world is that separating eubacteria from archaebacteria [111,116,122].

That suggests that the last universal common ancestor (LUCA), which we posit to be a geochemically confined ancestor (figure 1), possessed among other things genes, proteins, the code, along with biosynthetic pathways for amino acids, bases and cofactors to support functional ribosomes and an ATPase that could tap the naturally chemiosmotic gradient at the vent–ocean interface. But it was far short of being a free-living cell. Many attributes of archaebacteria and eubacteria are so fundamentally different that, for lack of similar chemical intermediates in the pathway or for lack of subunit composition or sequence similarity, independent origin of the genes underlying those differences is the simplest explanation. Such differences include: (i) their membrane lipids (isoprene ethers versus fatty acid esters) [123], (ii) their cell walls (peptidoglycan versus S-layer) [124], (iii) their DNA maintenance machineries [116,125], (iv) the 31 ribosomal proteins that are present in archaebacteria but missing in eubacteria [126,127] (v) small nucleolar RNAs (homologues found in archaebacteria but not eubacteria) [128], (vi) archaebacterial versus eubacterial-type flagellae [129], (vii) their pathways for haem biosynthesis [130,131], (viii) eubacterial- versus archaebacterial-specific steps in the shikimate pathway [132,133], (ix) a eubacterial-type methylerythrol phosphate isoprene pathway versus an archaebacterial-type mevanolate isoprene pathway [134], (x) a eubacterial-type fructose-1,6-bisphosphate aldolase and bisphosphatase system versus the archaebacterial bifunctional aldolase-bisphosphatase [135], (xi) the typical eubacterial Embden–Meyerhoff (EM) and Entner–Doudoroff (ED) pathways of central carbohydrate metabolism versus the modified EM and ED pathways of archaebacteria [136], (xii) differences in cysteine biosynthesis [137], (xiii) different unrelated enzymes initiating riboflavin (and F_{420}) biosynthesis [138], and (xiv) in very good agreement with figure 2*b*, different, unrelated, independently evolved enzymes in core pterin biosynthesis [139], to name a few examples. The pterin biosynthesis example is relevant because the cofactors H_4F, H_4MPT and MoCo, which are central to the eubacterial and archaebacterial manifestations of the Wood–Ljungdahl pathway are pterins (figure 2*d*), suggesting that methyl synthesis occurred geochemically (non-enzymatically) for a prolonged period of biochemical evolution.

It has been argued that these deep differences between archaebacteria and eubacteria reflect comparatively recent environmental adaptations; a specific claim is that archaebacteria adapted very recently [140]. However, many eubacteria thrive in hyperthermophilic environments without going to the trouble of replacing all their membranes and walls, DNA replication, etc., while plenty of archaebacteria thrive in mesophilic environments such as the open oceans and soils. It is therefore not the case that archaebacteria are somehow better adapted to hyperthermophilic conditions whereas bacteria are better adapted to mesophilic conditions, even though the perception of archaebacteria as extremophiles persists. Alternatively, the differences of archaebacteria and eubacteria have been interpreted as adaptations to energy stress [141], but both acetogens and methanogens live at the lower end of the free energy spectrum whereby acetogens live from less energy than methanogens [142,143] but without having become archaebacteria, so the energy stress argument is clearly not robust. Given the almost equally wide environmental representation of both groups, both in energy-rich and in energy-poor environments [144], the deep divergence between them in all likelihood simply reflects the very early evolutionary divergence, in our view physical divergence, of replicating compartment contents within the mound that led to two stem lineages which became the common ancestors of eubacteria and archaebacteria, respectively (figure 1). Physiologically, those stem lineages gave rise to acetogens and methanogens. The similarities and differences in carbon metabolism (the Wood–Ljungdahl pathway) in acetogens and methanogens outlined earlier (figure 2) are paralleled by similarities and differences in their mechanisms of energy conservation (figure 3), which entail chemiosmotic coupling.

At first sight, the idea that chemiosmosis is a very ancient means of energy transduction might seem counterintuitive. More familiar to many is the old (and popular) doctrine that the most ancient pathway of energy metabolism is a fermentation such as glycolysis [77], an idea that goes back at least to Haldane [2] and hence arose long before anyone had a clue that biological energy can be harnessed in a manner that does not involve substrate-level phosphorylations and 'high-energy' bonds [149,150]. In modern life, all biological energy in the form of ATP comes ultimately from chemiosmotic coupling [151], the process of charge separation from the inside of the cell to the outside, and the harnessing of that electrochemical gradient via a coupling factor, an ATPase of the rotor–stator-type. It was not until the 1970s that it became generally apparent that Mitchell [152] was right, his Nobel prize coming in 1978, and it is hard to say when it became clear to microbiologists that all fermentative organisms are derived from chemiosmotic ancestors. We also note that Mitchell's consideration of the problem of the origin of life introduced key concepts of his later chemiosmotic hypothesis, including a definition of life as process, and the idea of vectorial catalysis across a membrane boundary that is inseparable either from the environment or from the organism itself [153].

The maxim that glycolysis is ancient might be an artefact of experience, since it was the first pathway both to be discovered and that we learned in college; in that sense, it really is the oldest. When one suggests that chemiosmotic coupling in methanogens or acetogens might be ancient, many listeners and readers shy away, mainly because the pathways are unfamiliar and often entail dreaded cofactor names. The familiar cytochrome complexes, quinones and quinone-based charge translocation are lacking in these apparently ancient life forms, which possess no complexes I, II or III, no terminal oxidases, and only a single coupling (ion pumping) site each. Their energy metabolic pathways differ so fundamentally from the familiar respiratory chain-type chemiosmotic pathways of *E. coli* or mitochondria that only recently have biochemists uncovered how their basic energetics works. Yet, despite these fundamental differences, and their unfamiliarity to most of us, the bioenergetics of methanogens and acetogens nonetheless feature bona fide chemiosmotic circuits, with coupling proteins, ion electrochemical gradients over membranes and the ATP synthase. These pathways are illustrated in figure 3, which recapitulates the schemes presented by Buckel & Thauer [143], Thauer *et al.* [40] and Poehlein *et al.* [146], for methanogens and acetogens that

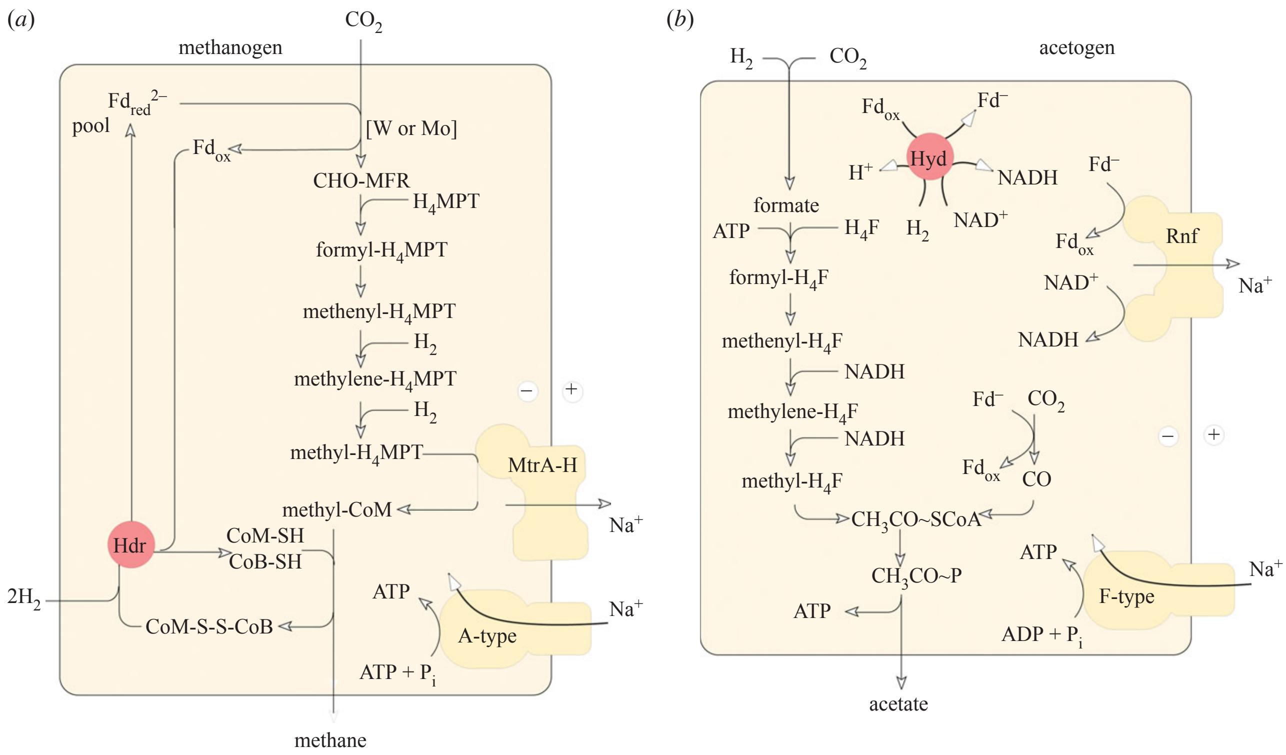


Figure 3. Energy metabolism of (*a*) methanogens and (*b*) acetogens without cytochromes. Redrawn for *Methanothermobacter marburgensis* from Thauer *et al*. [40], Kaster *et al*. [145] and Thauer & Buckel [143], redrawn for *Acetobacterium woodii* from Pohlein *et al*. [146] and Thauer & Buckel [143]. When we refer to acetogens and methanogens that lack cytochromes, we are referring to the physiology in those organisms, as the examples. The use of the symbol Fd^{2-} indicates that the ferredoxin in question has two FeS centres, both of which become reduced. MtrA-H, methyl transferase complex [40]. Rnf: Fd:NADH oxidoreductase, originally named for *Rhodobacter* nitrogen fixation [147]. Other abbreviations as in figure 2. Enzymes known to perform electron bifurcation are indicated in red: heterodisulfide reductase (Hdr) [145] and hydrogenase (Hyd) [148].

lack cytochromes. Acetogens and methanogens are unique among chemiosmotic organisms studied so far in that both types exist as forms that lack cytochromes and quinones.

In brief, methanogens that lack cytochromes, as *Methanobacterium marburgensis* (figure 3*a*) use their version of the Wood–Ljungdahl pathway to synthesize methyl-H_4MPT, but instead of donating it to CoFeS and CODH/ACS as in carbon metabolism (figure 2*b*), they donate it to the thiol moiety of CoM in an exergonic reaction ($\Delta G^{\circ\prime} = ca\ -30$ kJ mol^{-1}) catalysed by a membrane-bound methyltransferase (MtrA-H in figure 3*a*), which conserves energy via the pumping of Na^+ ions across the cytoplasmic membrane [40]. CoM-SH is regenerated from methyl-CoM by methyl-CoM reductase, which releases methane by condensing the CoM moiety with the thiol group of coenzyme B (CoB-SH) producing the heterodisulfide CoM-S-S-CoB. The CoM-SH and CoB-SH thiols are regenerated by reduction with electrons from H_2 by a heterodisulfide reductase (Hdr in figure 3*a*), in a highly exergonic reaction ($\Delta G^{\circ\prime} = -55$ kJ mol^{-1}) [145]. Hdr also generates Fd_{red}, which serves as the low-potential reductant for the formyl-MF dehydrogenase step at the beginning of the WL-pathway, and the significance of this Hdr reaction will be discussed shortly. Per mol of methane, 2 mol Na^+ ions are pumped, whereas the ATPase requires about 4 Na^+ per ATP [145]. The overall reaction is that given in reaction (2.4), whereby, per mol methane, the cell is able to condense about 0.5 mol of ATP from ADP and P_i [145]. Aside from the anhydride bond in the ATP that is the product of the pathway, there are no thioesters, no acyl phosphates and no ATP consumption involved in methane production. The main energy currency is low-potential-reduced ferredoxin, Fd_{red} [143].

Acetogens that lack cytochromes use their version of the WL-pathway to make methyl-H_4F, as sketched in figure 3*b*, which is modified from Pohlein *et al.* [146] and Buckel & Thauer [143] for *Acetobacterium woodii*. ATP investment (a concept familiar from glycoysis) is required at the endergonic formyl-H_4F synthetase step, formyl phosphate (figure 2*c*) being the 'activated' ATP-generated reaction intermediate, both in *E. coli* [154,155] and in the acetogenic enzyme [156]. The methyl group of methyl-H_4F is transferred via CoFeSP to CODH/ACS where the acetyl-CoA is released. The energy in the thioester bond is conserved as acetyl phosphate in the phosphotransacetylase reaction. Acetate kinase converts acetyl phosphate and ADP into acetate and ATP, thereby recovering the ATP investment at the formyl-H_4F synthetase step. Methyl synthesis consumes electrons from H_2 according to reaction (2.3). The *A. woodii* formate dehydrogenase probably uses H_2 directly as substrate [146], whereas the remaining two reduction steps are NADH-dependent [143,146]. The NADH stems from a soluble hydrogenase (Hyd) that, importantly, also generates Fd_{red} in the process [148]. Fd_{red} is the substrate for a membrane-bound protein called Rnf [147] that generates NADH and pumps, in the process, about one Na^+ ion per electron transferred, thus generating the ion gradient that is harnessed by the ATPase, which requires about 4 Na^+ ions per ATP. Per mol of acetate, about 0.25 mol of ATP is generated [143]. The ATP investment at formyl-H_4F synthetase step and return via acetate kinase is strictly 1 : 1, so there is no net gain of ATP via that route. The currency of energy conservation is again Fd_{red}, which powers pumping at Rnf to drive the ATPase.

6. Electron bifurcation

Both acetogens and methanogens reduce CO_2 to a methyl group using electrons from H_2. This reaction actually should not proceed at equimolar concentrations, because the midpoint redox potential, $E^{\circ\prime}$, of the H_2/H^+ couple is −414 mV, while the $E^{\circ\prime}$ for the CO_2/formate couple is −430 mV and for the formate/formaldehyde couple it is −580 mV. In simple terms, H_2 is not a strong enough electron donor to reduce CO_2 to the level of a methyl group; the electrons would have to go steeply uphill. This prompted Wächtershäuser [157, pp. 1790–1791] to conclude that H_2 can be 'excluded as the first source of electrons since its reducing potential is not sufficient for reducing CO_2'. So how do chemolithoautotrophs reduce CO_2 with electrons from H_2? The simple answer would be that they generate low-potential Fd_{red} with midpoint potentials of the order of −500 mV, but that is also still steeply uphill from H_2, bringing us to the crux of the issue: how do they generate Fd_{red} from H_2? The answer, which only recently became apparent and is an exciting development in energetics, is that they perform a reaction called flavin-based electron bifurcation [143,158]. This electron bifurcation occurs at the reaction catalysed by Hdr in the methanogens, and at the reaction catalysed by Hyd in the acetogens (figure 3).

This newly discovered mechanism couples an endergonic reaction to an exergonic one, analogously to the familiar case of coupling an unfavourable reaction to ATP hydrolysis so that the overall energetics of the reaction are favourable. But in electron bifurcation, no ATP is involved. Instead, it entails the splitting of electron pairs, from H_2, for example ($E^{\circ\prime} =$ −414 mV), at a flavin, such that one electron goes energetically uphill to generate reduced low-potential ferredoxins (Fd_{red}) with a midpoint potential close to −500 mV, with the energy for that uphill climb being provided by the second electron of the pair going downhill to a high-potential acceptor, such as the heterodisulfide CoM-S-S-CoB ($E^{\circ\prime} = -140$ mV) of methanogens [145] or NAD^+ ($E^{\circ\prime} = -320$ mV) at the electron-bifurcating hydrogenase of acetogens [148]. Electron bifurcation permits conservation of biochemical energy in the currency of reduced ferredoxin—a principle of energy conservation that departs from Lipmann's [149] high-energy bonds. Fd_{red} contains no cleavable 'high-energy' bonds, but electrons at low potential, so it is clearly a currency of chemical energy [143]. Oxidation of Fd_{red} by a pumping complex such as Rnf [147] or the Ech hydrogenase [159] couples the electron bifurcation mechanism to chemiosmotic energy conservation (figure 4).

The electrons in low-potential Fd_{red} are crucial for CO_2 reduction in organisms that use the acetyl-CoA pathway. By splitting the electron pair in H_2 with one going uphill to Fd the other going downhill to a high-potential acceptor, these organisms can reduce CO_2 under what might seem to be energetically hopeless conditions. The prevalence of electron bifurcation among strict anaerobes suggests that it is a very ancient biochemical mechanism [73,143], one that was apparently a prerequisite for a lifestyle of reducing CO_2 with electrons from H_2, which is how organisms that use the acetyl-CoA pathway for their carbon and energy metabolism survive.

7. Counting rocks in genomes

In anaerobes, Fd is usually one of the most abundant proteins in the cell, pointing to the central importance of this small, FeS cluster containing protein in their energy metabolism. But Fd is not the only abundant FeS protein in these organisms. The proteins of anaerobes are generally rich in transition metals and transition metal sulfide clusters at the catalytic centres of their proteins, catalytic clusters which are often replaced by NAD^+ and other organic cofactors in aerotolerant species. The early Earth was devoid of molecular oxygen [30]; hence the first life was anaerobic. There is a consistent and robust line of biochemical evolutionary reasoning that FeS clusters, chemolithoautotrophy and anaerobes tend to reflect antiquity [44,49,160,161]. That line of reasoning is still current [73,162,163] and is reinforced by geochemical data providing evidence for the antiquity of methanogenesis [164].

If FeS clusters are relics from the ancient past, it could be that the most ancient organisms preserved more FeS-containing proteins in their genomes. Major *et al.* [165] found that among the 120 genomes sampled then, the methanogens had by far the highest frequency of FeS proteins. Here, we have updated their survey with 1606 genomes, looking for the frequency of 4Fe-4S clusters of the type $CX_2CX_2CX_3C$. We observed the distribution shown in figure 5 which clearly shows that methanogens, acetogens and sulfate reducers have the highest frequencies of FeS proteins. This fits well with our general premises, but it does not prove anything. There is a point to be made, though, in that the case has recently been argued that ZnS, rather than transition metal sulfides, might have been the starting point of prokaryotic biochemistry [5]. However, it should be stressed that there is a big difference between the view that life emerged from a ZnS-mediated process and the view that life emerged from FeS and other transition metal sulfide-mediated processes. In the model of Mulkidjanian *et al.* [5], ZnS has the role of providing electrons (photochemically), a role that we ascribe to serpentinization (geochemically), but that is not such a big difference. The big difference is that Zn is a substrate in their model, providing two electrons per atom in the one-off photochemical reaction, not a catalyst (which FeNiS is in our model). Zn is not catalytic in the way that Fe, Ni, Mo, W or other transition metals are anyway, for which reason modern anaerobes lack ZnS centres participating in redox reactions of carbon or energy metabolism. The Zn model lacks redox catalysis and similarity to modern metabolism.

Following the physiological line of reasoning on evolution a bit further, Decker *et al.* [44] suggested that the next physiological group to arise following the acetogens and the methanogens were the sulfate reducers, strict anaerobes, many of which are autotrophs that use the acetyl-CoA pathway [167–171]. This notion would fit very well within the present considerations, including the circumstance that there are archaebacterial and eubacterial sulfate reducers [172]. It is furthermore supported by geochemical isotope evidence that attests to the existence of processes performed by sulfate reducers—sulfate/sulfite reduction and sulfur disproportionation—in the early Archaean era, around 3.4 billion years ago [173–175].

The ancestral metabolism performed by sulfate reducers was likely to have been sulfite reduction or sulfur disproportionation—and perhaps also disulfide disproportionation, as newer findings suggest [176]—since sulfur and sulfite would be produced abundantly from volcanic and hydrothermal SO_2, whereas sulfate, prior to the advent of oxygen, would probably have had only very limited, localized significance [177]. In line with that, the use of sulfate as a substrate by sulfate reducers involves soluble ATP consumption steps to

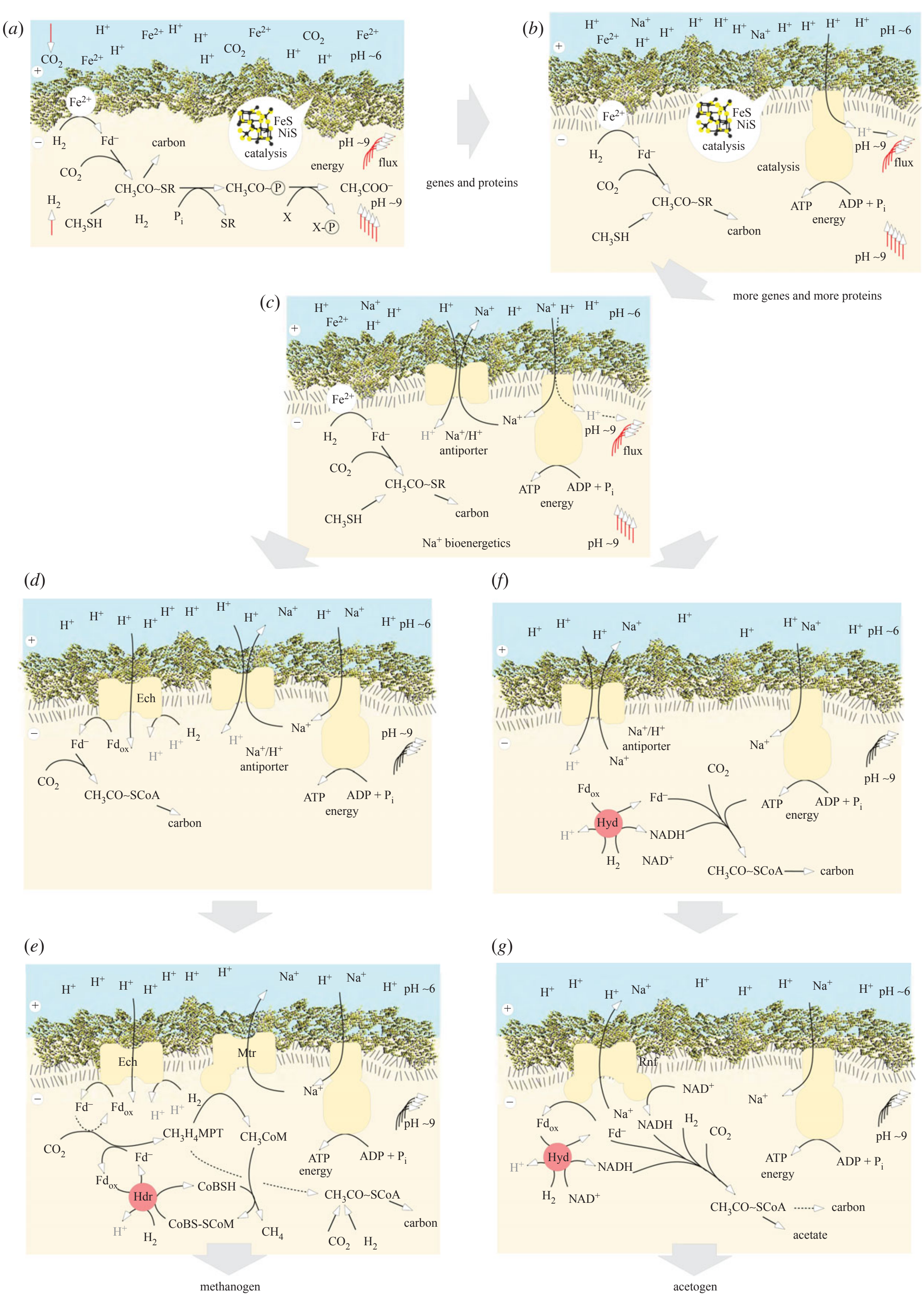


Figure 4. A hypothetical path for the events linking figures 2 and 3, redrawn from [43]. Ech: Energy converting hydrogenase [159]. Other abbreviations as in figures 2 and 3. See text.

produce sulfite [172]. On the early Earth, localized sulfate concentrations could be formed abiotically from atmospheric photolysis of SO_2, or biologically from sulfide-dependent anoxygenic photosynthesis or sulfur disproportionation, but it was not until the increased oxygenation of the atmosphere that oceanic sulfate levels started to increase and sulfate began to play a prominent role in the global sulfur cycle [178]. Dissimilatory sulfite reduction is a more widespread trait than sulfate reduction, and the enzyme responsible for it, the dissimilatory sulfite reductase (DsrAB) is clearly an enzyme of very ancient origin, and it is closely related to the assimilatory enzyme present in all domains of life [179]. Both assimilatory and dissimilatory enzymes result from a gene duplication event that preceded the divergence of the

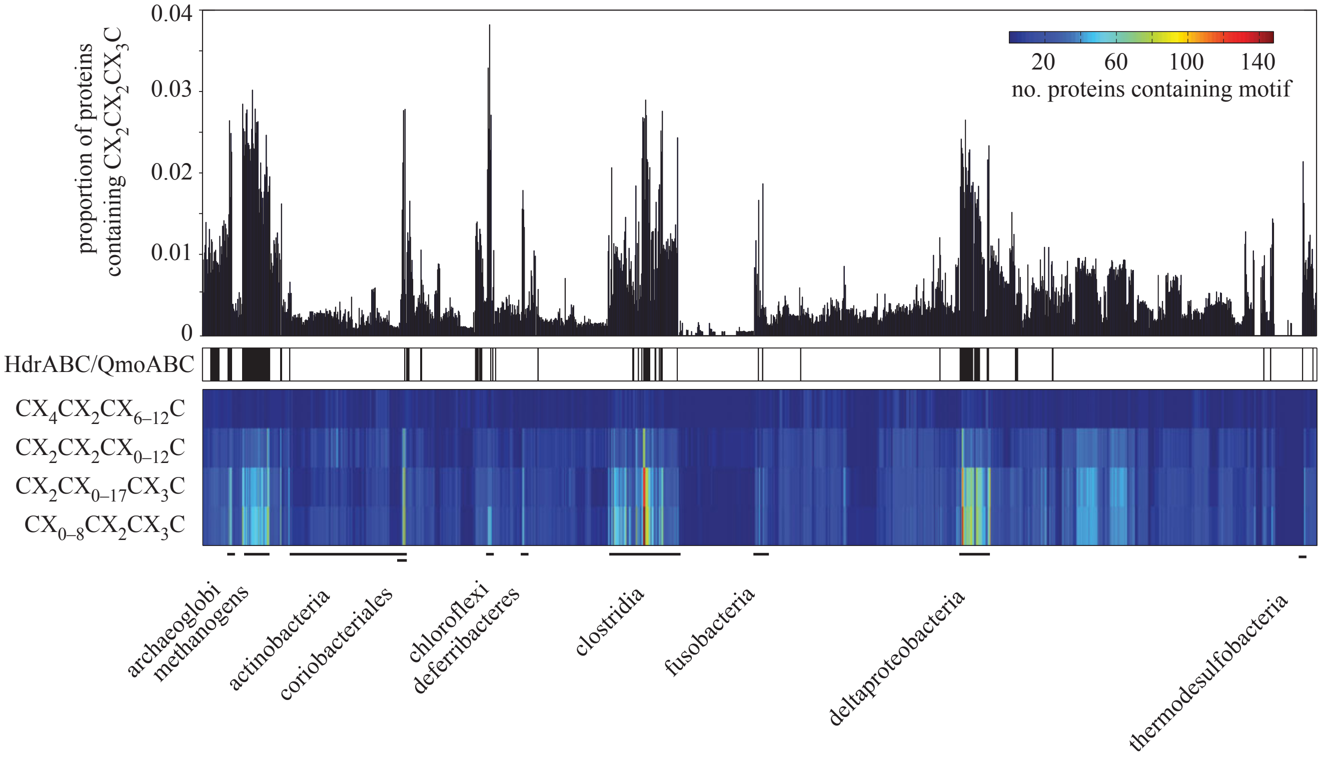


Figure 5. Occurrence of 4Fe–4S cluster motifs among 1606 prokaryotic genomes. The upper part of the figure represents the results of a search for the general form of the 4Fe–4S cluster-forming protein motif $CX_2CX_2CX_3C$. The proteins of 1606 prokaryotic genomes from the RefSeq database (v03.2012) [166] were under examination for this. The prokaryotes are represented by single bars and are ordered by their taxonomical classification. The height of each bar indicates the proportion of proteins within a single genome containing the motif. The lower part of the figure gives the absolute number of proteins containing one of the four additional motifs, which have different numbers of bridging amino acids. The order of the columns is the same as the corresponding bars in the upper part. The following prokaryotes having a high abundance of $CX_2CX_2CX_3C$ motif containing proteins were marked: archaeoglobi, methanogens (methanobacteria, methanococci, methanomicrobia), coriobacteriales (actinobacteria), dehalococcoidetes (chloroflexi), deferribacteres, clostridia, fusobacteria, deltaproteobacteria and thermodesulfobacteria. All 1606 taxa were also checked for the presence of heterodisulfide reductase subunits (HdrABC) or its relative, the quinone-interacting membrane-bound oxidoreductase subunits of sulfate reducers (QmoABC), as these might hint at the presence of electron bifurcation involving these proteins; black bars indicate taxa where at least one of them was present. Note that several other enzymes known to be involved in electron bifurcation [143] are not indicated.

archaebacterial and eubacterial domains [180–182], so a primordial sirohaem-containing sulfite reductase was most probably present in LUCA. Note that the presence of sirohaem does not mean the presence of haem in LUCA, because sirohaem synthesis merely entails the insertion of Fe^{2+} into sirohydrochlorin, an intermediate of anaerobic cobalamine synthesis (which acetogens and methanogens possess) and sirohaem is furthermore a precursor of the archaebacterial (alternative) haem biosynthetic pathway [183].

Something else that speaks rather strongly for the antiquity of methanogens, acetogens and sulfate reducers is their prevalence in the deep biosphere, where, as on the early Earth, there are few ways to make a living. Sulfate reducers, acetogens and methanogens appear to be the most common inhabitants of ecosystems situated deep within the crust [184–188]. They inhabit such environments today, suggesting that their ancestors might have as well. Curiously, though not involved in energy metabolism directly, primitive forms of Dsr-like proteins are apparently abundant in methanogens [189], with a role in sulfite assimilation or detoxification, further stressing the close ties between sulfate reducers and methanogens.

8. Electron bifurcation and sulfate reducers

Another link between methanogens and sulfate reducers is the occurrence in the latter of a number of Hdr-related proteins, suggestive of electron bifurcation [172]. For methanogens, the critical bifurcation reaction, described earlier, is catalysed by a complex of a heterodisulfide reductase with a hydrogenase, the MvhADG–HdrABC complex [145]. Homologues of HdrABC are found in other organisms, including at least one cytochrome-containing acetogen, *Moorella thermoacetica* [89], and many proteobacterial and clostridial sulfate reducers [172]. In both cases, it is suspected that HdrABC-mediated electron bifurcation is participating in core carbon and energy metabolism, but this has not been directly shown so far. For *Moorella*, HdrABC homologues are encoded directly next to the gene for methylene-H_4F reductase, which catalyses a very exergonic reaction and was therefore once suggested as a coupling site for energy conservation [39]. It remains possible that the *Moorella* methylene-H_4F reducatase does conserve energy, but not via ion pumping as originaly envisaged, rather via electron bifurcation [89] and Fd reduction instead. Recently, an electron-bifurcating ferredoxin- and NAD-dependent [FeFe]-hydrogenase (HydABC) from *Moorella* was shown to catalyse the reversible bifurcation reaction [190]. For the sulfate reducers, several organisms contain HdrABC or HdrA-like proteins, encoded next to genes for possible electron donor proteins such as an MvhADG hydrogenase, a NAD(P)H dehydrogenase or a formate dehydrogenase [172], strongly suggesting that electron bifurcation is also involved in their energy conservation.

Furthermore, several proteins directly implicated in sulfate/sulfite reduction are strikingly related to HdrABC subunits.

Sulfate reducers contain menaquinone, but its role in respiration was for long a mystery because the terminal reductases, APS reductase (AprBA) and DsrAB, are soluble cytoplasmic enzymes. A link between the two started to emerge with the identification of two membrane complexes having quinone-interacting subunits, QmoABC and DsrMKJOP, as probable electron donors to AprBA and DsrAB [191–193]. Both these complexes have Hdr-related subunits. The QmoABC complex is essential for sulfate reduction [194], and two of its subunits, QmoA and QmoB, are both flavoproteins closely related to the bifurcating subunit HdrA. In the DsrMKJOP complex, the two DsrMK subunits are closely related to the membrane-associated HdrED enzyme present in cytochrome-containing methanogens. DsrK contains the special catalytic FeS cluster that in HdrB is responsible for reduction of the heterodisulfide, which in the case of the sulfate reducers is proposed to be the small protein DsrC that is involved in sulfite reduction by DsrAB [195].

How energy is conserved by the QmoABC and DsrMKJOP membrane complexes has not been clearly established, but a recent proposal suggested a new mechanism of electron confurcation—bifurcation in the reverse direction [196]—involving menaquinone in the case of Qmo [197]. In this proposal, the endergonic reduction of APS by menaquinol is coupled to its exergonic reduction by a low-potential soluble electron donor (such as Fd). In contrast to electron bifurcation where a single-electron donor (NADH, H_2 or formate) is used to reduce two electron acceptors, one of high and one of low potential (most commonly Fd), in this quinone-linked confurcation two electron donors, of high (menaquinone) and low potential (Fd?), are used to reduce a single-electron acceptor. Soluble confurcation processes have already been described, such as those involving multimeric hydrogenases (NADH and Fd_{red} as electron donors for H_2 production) [196] and the NfnAB transhydrogenase (NADH and Fd_{red} as electron donors for $NADP^+$ reduction) [89,198]. Having menaquinone as one of the confurcation players would allow for transmembrane electron transfer and charge separation to effectively couple the process to chemiosmotic energy conservation. This could not occur if only a soluble low-potential donor was involved. It is possible that this, so far speculative, idea might also apply to other anaerobes for which the role of menaquinones is so far unclear, and where MK does not have a sufficiently low reduction potential to participate in core energy or respiratory metabolism. The cytochrome-containing acetogens such as *Moorella*, in particular, come to mind in this respect, because they contain menaquinone, but there is yet no obvious role for it in core metabolism [89].

The involvement of electron bifurcation in sulfate reducers points further to the antiquity of sulfate/sulfite reduction as an ancient anaerobic physiology and is consistent with the observation that a number of sulfate reducers use the acetyl-CoA pathway [199–201].

9. From harnessing to pumping: the same leap by two prokaryotes

Electron bifurcation in methanogens and acetogens permits generation of membrane potential from H_2 and CO_2, which can then be used to drive net ATP synthesis via the ATP synthase, or in the case of *M. marburgensis*, to reduce ferredoxin via the energy converting hydrogenase, Ech. Thus, electron bifurcation helps cells generate what alkaline hydrothermal vents geochemically provide for free: ion gradients. Given the foregoing, the continuous reduction of CO_2 to organics could have ultimately driven the emergence of RNA, proteins and genes within the vent pores: what we shall call protocells, meaning the organic contents of inorganic pores within the microporous labyrinth of alkaline hydrothermal vents, but not, as yet, fully functional or independently dividing cells. How (and why) did these confined protocells, dependent on natural pH gradients, begin to generate their own gradients, enabling them ultimately to leave the vents as self-sufficient, free-living cells?

One possible solution depends on the counterintuitive properties of membranes in the presence of natural ion gradients [43]. Apart from the remarkable sophistication of modern respiratory chains, a second reason that chemiosmotic coupling is usually considered to have arisen late relates to the permeability of early membranes. Chemiosmotic coupling today requires an ion-tight membrane, so that ions pumped out return mostly through the ATP synthase, driving ATP synthesis. If ions can instead return through the lipid phase of a leaky membrane, ATP synthesis is said to be uncoupled—literally, short-circuited—and most of the energy consumed in pumping is simply dissipated as heat. Thus, pumping protons across a leaky membrane costs a lot more energy than can be conserved as ATP, making it worse than useless. The problem at the origin of life is that early membranes would certainly have been leaky, hence pumping protons across them could offer no advantage, especially in the presence of small organic acids which naturally dissipate proton gradients. Pumping protons at the vent is pointless anyway, as the proton gradient already exists. A second related problem is that archaebacteria and eubacteria do not have homologous membranes; hence we have argued that their last ancestor, confined to vents, did not have a modern membrane. While simple phase separation of amphiphiles, fatty acids, alkanes and other hydrophobics would tend to produce hydrophobic layers within the vent pores, probably even lipid bilayers, these were likely to be very leaky to small ions.

But arguments against early chemiosmotic coupling share a common flaw: they assume that pumping arose before harnessing. In the presence of natural proton gradients, this is not the case. We have argued that the ATPase arose before pumping mechanisms. The passage of protons through the ATP synthase today is about 4–5 orders of magnitude faster than through the lipid phase of the membrane. But with a free, geochemically provided gradient, membrane permeability could be orders of magnitude greater, while still allowing net ATP synthesis. All that is required is that protons should pass through the ATP synthase as well as through the lipid phase (and that they be removed by the alkaline effluent). Any improvements in the ion-tightness of the membrane would therefore be beneficial, as they would funnel a larger proportion of protons through the ATPase. Hence if there were any genetic component to early membranes, we would predict that coupling would steadily improve, which is to say that organic membranes would tend to become less permeable to small ions over time. On the face of it, that seems a reasonable statement, but it leads to an unexpected problem.

The problem relates to the transfer of charged particles—protons—across the membrane. In vents, neither hydrothermal fluids nor ocean waters carry a net charge: the excess of H^+ in the ocean is balanced by HCO_3^-, whereas the excess of OH^- in hydrothermal fluids is balanced by Mg^{2+}. Thus, the transfer of a single proton across the membrane down a concentration gradient also transfers a positive charge, and this opposes the further transfer of charge—which is to say, more protons. Soon, a Donnan equilibrium is established, in which the electrical charge across the membrane balances the concentration gradient, annulling the proton-motive force. This problem is only unmasked if the membrane can actually hold an electric charge, which is to say, only if the membrane has become impermeable to small ions. Before that, however—if the membrane is leaky or discontinuous—the system is essentially open, and charge does not enter into it. Any H^+ crossing the membrane will react with hydrothermal OH^- to form water, which is swept away in the flow, negating the accrual of charge. In such an open system, the pH gradient is maintained not by pumping, but by the continuous circulation of hydrothermal fluids and ocean waters, continually juxtaposing phases in redox and pH disequilibrium. Thus, at an early stage, the proton-motive force can only exist in the presence of leaky membranes and an open system: far from being a problem, leakiness is actually a necessity.

What happens when the membrane tightens off to small ions and forms cell-like structures? Transfer of protons into an impermeable cell will result in a Donnan equilibrium that opposes further transfer unless counterions are also transferred, or unless the proton can be pumped out again. Pumping out a proton is problematic. Either the cell needs to invent machinery to pump out protons as soon as the membrane becomes impermeable (plainly impossible) or it needs to invent this machinery while the membrane is still permeable, i.e. before it needs it. This is also unlikely, as the cell would need to actively pump protons (costing energy) from an internal concentration of pH 9–10 (less than or equal to 1 nM) across a leaky membrane against a 1000-fold concentration gradient, which is immediately dissipated through the leaky membrane—a common objection to chemiosmotic coupling evolving early. The solution is surprisingly simple: a Na^+/H^+ antiporter, an otherwise unassuming protein that would hardly be a prime suspect for evolutionary heroics, except in alkaline hydrothermal vents.

In a world of genes and proteins, a Na^+/H^+ antiporter is far easier to invent than an ATPase. It is a single protein, and an ancient one that is also a component of both Ech and Complex I [202,203]. But it confers an inordinate benefit. An antiporter transfers a positive charge in the opposite direction to the proton, meaning that overall the passage of a proton down a concentration gradient is charge neutral. Thus, with an antiporter, there is no problem with Donnan equilibria: the system does not bung up. What is more, the passage of H^+ down a natural gradient drives the efflux of Na^+ ions. This has two beneficial effects, one immediate and the other longer term. The immediate benefit is improved coupling, assuming, as we do, that ancient ATPases were promiscuous for H^+ and Na^+, as modern ATPases of methanogens are [204]. Modern membranes are 2–3 orders of magnitude more permeable to H^+ than to Na^+, and there is no reason to suppose that early membranes would have been any different, especially in vents, where the presence of small organic acids uncouples H^+ gradients but not Na^+ gradients. So an antiporter transduces a natural proton gradient into a biochemical sodium gradient, which immediately improves coupling at essentially no energetic cost. Na^+ extrusion is powered by the free geochemical proton gradient, not by active pumping (which consumes an energy budget that could otherwise be spent on carbon assimilation). Any improvement in coupling, of course, requires that the ancient ATPase could use Na^+ ions, similar to how the modern one can [204,205], perhaps because hydronium ions (H_3O^+) have an identical charge and similar ionic radius to Na^+ (ionic radii for unhydrated $H_3O^+ = 115$ pm, and for $Na^+ =$ 117 pm). A number of apparently ancient bioenergetic proteins found in both methanogens and acetogens, notably the ATPase and Ech, are promiscuous for H^+ and Na^+, which might be a relic of early coupling.

The longer term benefit relates to optimizing enzyme function by ionic balance. Most modern cells have low intracellular Na^+ relative to K^+, unlike the oceans, where Na^+ concentration is much higher, 470 mM in oceans versus approximately 10 mM in cells [5]. Some universal and ancient proteins, notably those involved in translation, are optimized to this ionic balance, raising the question of how, if life arose in submarine alkaline hydrothermal vents, did intracellular Na^+ concentration become lowered to this optimum? An H^+/Na^+ antiporter would readily explain how, but as the secondary consequence of a more immediate bioenergetic pressure.

The Na^+/H^+ antiporter also helps to solve the larger problem of the origin of active pumping. In the presence of free geochemical proton gradients, protocells in vents can adapt to a sodium-motive force, all the while being powered by a proton-motive force. Eventually, cell membranes became tight to Na^+, if not H^+. So how did active pumping originate?

Here, an important point needs to be made, lest the reader accuse us of invoking preadaptation with regard to pumping. There is no advantage to active pumping at regions of the vent where gradients are sharp by means of effluent flux. In distal regions of the vent, however, where alkaline effluent flux is impaired, energy coupling, and hence proliferation would cease. Only such protocells that 'learned'—via the standard workings of natural variation and selection—to generate their own gradients in the context of their existing repertoire of redox reactions, could proliferate in such margins. And only such protocells could make the final transition to the free-living lifestyle. So there is clear proliferation benefit for the invention, by chance mutation, of proteins that can pump, and that benefit exists at regions of the vent where flux is impaired. This is likely the initial selective advantage of pumping.

What are the primordial active ion pumps? A simple solution is suggested by the sole coupling site of acetogens that lack cytochromes, Rnf. Electron transfer from ferredoxin to NAD^+ drives the extrusion of Na^+ via Rnf [147]. The problem of generating reduced ferredoxin in protoacetogens via flavin-based electron bifurcation at the hydrogenase step had been solved previously for the purpose of reducing CO_2. The circuit was nearly complete, only redox balance remained to be achieved. Reoxidizing the NADH generated at the hydrogenase and Rnf reactions so as to synthesize methyl groups closes the stoichiometric loop, yielding an H_2/CO_2 dependent, acetate excreting entity that, upon cellularization (bubbling off, one might say) was energetically free to leave the vents.

Methanogens found a different and completely unrelated route, while also making use of flavin-based electron bifurcation, suggesting independent origins of active pumping. In the case of methanogens, the Na^+-motive force is generated by pumping at the highly exergonic methyl transfer step ($\Delta G^{\circ\prime} = ca\ -30\ \mathrm{kJ\ mol^{-1}}$) from methyl-$H_4$MPT to CoM by the methyl transferase (Mtr) complex, the only coupling site in methanogens that lack cytochromes [145]. Reoxidation of CoB-SH with methyl-CoM to generate the heterodisulfide CoM-S-S-CoB and methane establishes a redox balanced Na^+ pumping circuit. In the protomethanogen lineage, this Na^+ gradient could also be tapped via Ech for net CO_2 to proceed. Regardless of the exact details, it appears that the deep divergence of overall design in methanogen and acetogen bioenergetics, despite their distinct commonalities, occurred within the vents.

From rather similar starting points (but with membranes that had already diverged), both methanogens and acetogens solved the problem of active pumping using an Na^+/H^+ antiporter, both relied on chemiosmotic ATP synthesis, and both evolved similar mechanisms of flavin-based electron bifurcation, using a similar set of hydrogenase proteins. But divergent stem lineages solved the details of the problem in different ways. And in both, flavin-based electron bifurcation gave rise to a balanced redox circuit and Na^+ gradient generation via a single coupling site. Other elements of the more familiar respiratory chain, notably quinones, the Q cycle, and cytochromes *b* and *c*, arose later. This interpretation is supported by the remarkable homology of several subunits of complex I to the proteins discussed here: FeFe trimeric hydrogenases, NiFe hydrogenases, ferredoxin-reducing subunits, Ech and Na^+/H^+ antiporters [202]. The only further subunits required for true complex I function were quinone and NADH-binding subunits. We discuss the origin of quinones and cytochromes in §13, but it seems likely that they did indeed arise after Ech and Na^+/H^+ antiporters. In this case, it is also noteworthy that complex I itself possesses some of the same Na^+/H^+ promiscuity [206] as Ech and the ATP synthase do. The fact that the proton-motive force is more universal today than the sodium-motive force probably derives from evolutionary interpolation of quinones—which are always coupled to proton-dependent redox reactions—on these earlier, necessarily promiscuous origins [205].

Thinking things through, one additional aspect comes to the fore that is not immediately related to energetics. Namely, if the transition to the free-living state was truly independent in the archaebacterial and eubacterial lineages, as we suggest, then prokarytic cell division well could be predicted to have arisen twice independently. A look at the comparative genomics of cell division in archeabacteria and eubacteria provides evidence for the ubiquity of the FtsZ-derived system in eubacteria, and widespread occurrence of FtsZ in archaebacteria, but in addition the occurrence of ESCRTIII-related cell division machineries in some archaebacterial lineages [207]. Thus, while one can say that the ESCRTIII-related route is specific to archaebacteria, it is not obviously present in their common ancestor, while FtsZ (which seems to be a main cell division protein in methanogens) probably was [207]. Assuming that cell division had to evolve before cells escaped to the free-living state, what might cell division within rigid inorganic confines have looked like?

The example of spore formation in the acetogen *Acetonema longum* [208] is possibly instructive. Sporulation is a specialized case of cell division where the daughter cell grows inside the mother cell: some members of the clostridia, such as *Epulopiscium*, use the cell within a cell strategy for normal cell division [209]. During that process in *Acetonema*, the inner membrane of the mother cell becomes the outer membrane of the spore (the daughter cell), which grows within the mother cell and which initially posseses two membranes, inner and outer, one of which is ultimately shed [208]. This general kind of cell division has, not unreasonably, been suggested as the ancestral source of the outer membrane in Gram-negative bacteria; it might also represent an ancestral state of cell division reflecting a growth process within physical confinement.

10. Not enough energy?

Recently, Nitschke & Russell [210, p. 482] have argued that the Wood–Ljungdahl pathway ‘appears inadequate to drive reductions, condensations and biosyntheses required of an emergent metabolic cooperative’. The gist of this argument is that the H_2–CO_2 couple does not present enough free energy to get life started (although methanogens and acetogens do just fine). Their inference is that high-potential electron acceptors such as NO or NO_3^- with midpoint potentials near or exceeding O_2 must have been involved at the very earliest stages of chemical evolution [210,211], and in the latest formulation entailing oxidant-driven methane oxidation as the first step towards the synthesis of organic molecules [212]. A severe problem with that view, though, is that it demands that existence of a 1 : 1 molar ratio of high-potential acceptors for every organically incorporated carbon atom at the onset of primordial biochemistry; that would shift the oxidation state of such an environment dramatically.

This is especially true when one recalls that at the onset of biochemistry, there was little or no specificity in catalysis, such that a vast molar excess of ‘non-biogenic’ (side product) reduced carbon compound was required for the synthesis of every ‘biogenic’ (biologically useful) one [43]. The consequence is that in the presence of abundant high-potential acceptors of the type that methane oxidation [212] would demand, organic synthesis would probably just ‘go up in smoke’ being pulled towards CO_2. The reason is because amino acid synthesis and cell mass synthesis, while requiring little if any energy input or even being exergonic under strictly anoxic hydrothermal vent conditions [36], become extremely endergonic even under very mildly oxidizing conditions, such as microoxic conditions corresponding to only 1/1000th of present oxygen levels [35].

Specifically, the synthesis of cell mass requires 13-fold greater energy input (or even more if NO_3^- is the nitrogen source) under microoxic than under anoxic conditions [35]. That meshes with the observations of Heijnen & van Dijken [213] that anaerobic chemolithoautotrophs such as methanogens require of the order of 30–40 kJ to synthesize a gram of cell mass, whereas aerobes require of the order of 80–170 kJ to synthesize a gram of cell mass [35]. Thus, the argument that high-potential electron acceptors for methane oxidation were needed to get organic synthesis started [212] has the problem that their presence would tend to preclude accumulation of reduced organic compounds such as amino acids [35], because the reducing conditions would be gone. One might counter that these microoxic conditions apply to the oceans, whereas the internal compartments of hydrothermal vents are highly reducing, gases such as NO are as likely to traverse hydrophobic walls as CO_2, and certainly more easily than protons, and would therefore inevitably alter the oxidation state within the vent. Indeed, McCollom & Amend

[35] point out that 'the primary control on the energetic requirements for biomass synthesis is the oxidation state of the environment' which is consistent with the observations of Amend & Shock [34], who pointed out that in environments of high or even moderate reducing potential, the autotrophic synthesis of amino acids requires little energy input, in contrast to oxidizing environments. Thermodynamic considerations would strongly favour a reducing environment for the synthesis of life's first building blocks over an environment in which the synthesis of reduced carbon compounds had to take place against the workings of strongly oxidizing agents. An additional complaint of Nitschke & Russell [210] that 'acetate has defied chemical synthesis in aqueous solution directly from CO_2 and H_2 in the laboratory' probably relates to electron bifurcation, whereby aqueous acetate synthesis from CO_2 with $Fe^{\pm 0}$ is facile [62].

And when it comes to the topic of strong oxidants, there is a lot to be said for molybdenum. Mo is involved in many interesting reactions, among them CO_2 reduction in the initial steps of the acetogen and methanogen methyl synthesis pathways (figure 2) and N_2 reduction in nitrogenase, which was recently shown to possess a carbon atom with only Fe ligands in the centre of its FeS cluster [214]. Mo is also involved in some reactions where it was long thought that only a strong oxidant such as O_2 could do the job. A recent example is the characterization of a molybdoenzyme from *Sterolibacterium denitrificans* that anaerobically hydroxylates a tertiary carbon atom in sterol [215], a reaction hitherto thought to require O_2. Another example is the anaerobic hydroxylation of ethylbenzene by ethylbenzene dehydrogenase, where Mo(VI) performs the oxidative step in the reaction mechanism [216].

11. A tree of tips

One might ask how phylogenomics stacks up against these concepts. Groups trying to work out the 'phylogeny' of prokaryotes [217] have come up with various schemes to classify groups, mainly based on the identification of some core set of genes that are concatenated and used to create a phylogenetic tree, a procedure laden with problems [218]. While such phylogenomic studies are worthwhile undertakings, they come with the heftiest of caveats. This is because even with massive amounts of data and refined molecular phylogenetic methods, it is a challenge to get the orders of mammals [219] or the orders of flowering plants [220] satisfactorily sorted out, and those processes span 'only' about 200 Myr into evolutionary time. Much, much harder are the early eukaryotic groups, which go back about 1.5 billion years [221], or even more difficult, linking eukaryotes to prokaryotes [222–224]. With that in mind, dismal appear the prospects of getting branching patterns sorted out for prokaryotes, which have been around for 3.8 billion years or more [30,164], assuming that there are any real branches in that phylogeny to be recovered in the first place [225], and the goal probably pushes phylogenetics beyond its limits.

What if we poll the genes that are used in such concatenation studies to see whether they individually tend to support the trees that they produce in concatenated analyses? We did that in figure 6. The result shows that the 48 core genes that can be distilled to be present in all of a small sample of 100 genomes recover, individually, the deep split between archaebacteria and eubacteria, and they recover recent phylogenetic signals at the tips of the tree (figure 6). The tree of life as currently constructed is based only on about 30 genes [217], which correspond to about 1 per cent of the genes found in an average prokaryotic genome; in that sense the tree of life is more the tree of 1 per cent [227]. But by viewing the tree in terms of how often different genes find the same branch for the concatenated tree they purport to uphold, what we see is not a tree of life, but a tree of tips [228]. In the deeper branches, not a single individual gene branches in the same way as the concatenated tree does. This underscores what Doolittle has been saying for some time [225,229], namely that scientists are unable to muster positive evidence that any given gene or set of genes has had the same history all the way back to the beginning of life's first cells.

Thus, while concatenated gene sets (analysed individually or in some concatenated manner) readily provide groups (a classification), the branching order of those groups remains obscure. But more importantly, even if one were, by some means, to get a branching pattern sorted (or agreed upon by consensus) for some core prokaryotic gene set, there remain incontrovertible and fundamental differences between classification of eukaryotic groups and classification of prokaryotic groups: the former is largely natural, but the latter is not [229]. For example, we can generate an approximate phylogeny for the evolution of vertebrates and map onto that phylogeny the origin of jaws, fins, lungs, feet, wings and so forth in a meaningful manner. But the same is not possible for prokaryotes because they readily spread their genes via lateral gene transfer (LGT), and what we find in modern genomes are useful collections of genes that represent viable strategies for survival in existing niches [225]. In other words, we cannot take any given prokaryote phylogeny and map out the linear evolution of physiological traits, because their evolutionary patterns are too complex, often marked by very sparse distributions.

Notwithstanding, it is possible to generate an argument, based on phylogenetic trees, that acetogens and methanogens are among the true latecomers in evolution, recently evolved and derived specialists [212]. But the branches in phylogenetic trees built from sequence data depend on many things, mainly things that go on inside computers [222]. And for each tree that suggests methanogens and acetogens to be latecomers [211], there is one to show that methanogens [222,224,230] and clostridia (acetogens) [217] branch at the base of their respective domains.

Why is the split between eubacteria and archaebacteria in figure 6 (and in many other such trees) so deep and so clear, while the next branches disappear? Figure 6 summarizes trees of universal genes. While they do not recover the middle ages of prokaryote evolution, they do tell us what we already know, that archaebacteria and eubacteria are very different. Under our current premises, their divergence corresponds to a comparatively short period of time, less than *ca* 400 Myr, that separates the advent of water at 4.2 Ga and evidence for life at 3.8 Ga [11,30]. Perhaps the simplest interpretation for this deep divergence is that it reflects the circumstance that the stem archaebacteria and stem eubacteria diverged at a time when proteins were still being invented and those that existed were getting better at what they do. In modern evolution, proteins evolve neutrally on the whole, functional constraints causing sequence conservation are the norm, and have been for most of Earth's history.

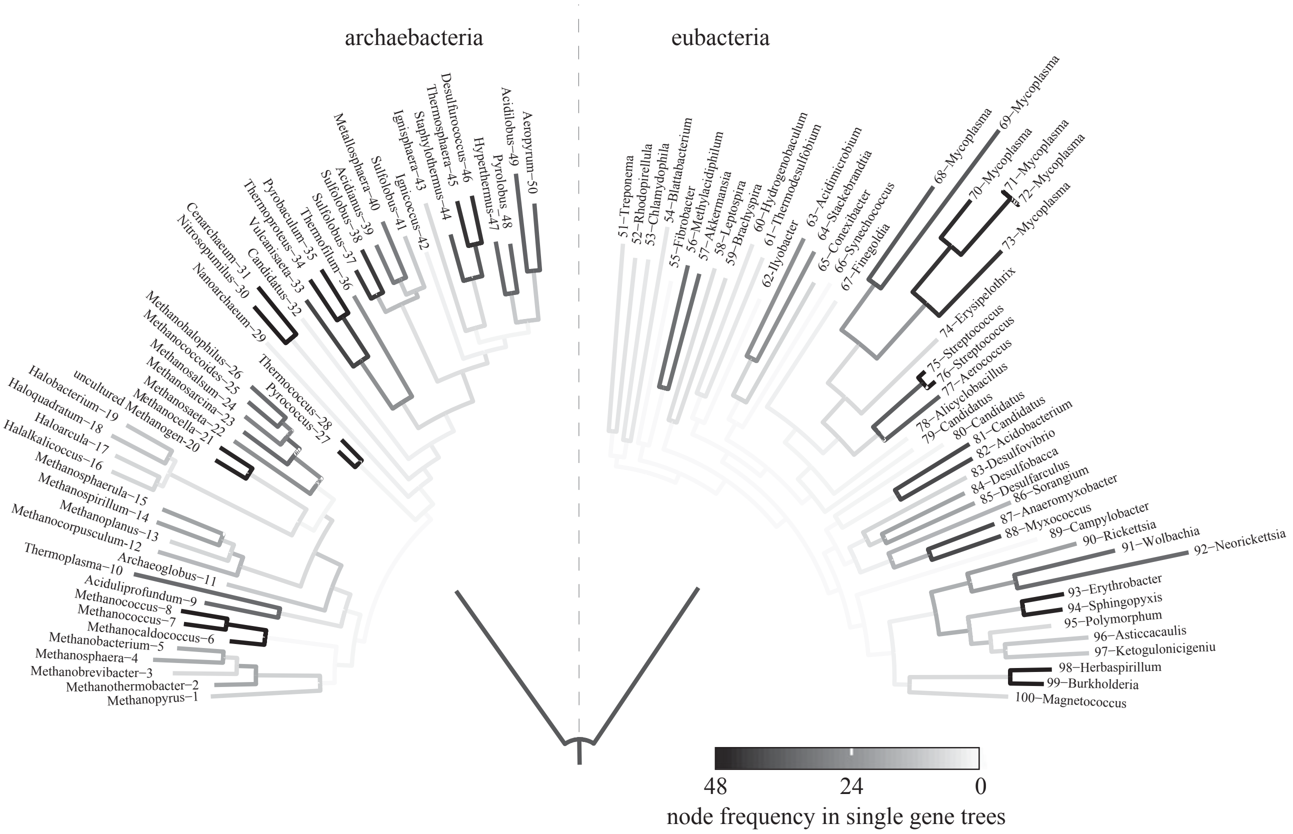


Figure 6. The 'amazing disappearing tree' of 48 universal genes for a 100 species set. A tree generated from a concatenated alignment of 48 universal genes, compared with its underlying single gene trees. The species sample comprises 50 archaebacteria and 50 eubacteria. To estimate the inconsistency between single gene trees and the concatenated tree, the frequency of each node in the concatenated tree was compared with its frequency within the single gene trees. The transparency of the branches reflects how often the associated node was present within the single gene trees. The 48 universal genes consist of the 31 genes that were previously identified as universal [226], and later used in phylogenetic analysis [217] namely (ArgRS, RNApol(a), LeuRS, metal-dependent protease, PheRS, GTPase, SecY, Rpl1, Rpl11, Rpl13, Rpl14, Rpl15, Rpl16/L10E, Rpl18, Rpl22, Rpl3, Rpl5, Rpl6, Rps11, Rps12, Rps13, Rps15/13E, Rps17, Rps2, Rps3, Rps4, Rps5, Rps7, Rps8, Rps9, Rps, SerRS), plus 17 additional genes (PRPP, AlaRS, PCNA homologue, RNApol(b), HisRS, Met-aminopeptidase, MetRS, PheRS beta subunit, ProRS, RecA, Rpl4, ThrRS, EfG, translation release factor, eIF5A, TyrRS, ValRS) that are present in this prokaryote sample, which contains no members with highly reduced genomes. The taxa were chosen for broad sampling. For this, proteomes of 1606 prokaryotes were retrieved from the RefSeq database (v03.2012) [166]. Pairwise sequence comparisons were run for the ribosomal protein L3. Based on these results, all prokaryotes were clustered by a hierarchical clustering algorithm. From each cluster 100 sample taxa (50 archaeabacteria and 50 eubacteria) were chosen. A complete list of genomes sampled is available in the electronic supplementary material.

But at the origin of genes, evolution was anything but neutral—it was fierce competition for organization of available matter into things that could organize matter best.

The examples encountered so far for independent origin of genes in archaebacteria and eubacteria clearly indicate that the invention of the full gene repertoire needed to exist as a free-living cell had not gone to completion in the confined universal ancestor. In the two stem prokaryote lineages, proteins were getting better at what they do (positive selection quickly fixes new mutations) and coming to rest in different regions of the fitness landscape on the plane of sequence space. This causes rapid sequence evolution, which translates into many accumulated sequence differences per unit time, which translates into long branches. Once proteins reached a certain level of optimization at what they do, functional constraints causing sequence conservation (slower mutation accumulation) became the norm.

12. What was present in the common ancestor?

We can look for proteins that are of sufficiently wide distribution to assume that they might have been present in the common ancestor of archaebacteria and eubacteria (table 1). If we opt for strict criteria as 'present in all genomes' then we would come up with a list of the roughly 31 proteins that various groups [217,226,232] have examined. An extended version of that list would also include many proteins that are less than 20 per cent identical in many pairwise comparisons. Our criteria for 'widely distributed' are more relaxed with respect to gene presence or absence but a bit more strict with respect to sequence identity. The protein families used by Nelson-Sathi *et al.* [231] to construct deep phylogenetic trees are convenient. The criterion of sharing at least 30 per cent amino acid identity was applied, because alignment and phylogeny procedures produce severe artefacts with amino acid sequences that share substantially less identity [233]. Concerning gene presence or absence, one can relax the stringency a bit, allowing for some loss, and ask whether a protein is present in at least one representative of various higher prokaryotic taxa corresponding to phylum, for example. In the study of Nelson-Sathi *et al.* [231], there were 11 higher archaebacterial taxa (75 genomes total) and 30 higher eubacterial taxa (1143 genomes total) sampled. Table 1 shows how many genes fulfil the distribution criteria for being present in (only!) at least one member each of 11/11 archaebacterial groups *and* in at least one member each of 20/30 higher prokaryotic

Table 1. Genes generously universal across the great prokaryotic (archaebacterial-eubacterial) divide at varying taxonomic stringence. The 75 archaebacterial genomes fall into 11 major taxonomic groups, the 1143 eubacterial genomes fall into 30 major taxonomic groups, as given in detail in [231]. To be scored as present within a taxonomic group, the proteins are required to be approximately 30 per cent identical, which is stringent, but we generously allow for loss. Thus, a gene counted as present is present in at least 11/75 and 20/1143 genomes. Most numbers are much larger, of course. The annotations of those particular 106 gene families are listed by functional category in table 2.

	eubacterial						
	all 30	≥25	≥20	≥15	≥10	≥5	≥2
archaebacterial	—	—	—	—	—	—	—
all 11	54	93	**106**[a]	117	120	125	130
≥10	72	152	172	191	201	213	224
≥9	86	205	237	262	276	294	308
≥8	107	254	302	341	359	385	407

[a]There are 106 gene families present in at least *one member each* of 20 out of 30 major eubacterial taxonomic groups (roughly corresponding to NCBI taxonomy phyla) and present in at least *one member each* of all 11 major archaebacterial taxonomic groups sampled and roughly 30% identical in all comparisons.

taxa *and* sharing more than 30 per cent amino acid identity. Table 2 shows which genes those are, by annotation for the case of 106 genes that are present in at least one member each of 11/11 archaebacterial groups *and* in at least one member each of 20/30 higher prokaryotic taxa *and* that share more than 30 per cent amino acid identity.

The list depicts what sorts of functions might have been present as genetically encoded functions in the common ancestor, allowing generously for loss and keeping in mind that we cannot readily tell how much LGT has contributed to those gene distributions. With those caveats, the list tends to reflect a last common ancestor that had ribosomes, the genetic code, bits and pieces of cofactor biosynthesis, bits and pieces of amino acid biosynthesis, bits and pieces of nucleotide biosynthesis, and a fully fledged ATPase that could convert an ion gradient into chemically accessible high-energy bonds. The paucity of obvious components that would generate an ion gradient is striking. The list in table 2 is remarkably consistent with the view that the genes listed arose in an environment (an alkaline hydrothermal vent), where geochemically produced proton gradients existed and only had to be tapped. The presence of several tRNA modifying enzymes in the list is also congruent with our arguments about modified bases.

13. Haem and menaquinone

The simplicity of the acetogen and methanogen bioenergetic pathways, not involving quinones or cytochromes and being replete in FeS centres, as well as their occurrence in strictly anaerobic chemolithoautotrophs, tends to speak in favour of their antiquity. An alternative view [211] has it that menaquinone (MK)-based-dependent proton gradient generation is the ancestral state of membrane bioenergetics, with a good portion of that argument resting on the wide distribution of MK among archaebacterial and eubacterial cells, and the occurrence of several ion translocating enzymes of bioenergetic membranes in some archaebacterial and eubacterial lineages. There are several problems with that view, though. Three problems with the argument for the 'antiquity based on ubiquity' of MK is that (i) methanogens do not have it at all, except derived methanogens that have acquired the genes for MK biosynthesis from eubacteria [231], (ii) members of the sulfolobales and acidobales synthesize benzothiophenes (quinone derivatives), but have no known MK pathway [234], and (iii) the distribution of the MK biosynthesis pathways across archaebacteria is generally sparse (figure 7), whereby tests of the two possibilities underlying that sparse distribution, differential loss versus lineage-specific acquisitions, have not been reported. Another problem with the view that quinone-cytochrome-based ion-pumping systems are ancestral is that haems are lacking in all methanogens except the evolutionarily derived group of the methanosarcinales and that the archaebacteria haem biosynthesis pathway is unrelated to and arose independently from the eubacterial pathway [130,131]. We have plotted the distribution of the two haem and the two MK biosynthetic pathways across groups in figure 7. Besides the exceptions mentioned earlier, an interesting observation is the absence of both MK pathways in some quinone-containing organisms as for instance thermotogae and tenericutes. This might suggest the existence of an additional MK biosynthetic pathway. Of course one could also argue that non-enzymatic haem synthesis, which does work [238], is the ancestral state used by the common ancestor. However, a more problematic aspect of the cytochromes first model is that, in that view, both methanogens and acetogens are seen as very late and highly derived evolutionary lineages [212], whereas isotope evidence has it that methanogens are ancient [164] and comparative physiology would have it that the core biochemistry of both methanogens and acetogens is ancient [44,73,160].

If our premise is correct that acetogens and methanogens that lack cytochromes and quinones are the most ancient kinds of cells, and that the primordial organisms did not have basically everything and could not do basically everything in the bioenergetic sense, a question arises, namely who invented cytochromes and respiratory chains? Some acetogens have cytochromes, but at least those of *Moorella thermoacetica* [239] are lateral acquisitions, probably from eubacterial sulfate reducers, as sequence comparisons readily reveal. Some methanogens also have cytochromes but they are also lateral acquisitions [231]. The deltaproteobacterial

Table 2. Functional annotation (using COG) of generously universal genes. List of 106 genes present in at least one member of all 11 archaeal groups and at least one member of 20 eubacterial groups (table 1).

no.	COG-id	category	product
I. information storage and processing			
1	COG0008	(J)	glutamyl- and glutaminyl-tRNA synthetases[a]
2	COG0012	(J)	predicted GTPase, probable translation factor[a]
3	COG0016	(J)	phenylalanyl-tRNA synthetase alpha subunit[a]
4	COG0017	(J)	aspartyl/asparaginyl-tRNA synthetases[a]
5	COG0024	(J)	methionine aminopeptidase[a]
6	COG0030	(J)	dimethyladenosine transferase (rRNA methylation)[a]
7	COG0048	(J)	ribosomal protein S12[a]
8	COG0051	(J)	ribosomal protein S10[a]
9	COG0060	(J)	isoleucyl-tRNA synthetase[a]
10	COG0080	(J)	ribosomal protein L11[a]
11	COG0081	(J)	ribosomal protein L1[a]
12	COG0086	(K)	DNA-directed RNA polymerase, beta' subunit/160 kD subunit[a]
13	COG0090	(J)	ribosomal protein L2[a]
14	COG0092	(J)	ribosomal protein S3[a]
15	COG0093	(J)	ribosomal protein L14[a]
16	COG0094	(J)	ribosomal protein L5[a]
17	COG0099	(J)	ribosomal protein S13[a]
18	COG0100	(J)	ribosomal protein S11[a]
19	COG0124	(J)	histidyl-tRNA synthetase[a]
20	COG0185	(J)	ribosomal protein S19[a]
21	COG0442	(J)	prolyl-tRNA synthetase[a]
22	COG0480	(J)	translation elongation factors (GTPases)[a]
23	COG0525	(J)	valyl-tRNA synthetase[a]
24	COG0532	(J)	translation initiation factor 2 (IF-2; GTPase)[a]
25	COG0621	(J)	2-methylthioadenine synthetase[a]
26	COG5256	(J)	translation elongation factor EF-1 alpha (GTPase)[a]
27	COG5257	(J)	translation initiation factor 2, gamma subunit (eIF-2 gamma; GTPase)[a]
28	COG0049	(J)	ribosomal protein S7
29	COG0013	(J)	alanyl-tRNA synthetase
30	COG0072	(J)	phenylalanyl-tRNA synthetase beta subunit
31	COG0087	(J)	ribosomal protein L3
32	COG0089	(J)	ribosomal protein L23
33	COG0096	(J)	ribosomal protein S8
34	COG0097	(J)	ribosomal protein L6P/L9E
35	COG0098	(J)	ribosomal protein S5
36	COG0102	(J)	ribosomal protein L13
37	COG0103	(J)	ribosomal protein S9
38	COG0130	(J)	pseudouridine synthase
39	COG0162	(J)	tyrosyl-tRNA synthetase
40	COG0164	(L)	ribonuclease HII
41	COG0180	(J)	tryptophanyl-tRNA synthetase
42	COG0343	(J)	queuine/archaeosine tRNA-ribosyltransferase
43	COG0522	(J)	ribosomal protein S4 and related proteins
44	COG1093	(J)	translation initiation factor 2, alpha subunit (eIF-2 alpha)

(*Continued*.)

Table 2. (*Continued.*)

no.	COG-id	category	product
45	COG1514	(J)	2′-5′ RNA ligase
46	COG2511	(J)	archaeal Glu-tRNAGln amidotransferase subunit E (contains GAD domain)
47	COG2890	(J)	methylase of polypeptide chain release factors
II. cellular process and signalling			
1	COG0396	(O)	ABC-type transport system involved in Fe-S cluster assembly, ATPase component[a]
2	COG0459	(O)	chaperonin GroEL (HSP60 family)[a]
3	COG0464	(O)	ATPases of the AAA+ class[a]
4	COG0489	(D)	ATPases involved in chromosome partitioning[a]
5	COG0492	(O)	thioredoxin reductase[a]
6	COG0533	(O)	metal-dependent proteases with possible chaperone activity[a]
7	COG0037	(D)	predicted ATPase of the PP-loop superfamily implicated in cell cycle control
8	COG0541	(U)	signal recognition particle GTPase
9	COG1180	(O)	pyruvate-formate lyase-activating enzyme
III. metabolism			
1	COG0020	(I)	undecaprenyl pyrophosphate synthase[a]
2	COG0078	(E)	ornithine carbamoyltransferase[a]
3	COG0082	(E)	chorismate synthase[a]
4	COG0112	(E)	glycine/serine hydroxymethyltransferase[a]
5	COG0126	(G)	3-phosphoglycerate kinase[a]
6	COG0127	(F)	xanthosine triphosphate pyrophosphatase[a]
7	COG0136	(E)	aspartate-semialdehyde dehydrogenase[a]
8	COG0142	(H)	geranylgeranyl pyrophosphate synthase[a]
9	COG0148	(G)	enolase[a]
10	COG0169	(E)	shikimate 5-dehydrogenase[a]
11	COG0171	(H)	NAD synthase[a]
12	COG0461	(F)	orotate phosphoribosyltransferase[a]
13	COG0498	(E)	threonine synthase[a]
14	COG0504	(F)	CTP synthase (UTP-ammonia lyase)[a]
15	COG0519	(F)	GMP synthase, PP-ATPase domain/subunit[a]
16	COG0540	[F)	aspartate carbamoyltransferase, catalytic chain[a]
17	COG1109	(G)	phosphomannomutase[a]
18	COG1155	(C)	archaeal/vacuolar-type H+-ATPase subunit A[a]
19	COG1156	(C)	archaeal/vacuolar-type H+-ATPase subunit B[a]
20	COG0002	(E)	acetylglutamate semialdehyde dehydrogenase
21	COG0005	(F)	purine nucleoside phosphorylase
22	COG0028	(EH)	thiamine pyrophosphate-requiring enzymes (acetolactate synthase, pyruvate dehydrogenase (cytochrome), glyoxylate carboligase, phosphonopyruvate decarboxylase)
23	COG0059	(EH)	ketol-acid reductoisomerase
24	COG0065	(E)	3-isopropylmalate dehydratase large subunit
25	COG0066	(E)	3-isopropylmalate dehydratase small subunit
26	COG0105	(F)	nucleoside diphosphate kinase
27	COG0119	(E)	isopropylmalate/homocitrate/citramalate synthases
28	COG0125	(F)	thymidylate kinase
29	COG0129	(EG)	dihydroxyacid dehydratase/phosphogluconate dehydratase
30	COG0137	(E)	argininosuccinate synthase

(*Continued.*)

Table 2. (*Continued.*)

no.	COG-id	category	product
31	COG0160	(E)	4-aminobutyrate aminotransferase and related aminotransferases
32	COG0174	(E)	glutamine synthetase
33	COG0179	(Q)	2-keto-4-pentenoate hydratase/2-oxohepta-3-ene-1,7-dioic acid hydratase (catechol pathway)
34	COG0252	(EJ)	L-asparaginase/archaeal Glu-tRNAGln amidotransferase subunit D
35	COG0460	(E)	homoserine dehydrogenase
36	COG0462	(FE)	phosphoribosylpyrophosphate synthetase
37	COG0473	(CE)	isocitrate/isopropylmalate dehydrogenase
38	COG0499	(H)	S-adenosylhomocysteine hydrolase
39	COG0528	(F)	uridylate kinase
40	COG1013	(C)	pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, beta subunit
41	COG1053	(C)	succinate dehydrogenase/fumarate reductase, flavoprotein subunit
42	COG1324	(P)	uncharacterized protein involved in tolerance to divalent cations
IV. unknown			
1	COG1163	(R)	predicted GTPase[a]
2	COG1245	(R)	predicted ATPase, RNase L inhibitor (RLI) homologue[a]
3	COG1782	(R)	predicted metal-dependent RNase, consists of a metallo-beta-lactamase domain and an RNA-binding KH domain
4	COG1690	(S)	uncharacterized conserved protein

[a]Present in at least one member of all 11 archaeal groups and at least one member of all 30 eubacterial groups.

sulfate reducers have regular cytochromes *b*, as well as multihaem cytochromes *c* [172] synthesized via the alternative archaeal haem pathway. Sulfate reducers also have MK, but they also have some things in common with methanogens in terms of Hdr-related proteins. Sulfate reducers do not, however, have membrane integral cytochrome *bc* or cytochrome b_6f complexes [172], and it is not yet clear how they use their MK pool, although a participation of electron confurcation, as discussed earlier, seems possible [197]. The circumstance that sulfate reduction is mainly a cytosolic process and that it likely involves flavin-based electron bifurcation (or confurcation) speaks for its antiquity. Indeed, given that there was surely no lack of sulfur substrates for sulfate reducers [179] at the vent–ocean interface, autotrophic sulfate reducers that use the acetyl-CoA pathway could, in principle, harbour a physiology that is nearly as ancient as that of acetogens and methanogens, requiring merely the additional evolutionary invention of cytochromes and quinones. The links between sulfur metabolism and methanogenesis point to the antiquity of both [163].

The prokaryotes who invented cytochrome *bc* type containing respiratory chains probably also invented the Q-cycle [240], the membrane phase and quinone-based analogue of flavin-based electron bifurcation, which is a cytosolic process so far. Among prokaryotes, flavins are far more universal than quinones. This suggests a possible sequence of events in the early evolution of energy conservation: (i) thioester-dependent substrate-level phosphorylations, (ii) chemiosmotic harnessing and the universality of ATP as energy currency, (iii) harnessing of Na^+ gradients generated by H^+/Na^+ antiporters, (iv) flavin-based bifurcation-dependent ion gradient generation, (v) quinone-based (and, eventually, Q-cycle containing) proton gradient generation involving membrane integral cytochrome complexes and bona fide respiratory chains. All of these processes ultimately depend, even today, upon CO_2 reduction with low-potential Fd_{red} (generated either chemosynthetically or photosynthetically), placing a reaction of the type 'reduced iron → reduced carbon' at the beginning of bioenergetic evolution, as outlined in figure 8. The evolutionary advent of quinones affected chemiosmotic pumping efficiency in at least two ways: (i) via quinone-dependent electron bifurcation in the Q cycle, both in cytochrome bc_1 and in cytochrome b_6f complexes [243], and (ii) in complex I. The recent structure of complex I reveals how redox-dependent quinone movement introduces conformational changes across several adjacent antiporter subunits, causing them to pump in concert [244]. This generates a greater $H^+/2e^-$ stoichiometry than the hydrogenase-related precursors of complex I which, despite a similar subunit composition [202], pump without the help of quinones.

14. Conclusion

Here, we have specified some links between the energy releasing chemistry of a Hadean alkaline hydrothermal vent and the energy metabolism of particular groups of modern microbes. Energetic aspects of the origin of life are less widely discussed than RNA-oriented genetic aspects, and one can approach the problem either from a very general perspective, without linking early energetic models to modern bioenergetic configurations, or from a more specific perspective that aims to forge tangible connections between ancient chemical environments and modern microbial physiology. The latter approach requires stating some premises about what is ancient in modern biochemistry, and the further one delves into the physiology of acetogens and methanogens that lack cytochromes, the more ancient they look. Moreover, the more closely we compare their chemistry to processes at alkaline hydrothermal vents, the

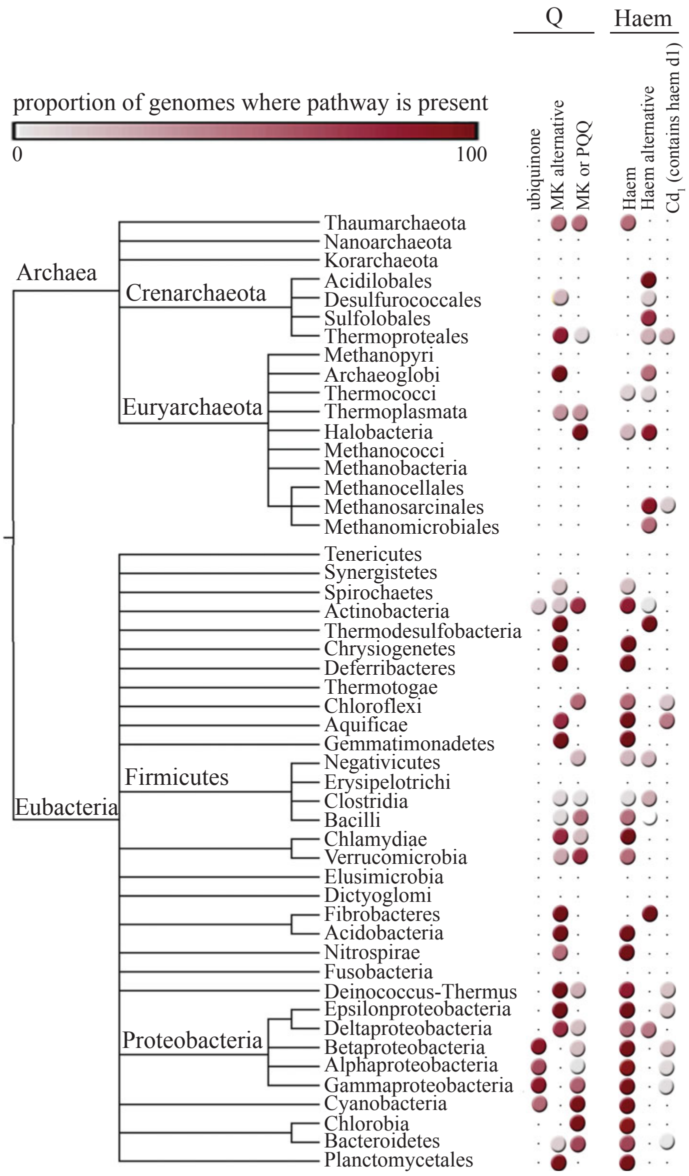


Figure 7. Distribution of quinone and haem biosynthetic pathways among 1606 prokaryotic genomes. The left part of the figure represents the organization of the selected taxonomic groups from the 1606 completed sequenced genomes (117 archaeal and 1489 eubacterial). The right part of the figure represents the proportion of genomes within a taxa where at least 70% of the genes involved in the pathway are present. Each column represents a different pathway. Homologous proteins involved in the several steps of ubiquinone (*ubiC*, *ubiA*, *ubiD/ubiX*, *ubiB*, *ubiH*, *ubiE*, *ubiF* and *ubiG*), menaquinone (MK) alternative (*MqnA*, *MqnB*, *MqnC* and *MqnD*) [235], menaquinone or phylloquinone (PQQ) (*MenF*, *MenD*, *MenH*, *MenC*, *MenE*, *MenB*, *MenA* and *UbiE/MenG*) [236], haem (*HemE*, *HemF/HemN*, *HemY/HemG* and *HemH*) and haem alternative (*AhbA*, *AhbB*, *AhbC* and *AhbD*) biosynthesis pathway were identified by BLAST [237]. The BLAST results were filtered for E values better than 10^{-10} and amino acid identities greater than or equal to 25 per cent. Owing to the high similarity between genes involved in haem d1 biosynthesis with genes from the haem alternative pathway [131], BLAST searches for the presence of cd1 nitrite reductase (the only enzyme containing haem d1) were also performed. In the genomes where both haem alternative pathway genes and cd1 nitrite reductase were present, the former were considered to be involved in haem d1 biosynthesis. Quinone biosynthesis distribution: ubiquinone is only present in the Eubacteria domain, mainly in proteobacteria (beta, alpha and gamma classes) and a few actinobacteria. It is an oxygen-dependent pathway being confined to aerobic organisms. The cyanobacterial ubiquinone hits reflect the presence of genes probably involved in plastoquinone biosynthesis instead. The two MK biosynthesis pathways are present in both prokaryotic domains although the MK alternative pathway has a broader distribution. The MK alternative pathway is the main pathway in both anaerobic and aerobic organisms such as archaeoglobi, thermoproteales, chrysiogenetes, deferribacteres, aquifecales, gemmatimonadetes, chlamydiae, fibrobacteres, acidobacteria, deinococcus-thermus, epsilonproteobacteria and deltaproteobacteria. The MK (and phylloquinone) 'classical' pathway is present in halobacteria, but it was acquired in their common ancestor by lateral gene transfer [231]. The classical pathway is also present in actinobacteria, gammaproteobacteria, cyanobacteria (PQQ and MK), chlorobia and bacteroidetes. Sulfolobales have benzothiophene quinone derivatives instead of typical quinones. Haem biosynthesis distribution: with very few exceptions (five out of 117 archaeal organisms surveyed here), the classical haem pathway is only present in the eubacteria domain. On the contrary, the alternative haem pathway is mostly confined to archaeal haem containing taxa and a few mostly anaerobic eubacteria (thermodesulfobacteria, gemmatimonadetes, clostridia, fibrobacterales and deltaproteobacteria). Interestingly, in some methanomicrobiales (organisms that do not contain cytochromes) genes coding for enzymes involved in the alternative haem pathway are present (*Methanoculleus marisnigri* JR1, *Methanoplanus petrolearius* DSM11571 and *Methanosphaerula palustris*). The role of the genes in these organisms is not clear and they might be involved in F_{430} synthesis instead.

stronger the similarities become, and in terms of their antiquity the sulfate reducers lag not far behind. The circumstances that the acetogens and methanogens share the Wood–Ljungdahl pathway, the most ancient of six CO_2 fixation pathways known [73], that they have cytochrome- and quinone-lacking forms (the only such groups among chemiosmotic prokaryotes

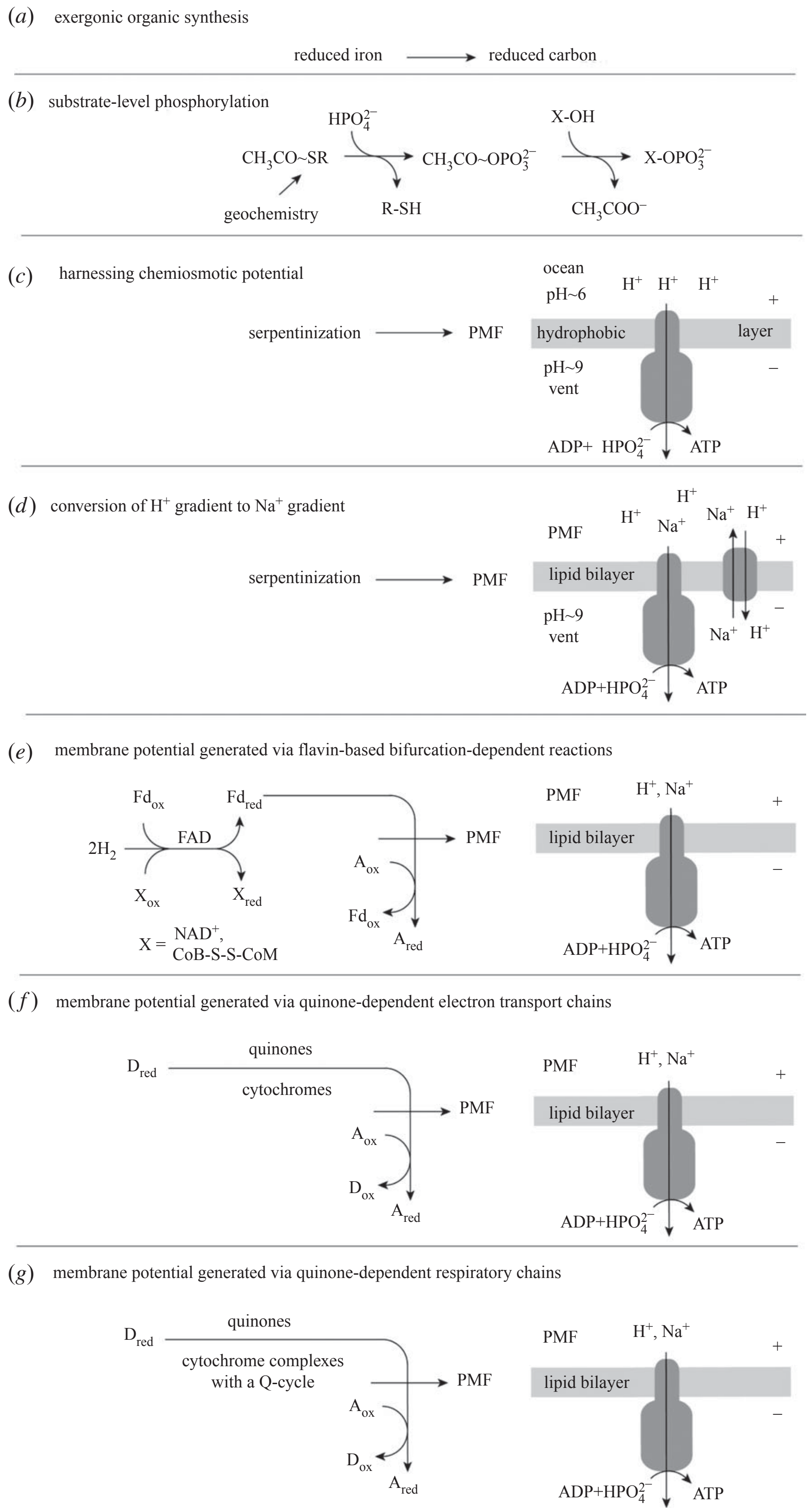


Figure 8. A summary diagram outlining a possible sequence of events in early bioenergetic evolution starting with (*a*) as the most ancient and ending with (*g*) as the most recent. Of course, respiratory chains are, generally speaking, ancient, just not as ancient as the bioenergetic processes in acetogens and methanogens that lack cytochromes or those in sulfate reducers, in our view (see text). The scheme in (*f*) could correspond to the situation in sulfate reducers; the scheme in (*g*) could correspond to the situation in *Paracoccus* [241] or *Rhodobacter* [242].

known), that they are replete with FeS proteins [50,60]) as seen in figure 5, that they live in habitats hardly different from those on the early Earth [11,30], that their energy metabolism is centred around flavin-based electron bifurcation [143]—the soluble precursor of the Q-cycle—and that the energetic pillar of their energy metabolism is low-potential-reduced ferredoxin, are all best interpreted in our view as evidence reflecting their ancient and primordial nature. From the standpoint of energetics, it appears possible that life could have evolved from gases (H_2, CO_2, CO, N_2, H_2S, SO_2) that reacted, with the help of transition metals, at the solid catalyst phase to produce aqueous organic compounds. Current views on the changes in free energy of biological systems, which are necessarily negative [14,29,39], as well as changes in entropy, which are close to zero [245,246], are compatible with that view. The antiquity of anaerobic chemolithoautotrophs seems as evident today as it did 40 years ago [44]. The ubiquity of chemiosmotic coupling as the ultimate source of net energy conservation [151] throughout the microbial world has, by comparison, come as a surprise. In the search for life's start, it is that peculiar energetic attribute of living things—the strict dependence upon chemiosmotic coupling—that makes alkaline hydrothermal vents [9] special. They go a long way towards closing the gap between acetogens, methanogens and the elements on early Earth.

We thank Rolf Thauer, Wolfgang Buckel, Manuela Pereira and Lee Hansen for comments on the draft, Bettina Siebers and Christiane Dahl for helpful discussions, and Re-Young Yu for help with preparing the manuscript. For funding, N.L. thanks the Leverhulme Trust, J.F.A. thanks the Leverhulme Trust for research grant F/07 476/AQ, I.C.P thanks PTDC/QUI-BIQ/100591/2008 funded by FCT, Portugal and FEDER program, and W.F.M. thanks the European Research Council.

References

1. Stüeken EE, Anderson RE, Bowman JS, Brazelton WJ, Colangelo-Lillis J, Goldman AD, Som SM, Baross JA. 2013 Did life originate from a global chemical reactor? *Geobiology* **11**, 101–126. (doi:10.1111/gbi.12025)
2. Haldane JBS. 1929 The origin of life. *Rationalist Ann.* **3**, 3–10.
3. de Duve C. 1994 *Vital dust: life as a cosmic imperative*. New York, NY: Harper Collins.
4. Ricardo A, Carrigan MA, Olcott AN, Benner SA. 2004 Borate minerals stabilize ribose. *Science* **303**, 196. (doi:10.1126/science.1092464)
5. Mulkidjanian AY, Bychkov AY, Dibrova DV, Galperin MY, Koonin EV. 2012 Origin of first cells at terrestrial, anoxic geothermal fields. *Proc. Natl Acad. Sci. USA* **109**, E821–E830. (doi:10.1073/pnas.1117774109)
6. Orgel LE. 2004 Prebiotic adenine revisited: eutectics and photochemistry. *Orig. Life Evol. Biosph.* **34**, 361–369. (doi:10.1023/B:ORIG.0000029882.52156.c2)
7. Brasier MD, Matthewman R, McMahon S, Wacey D. 2011 Pumice as a remarkable substrate for the origin of life. *Astrobiology* **11**, 725–735. (doi:10.1089/ast.2010.0546)
8. Baross JA, Hoffman SE. 1985 Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Orig. Life Evol. B* **15**, 327–345. (doi:10.1007/BF01808177)
9. Russell MJ, Hall AJ. 1997 The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. *J. Geol. Soc. Lond.* **154**, 377–402. (doi:10.1144/gsjgs.154.3.0377)
10. Kelley DS, Baross JA, Delaney JR. 2002 Volcanoes, fluids, and life at mid-ocean ridge spreading centers. *Annu. Rev. Earth Planet. Sci.* **30**, 385–491. (doi:10.1146/annurev.earth.30.091201.141331)
11. Sleep NH, Bird DK, Pope EC. 2011 Serpentinite and the dawn of life. *Phil. Trans. R. Soc. B* **366**, 2857–2869. (doi:10.1098/rstb.2011.0129)
12. Martin W, Baross J, Kelley D, Russell MJ. 2008 Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* **6**, 805–814.
13. Kelley DS *et al.* 2005 A serpentinite-hosted ecosystem: the Lost City hydrothermal field. *Science* **307**, 1428–1434. (doi:10.1126/science.1102556)
14. Amend JP, LaRowe DE, McCollom TM, Shock EL. 2013 The energetics of organic synthesis inside and outside the cell. *Phil. Trans. R. Soc. B* **368**, 20120255. (doi:10.1098/rstb.2012.0255)
15. Allen JF. 2010 Redox homeostasis in the emergence of life. On the constant internal environment of nascent living cells. *J. Cosmol.* **10**, 3362–3373.
16. Amend JP, McCollom TM, Hentscher M, Bach W. 2011 Catabolic and anabolic energy for chemolithoautotrophs in deep-sea hydrothermal systems hosted in different rock types. *Geochim. Cosmochim. Acta* **75**, 5736–5748. (doi:10.1016/j.gca.2011.07.041)
17. Kelley DS *et al.* 2001 An off-axis hydrothermal vent field near the Mid-Atlantic Ridge at 30° N. *Nature* **412**, 145–149. (doi:10.1038/35084000)
18. Ludwig KA, Shen CC, Kelley DS, Cheng H, Edwards RL. 2011 U-Th systematics and Th-230 ages of carbonated chimneys at the lost city hydrothermal field. *Geochim. Cosmochim. Acta* **75**, 1869–1888. (doi:10.1016/j.gca.2011.01.008)
19. Proskurowski G, Lilley MD, Seewald JS, Früh-Green GL, Olson EJ, Lupton JE, Sylva SP, Kelley DS. 2008 Abiogenic hydrocarbon production at Lost City hydrothermal field. *Science* **319**, 604–607. (doi:10.1126/science.1151194)
20. Amend JP, Shock EL. 2001 Energetics of overall metabolic reactions of thermophilic and hyperthermophilic Archaea and Bacteria. *FEMS Microbiol. Rev.* **25**, 175–243. (doi:10.1111/j.1574-6976.2001.tb00576.x)
21. Bach W, Paulick H, Garrido CJ, Ildefonse B, Meurer WP, Humphris SE. 2006 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.* **33**, L13306. (doi:10.1029/2006GL025681)
22. Sleep NH, Meibom A, Fridriksson T, Coleman RG, Bird DK. 2004 H_2-rich fluids from serpentinization: geochemical and biotic implications. *Proc. Natl Acad. Sci. USA* **101**, 12 818–12 823. (doi:10.1073/pnas.0405289101)
23. Russell MJ, Hall AJ, Martin W. 2010 Serpentinization as a source of energy at the origin of life. *Geobiology* **8**, 355–371. (doi:10.1111/j.1472-4669.2010.00249.x)
24. Schulte M, Blake D, Hoehler T, McCollom T. 2006 Serpentinization and its implications for life on the early Earth and Mars. *Astrobiology* **6**, 364–376. (doi:10.1089/ast.2006.6.364)
25. Klein F, Bach W. 2009 Fe–Ni–Co–O–S phase relations in peridotite–seawater interactions. *J. Petrol.* **50**, 37–59. (doi:10.1093/petrology/egn071)
26. Bradley AS, Summons RE. 2010 Multiple origins of methane at the Lost City hydrothermal field. *Earth Planetary Sci. Lett.* **297**, 34–41. (doi:10.1016/j.epsl.2010.05.034)
27. Lang SQ, Butterfield DA, Schulte M, Kelley DS, Lilley MD. 2010 Elevated concentrations of formate, acetate and dissolved organic carbon found at the Lost City hydrothermal field. *Geochim. Cosmochim. Acta* **74**, 941–952. (doi:10.1016/j.gca.2009.10.045)
28. Shock EL, Schulte MD. 1998 Organic synthesis during fluid mixing in hydrothermal systems. *J. Geophys. Res.* **103**, 28 513–28 527. (doi:10.1029/98JE02142)
29. Shock EL, McCollom T, Schulte MD. 1998 The emergence of metabolism from within hydrothermal systems. In *Thermophiles: the keys to molecular evolution and the origin of life* (eds J Wiegel, MWW Adams), pp. 59–76. Washington: Taylor and Francis.
30. Arndt N, Nisbet E. 2012 Processes on the young earth and the habitats of early life. *Annu. Rev. Earth Planetary Sci.* **40**, 521–549. (doi:10.1146/annurev-earth-042711-105316)
31. Trail D, Watson EB, Tailby ND. 2011 The oxidation state of Hadean magmas and implications for early Earth's atmosphere. *Nature* **480**, 79–82. (doi:10.1038/nature10655)
32. Zahnle K, Arndt N, Cockell C, Halliday A, Nisbet E, Selsis F, Sleep NH. 2007 Emergence of a habitable planet. *Planet. Space Sci. Rev.* **129**, 35–78. (doi:10.1007/s11214-007-9225-z)
33. Sleep NH. 2010 The Hadean-Archaean environment. *Cold Spring Harb. Perspect. Biol.* **2**, a002527. (doi:10.1101/cshperspect.a002527)
34. Amend JP, Shock EL. 1998 Energetics of amino acids synthesis in hydrothermal ecosystems. *Science* **281**, 1659–1662. (doi:10.1126/science.281.5383.1659)
35. McCollom TM, Amend JP. 2005 A thermodynamic assessment of energy requirements for biomass synthesis by chemolithoautotrophic microorganisms in oxic and anoxic environments. *Geology* **3**, 135–144.
36. Amend JP, McCollom TM. 2009 Energetics of biomolecule synthesis on early Earth. In *Chemical*

evolution II: from the origins of life to modern society (eds L Zaikowski, JM Friedrich, S Russell Seidel), pp. 63–94. Washington, DC: American Chemical Society.

37. McCollom TM, Shock EL. 1997 Geochemical constraints on chemolithoautotrohic metabolism in seafloor hydrothermal systems. *Geochim. Cosmochim. Acta* **20**, 4375–4391. (doi:10.1016/S0016-7037(97)00241-X)
38. Shock EL. 1990 Geochemical constraints on the origin of organic compounds in hydrothermal systems. *Orig. Life Evol. Biosph.* **20**, 331–367. (doi:10.1007/BF01808115)
39. Thauer RK, Jungermann KK, Decker K. 1977 Energy-conservation in chemotropic anaerobic bacteria. *Bacteriol. Rev.* **41**, 100–180.
40. Thauer RK, Kaster AK, Seedorf H, Buckel W, Hedderich R. 2008 Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat. Rev. Microbiol.* **6**, 579–591. (doi:10.1038/nrmicro1931)
41. Stouthamer AH. 1978 Energy-yielding pathways. In *The bacteria, vol VI: bacterial diversity* (ed. IC Gunsalus). New York, NY: Academic Press.
42. Martin W, Russell MJ. 2007 On the origin of biochemistry at an alkaline hydrothermal vent. *Phil. Trans. R. Soc. B* **367**, 1887–1925. (doi:10.1098/rstb.2006.1881)
43. Lane N, Martin WF. 2012 The origin of membrane bioenergetics. *Cell* **151**, 1406–1416. (doi:10.1016/j.cell.2012.11.050)
44. Decker K, Jungermann K, Thauer RK. 1970 Energy production in anaerobic organisms. *Angew. Chem. Internat. Edit.* **9**, 138–158. (doi:10.1002/anie.197001381)
45. Morowitz HJ. 1968 *Energy flow in biology*. New York, NY: Academic Press.
46. Martin W, Russell M. 2003 On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Phil. Trans. R. Soc. Lond. B* **358**, 59–85. (doi:10.1098/rstb.2002.1183)
47. Whitfield J. 2009 Nascence man. *Nature* **459**, 316–319. (doi:10.1038/459316a)
48. Boyce AJ, Coleman ML, Russell MJ. 1983 Formation of fossil hydrothermal chimneys and mounds from Silvermines, Ireland. *Nature* **306**, 545–550. (doi:10.1038/306545a0)
49. Wächtershäuser G. 1992 Groundworks for an evolutionary biochemistry—the iron–sulfur world. *Prog. Biophys. Mol. Biol.* **58**, 85–201. (doi:10.1016/0079-6107(92)90022-X)
50. Drennan CL, Peters JW. 2003 Surprising cofactors in metalloenzymes. *Curr. Opin. Struct. Biol.* **13**, 220–226. (doi:10.1016/S0959-440X(03)00038-1)
51. Cody GD. 2004 Transition metal sulfides and the origin of metabolism. *Annu. Rev. Earth Planet Sci.* **32**, 569–599. (doi:10.1146/annurev.earth.32.101802.120225)
52. Crabtree RH. 1997 Prebiotic chemistry—where smokers rule. *Science* **276**, 222. (doi:10.1126/science.276.5310.222)
53. Groysman S, Holm RH. 2009 Biomimetic chemistry of iron, nickel, molybdenum, and tungsten in sulfur-ligated protein sites. *Biochemistry* **48**, 2310–2320. (doi:10.1021/bi900044e)
54. Rees DC. 2002 Great metalloclusters in enzymology. *Annu. Rev. Biochem.* **71**, 221–246. (doi:10.1146/annurev.biochem.71.110601.135406)
55. Huber C, Kraus F, Hanzlik M, Eisenreich W, Wächtershäuser G. 2012 Elements of metabolic evolution. *Chem. Eur. J.* **18**, 2063–2080. (doi:10.1002/chem.201102914)
56. Ragsdale SW. 2009 Nickel-based enzyme systems. *J. Biol. Chem.* **284**, 18 571–18 575. (doi:10.1074/jbc.R900020200)
57. Lindahl PA. 2012 Metal–metal bonds in biology. *J. Inorg. Biochem.* **106**, 172–178. (doi:10.1016/j.jinorgbio.2011.08.012)
58. Duffus BR, Hamilton TL, Shepard EM, Boyd ES, Peters JW, Broderick JB. 2012 Radical AdoMet enzymes in complex metal cluster biosynthesis. *Biochim. Biophys. Acta* **1824**, 1254–1263. (doi:10.1016/j.bbapap.2012.01.002)
59. Schwarz G, Mendel RR, Ribbe MW. 2009 Molybdenum cofactors, enzymes and pathways. *Nature* **460**, 839–847. (doi:10.1038/nature08302)
60. Bender G, Pierce E, Hill JA, Darty JE, Ragsdale SW. 2011 Metal centers in the anaerobic microbial metabolism of CO and CO_2. *Metallomics* **3**, 797–815. (doi:10.1039/c1mt00042j)
61. Klein F, Bach W, Jöns N, McCollom T, Moskowitz B, Berquó T. 2009 Iron partitioning and hydrogen generation during serpentinization of abyssal peridotites from 15°N on the Mid-Atlantic Ridge. *Geochim. Cosmochim. Acta* **73**, 6868–6893. (doi:10.1016/j.gca.2009.08.021)
62. He C, Tian G, Liu Z, Feng S. 2010 A mild hydrothermal route to fix carbon dioxide to simple carboxylic acids. *Org. Lett.* **12**, 649–651. (doi:10.1021/ol9025414)
63. Brandes JA, Boctor NZ, Cody GD, Cooper BA, Hazen RM, Yoder Jr HS. 1998 Abiotic nitrogen reduction on the early Earth. *Nature* **395**, 365–367. (doi:10.1038/26450)
64. Dorr M *et al*. 2003 A possible prebiotic formation of ammonia from dinitrogen on iron sulfide surfaces. *Angew. Chem. Int. Ed.* **42**, 1540–1543. (doi:10.1002/anie.200250371)
65. Smirnov A, Hausner D, Laffers R, Strongin DR, Schoonen MAA. 2008 Abiotic ammonium formation in the presence of Ni-Fe metals and alloys and its implications for the Hadean nitrogen cycle. *Geochem. Trans.* **9**, 5. (doi:10.1186/1467-4866-9-5)
66. Yamagata Y, Wanatabe H, Saitoh M, Namba T. 1991 Volcanic production of polyphosphates and its relevance to prebiotic evolution. *Nature* **352**, 516–519. (doi:10.1038/352516a0)
67. Macleod G, McKeown C, Hall AJ, Russell MJ. 1994 Hydrothermal and oceanic pH conditions of possible relevance to the origin of life. *Orig. Life Evol. Biosph.* **24**, 19–41. (doi:10.1007/BF01582037)
68. Holm N, Dumont M, Ivarsson M, Konn C. 2006 Alkaline fluid circulation in ultramafic rocks and formation of nucleotide constituents: a hypothesis. *Geochem. Trans.* **7**, 7. (doi:10.1186/1467-4866-7-7)
69. Heinen W, Lauwers AM. 1996 Organic sulfur compounds resulting from the interaction of iron sulfide, hydrogen sulfide and carbon dioxide in an anaerobic aqueous environment. *Orig. Life Evol. Biosph.* **26**, 131–150. (doi:10.1007/BF01809852)
70. Huber C, Wächtershäuser G. 1997 Activated acetic acid by carbon fixation on (Fe,Ni)S under primordial conditions. *Science* **276**, 245–247. (doi:10.1126/science.276.5310.245)
71. Buckel W, Eggerer H. 1965 On the optical determination of citrate synthase and acetyl-coenzyme A. *Biochem. Z* **343**, 29–43.
72. Fuchs G. 1994 Variations of the acetyl-CoA pathway in diversely related microorganisms that are not acetogens. In *Acetogenesis* (ed. G Drake), pp. 506–538. New York, NY: Chapman and Hall.
73. Fuchs G. 2011 Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu. Rev. Microbiol.* **65**, 631–658. (doi:10.1146/annurev-micro-090110-102801)
74. Wood HG. 1991 Life with CO or CO_2 and H_2 as a source of carbon and energy. *FASEB J.* **5**, 156–163.
75. Ljungdahl LG. 2009 A life with acetogens, thermophiles, and cellulolytic anaerobes. *Annu. Rev. Microbiol.* **63**, 1–25. (doi:10.1146/annurev.micro.091208.073617)
76. Berg IA, Kockelkorn D, Ramos-Vera WH, Say SF, Zarzycki J, Hugler M, Alber BE, Fuchs G. 2010 Autotrophic carbon fixation in archaea. *Nat. Rev. Microbiol.* **8**, 447–460. (doi:10.1038/nrmicro2365)
77. de Duve C. 1991 *Blueprint for a cell - the nature and origin of life*. Burlington, NC: Neil Patterson.
78. Ferry JG, House CH. 2006 The step-wise evolution of early life driven by energy conservation. *Mol. Biol. Evol.* **23**, 1286–1292. (doi:10.1093/molbev/msk014)
79. Wolfe AJ. 2005 The acetate switch. *Microbiol. Mol. Biol. Rev.* **69**, 12–50. (doi:10.1128/MMBR.69.1.12-50.2005)
80. Wolfe AJ. 2010 Physiologically relevant small phosphodonors link metabolism to signal transduction. *Curr. Opin. Microbiol.* **13**, 204–209. (doi:10.1016/j.mib.2010.01.002)
81. Weber AL. 1981 Formation of pyrophosphate, tripolyphosphate, and phosphorylimidazole with the thioester, *N,S*-diacetyl-cysteamine, as the condensing agent. *J. Mol. Evol.* **18**, 24–29. (doi:10.1007/BF01733208)
82. Seewald JS, Zolotov MY, McCollom TM. 2006 Experimental investigation of single carbon compounds under hydrothermal conditions. *Geochim. Cosmochim. Acta* **70**, 446–460. (doi:10.1016/j.gca.2005.09.002)
83. Maupin Furlow JA, Ferry JG. 1996 Analysis of the CO dehydrogenase/acetyl-coenzyme A synthase operon of *Methanosarcina thermophila*. *J. Bacteriol.* **178**, 6849–6856.
84. Rother M, Metcalf WW. 2004 Anaerobic growth of *Methanosarcina acetivorans* C2A on carbon monoxide: an unusual way of life for a methanogenic archaeon. *Proc. Natl Acad. Sci. USA* **101**, 16 929–16 934. (doi:10.1073/pnas.0407486101)

85. Maden BEH. 2000 Tetrahydrofolate and tetrahydromethanopterin compared: functionally distinct carriers in C1 metabolism. *Biochem. J.* **350**, 609–629. (doi:10.1042/0264-6021:3500609)
86. Stryer L. 1975 *Biochemistry*. San Francisco, NC: Freemann.
87. Ownby K, Xu H, White RH. 2005 A *Methanocaldococcus jannaschii* archaeal signature gene encodes for a 5-formaminoimidazole-4-carboxamide-1-β-D-ribofuranosyl 5′-monophosphate synthetase: a new enzyme in purine biosynthesis. *J. Biol. Chem.* **280**, 10 881–10 887. (doi:10.1074/jbc.M413937200)
88. Svetlitchnaia T, Svetlitchnyi V, Meyer O, Dobbek H. 2006 Structural insights into methyltransfer reactions of a corrinoid iron-sulfur protein involved in acetyl-CoA synthesis. *Proc. Natl Acad. Sci. USA* **103**, 14 331–14 336. (doi:10.1073/pnas.0601420103)
89. Huang H, Wang S, Moll J, Thauer RK. 2012 Electron bifurcation involved in the energy metabolism of the acetogenic bacterium *Moorella thermoacetica* growing on glucose or H_2 plus CO_2. *J. Bacteriol.* **194**, 3689–3699. (doi:10.1128/JB.00385-12)
90. Bartoschek S, Vorholt JA, Thauer RK, Geierstanger BH, Griesinger C. 2000 *N*-carboxymethanofuran (carbamate) formation from methanofuran and CO_2 in methanogenic archaea. Thermodynamics and kinetics of the spontaneous reaction. *Eur. J. Biochem.* **267**, 3130–3138. (doi:10.1046/j.1432-1327.2000.01331.x)
91. Hochheimer A, Schmitz RA, Thauer RK, Hedderich R. 1995 The tungsten formylmethanofuran dehydrogenase from *Methanobacterium thermoautotrophicum* contains sequence motifs characteristic for enzymes containing molybdopterin dinucleotide. *Eur. J. Biochem.* **234**, 910–920. (doi:10.1111/j.1432-1033.1995.910_a.x)
92. Kletzin A, Adams MWW. 1996 Tungsten in biological systems. *FEMS Microbiol. Rev.* **18**, 5–63. (doi:10.1111/j.1574-6976.1996.tb00226.x)
93. McCollom TM, Seewald JS. 2006 Carbon isotope composition of organic compounds produced by abiotic synthesis under hydrothermal conditions. *Earth Planet Sci. Lett.* **243**, 74–84. (doi:10.1016/j.epsl.2006.01.027)
94. Morowitz HJ. 1992 *Beginnings of cellular life: metabolism recapitulates biogenesis*. New Haven, CT: Yale University Press.
95. Huber C, Wächtershäuser G. 2003 Primordial reductive amination revisited. *Tetrahedron Lett.* **44**, 1695–1697. (doi:10.1016/S0040-4039(02)02863-0)
96. Huber C, Eisenreich W, Hecht S, Wächtershäuser G. 2003 A possible primordial peptide cycle. *Science* **301**, 938–940. (doi:10.1126/science.1086501)
97. Leman L, Orgel L, Ghadiri MR. 2004 Carbonyl sulfide-mediated prebiotic formation of peptides. *Science* **306**, 283–286. (doi:10.1126/science.1102722)
98. Vaidya N, Manapat ML, Chen IA, Xulvi-Brunet R, Hayden EJ, Lehman N. 2012 Spontaneous network formation among cooperative RNA replicators. *Nature* **491**, 72–77. (doi:10.1038/nature11549)
99. Cantara WA, Bilbille Y, Kim J, Kaiser R, Leszczynska G, Malkiewicz A, Agris PF. 2012 Modifications modulate anticodon loop dynamics and codon recognition of *E. coli* $tRNA^{Arg1,2}$. *J. Mol. Biol.* **416**, 579–597. (doi:10.1016/j.jmb.2011.12.054)
100. Grosjean H, Gupta R, Maxwell ES. 2008 Modified nucleotides in archaeal RNAs. In *Archaea: new models for prokaryotic biology*, pp. 171–196. Norfolk, UK: Caister Academic Press.
101. Giege R, Jühling F, Pütz J, Stadler P, Sauter C, Florentz C. 2012 Structure of transfer RNAs: similarity and variability. *Wiley Interdiscip. Rev. RNA* **3**, 37–61. (doi:10.1002/wrna.103)
102. Grosjean H, de Crecy-Lagard V, Marck C. 2010 Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS Lett.* **584**, 52–264.
103. Limbach PA, Crain PF, McCloskey JA. 1994 Summary: the modified nucleosides of RNA. *Nucleic Acids Res.* **22**, 2183–2196. (doi:10.1093/nar/22.12.2183)
104. Hordijk W, Steel M, Kauffman S. 2012 The structure of autocatalytic sets: evolvability, enablement, and emergence. *Acta Biotheor.* **60**, 379–392. (doi:10.1007/s10441-012-9165-1)
105. Nowak MA, Ohtsuki H. 2008 Prevolutionary dynamics and the origin of evolution. *Proc. Natl Acad. Sci. USA* **105**, 14 924–14 927. (doi:10.1073/pnas.0806714105)
106. Vasas V, Fernando C, Santos M, Kauffman S, Szathmary E. 2012 Evolution before genes. *Biol. Direct* **7**, 1. (doi:10.1186/1745-6150-7-1)
107. Lane N, Allen JF, Martin W. 2010 How did LUCA make a living? Chemiosmosis and the origin of life. *BioEssays* **32**, 271–280. (doi:10.1002/bies.200900131)
108. Peters JW, Williams D. 2012 The origin of life: look up and look down. *Astrobiology* **12**, 1087–1092. (doi:10.1089/ast.2012.0818)
109. Martin WF. 2012 Hydrogen, metals, bifurcating electrons, and proton gradients: the early evolution of biological energy conservation. *FEBS Lett.* **586**, 485–493. (doi:10.1016/j.febslet.2011.09.031)
110. Koonin EV, Novozhilov AS. 2009 Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99–111. (doi:10.1002/iub.146)
111. Koonin EV. 2011 *The logic of chance: the nature and origin of biological evolution*. Upper Saddle River, NJ: FT Press.
112. Copley SD, Smith E, Morowitz HJ. 2005 A mechanism for the association of amino acids with their codons and the origin of the genetic code. *Proc. Natl Acad. Sci. USA* **102**, 4442–4447. (doi:10.1073/pnas.0501049102)
113. Woese CR, Fox GE. 1977 The concept of cellular evolution. *J. Mol. Evol.* **10**, 1–6. (doi:10.1007/BF01796132)
114. Hury J, Nagaswamy U, Larios-Sanz M, Fox GE. 2006 Ribosome origins: the relative age of 23S rRNA domains. *Orig. Life Evol. Biosph.* **364**, 21–29.
115. Novoa EM, Pavon-Eternod M, Pan T, de Pouplana LR. 2012 A role for tRNA modifications in genome structure and codon usage. *Cell* **149**, 202–213. (doi:10.1016/j.cell.2012.01.050)
116. Koonin EV, Martin W. 2005 On the origin of genomes and cells within inorganic compartments. *Trends Genet.* **21**, 647–654. (doi:10.1016/j.tig.2005.09.006)
117. Noller HF. 2012 Evolution of protein synthesis from an RNA world. *Cold Spring Harb. Perspect. Biol.* **4**, a003681. (doi:10.1101/cshperspect.a003681)
118. Wolf YI, Brenner SE, Bash PA, Koonin EV. 1999 Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9**, 17–26.
119. Stock D, Leslie AGW, Walker JE. 1999 Molecular architecture of the rotary motor in ATP synthase. *Science* **286**, 1700–1705. (doi:10.1126/science.286.5445.1700)
120. Chi A, Kemp RG. 2000 The primordial high energy compound: ATP or inorganic pyrophosphate. *J. Biol. Chem.* **275**, 35 677–35 679. (doi:10.1074/jbc.C000581200)
121. Mulkidjanian AY, Makarova KS, Galperin MY, Koonin EV. 2007 Inventing the dynamo machine: on the origin of the F-type and V-type membrane ATPases from membrane RNA/protein translocases. *Nat. Rev. Microbiol.* **5**, 892–899. (doi:10.1038/nrmicro1767)
122. Dagan T, Roettger M, Bryant D, Martin W. 2010 Genome networks root the tree of life between prokaryotic domains. *Genome Biol. Evol.* **2**, 379–392. (doi:10.1093/gbe/evq025)
123. Koga Y, Morii H. 2005 Recent advances in structural research on ether lipids from Archaea including its comparative and physiological aspects. *Biosci. Biotechnol. Biochem.* **69**, 2019–2034. (doi:10.1271/bbb.69.2019)
124. Albers SV, Meyer BH. 2011 The archaeal cell envelope. *Nat. Rev. Microbiol.* **9**, 414–426. (doi:10.1038/nrmicro2576)
125. Leipe DD, Aravind L, Koonin EV. 1999 Did DNA replication evolve twice independently? *Nucleic Acids Res.* **27**, 3389–3401. (doi:10.1093/nar/27.17.3389)
126. Wang J, Dasgupta I, Fox GE. 2009 Many nonuniversal archaeal ribosomal proteins are found in conserved gene clusters. *Archaea* **2**, 241–251. (doi:10.1155/2009/971494)
127. Lecompte O, Ripp R, Thierry JC, Moras D, Poch O. 2002 Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.* **30**, 5382–5390. (doi:10.1093/nar/gkf693)
128. Omer AD, Lowe TM, Russell AG, Ebhardt H, Eddy SR, Dennis PP. 2000 Homologs of small nucleolar RNAs in Archaea. *Science* **288**, 517–522. (doi:10.1126/science.288.5465.517)
129. Jarrell KF, Albers SV. 2012 The archaellum: an old motility structure with a new name. *Trends Microbiol.* **20**, 307–312. (doi:10.1016/j.tim.2012.04.007)
130. Storbeck S, Rolfes S, Raux-Deery E, Warren MJ, Jahn D, Layer G. 2010 A novel pathway for the biosynthesis of heme in Archaea: genome-based bioinformatic predictions and experimental evidence. *Archaea* **2010**, 1–15. (doi:10.1155/2010/175050)
131. Bali S, Lawrence AD, Lobo S, Saraiva LM, Golding BT, Palmer DJ, Howard MJ, Ferguson SJ, Warren MJ.

2011 Molecular hijacking of siroheme for the synthesis of heme and d1 heme. *Proc. Natl Acad. Sci. USA* **108**, 18 260 – 18 265. (doi:10.1073/pnas.1108228108)

132. Daugherty M, Vonstein V, Overbeek R, Osterman A. 2001 Archaeal shikimate kinase, a new member of the GHMP-kinase family. *J. Bacteriol.* **183**, 293 – 300.

133. White RH. 2004 *l*-Aspartate semialdehyde and a 6-deoxy-5-ketohexose 1-phosphate are the precursors to the aromatic amino acids in *Methanocaldococcus jannaschii*. *Biochemistry* **43**, 7618 – 7627. (doi:10.1021/bi0495127)

134. Lange BM, Rujan T, Martin W, Croteau R. 2000 Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proc. Natl Acad. Sci. USA* **97**, 13 172 – 13 177. (doi:10.1073/pnas.240454797)

135. Say RF, Fuchs G. 2010 Fructose 1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. *Nature* **464**, 1077 – 1081. (doi:10.1038/nature08884)

136. Siebers B, Schonheit P. 2005 Unusual pathways and enzymes of central carbohydrate metabolism in Archaea. *Curr. Opin. Microbiol.* **8**, 695 – 705. (doi:10.1016/j.mib.2005.10.014)

137. Grochowski LL, White RH. 2008 Promiscuous anaerobes: new and unconventional metabolism in methanogenic archaea. *Ann. NY Acad. Sci.* **1125**, 190 – 214. (doi:10.1196/annals.1419.001)

138. Grochowski LL, Xu H, White RH. 2009 An iron (II) dependent formamide hydrolase catalyzes the second step in the archaeal biosynthetic pathway to riboflavin and 7,8-didemethyl-8-hydroxy-5-deazariboflavin. *Biochemistry* **48**, 4181 – 4188. (doi:10.1021/bi802341p)

139. De Crecy-Lagard V, Phillips G, Grochowsky LL, El Yacoubi B, Jenney F, Adams MWW, Murzin AG, White RH. 2012 Comparative genomics guided discovery of two missing archaeal enzyme families involved in the biosynthesis of the pterin moiety of tetrahydromethanopterin and tetrahydrofolate. *ACS Chem. Biol.* **7**, 1807 – 1816. (doi:10.1021/cb300342u)

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

141. Valentine DL. 2007 Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nat. Rev. Microbiol.* **5**, 316 – 323. (doi:10.1038/nrmicro1619)

142. Deppenmeier U, Müller V. 2007 Life close to the thermodynamic limit: how methanogenic Archaea conserve energy. *Results Probl. Cell Differ.* **45**, 121 – 152.

143. Buckel W, Thauer RK. 2013 Energy conservation via electron bifurcating ferredoxin reduction and proton/Na$^+$ translocating ferredoxin oxidation. *Biochim. Biophys. Acta* **1827**, 94 – 113. (doi:10.1016/j.bbabio.2012.07.002)

144. Hoehler TM, Jörgensen BB. 2013 Microbial life under extreme energy limitation. *Nat. Rev. Microbiol.* **11**, 83 – 94. (doi:10.1038/nrmicro2939)

145. Kaster AK, Moll J, Parey K, Thauer RK. 2011 Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic Archaea. *Proc. Natl Acad. Sci. USA* **108**, 2981 – 6298. (doi:10.1073/pnas.1016761108)

146. Poehlein A *et al*. 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. (doi:10.1371/journal.pone.0033439)

147. Biegel E, Müller V. 2010 Bacterial Na$^+$-translocating ferredoxin: NAD$^{(+)}$ oxidoreductase. *Proc. Natl Acad. Sci. USA* **107**, 18 138 – 18 142. (doi:10.1073/pnas.1010318107)

148. Schuchmann K, Müller V. 2012 A bacterial electron-bifurcating hydrogenase. *J. Biol. Chem.* **287**, 31 165 – 31 171. (doi:10.1074/jbc.M112.395038)

149. Lipmann F. 1941 Metabolic generation and utilization of phosphate bond energy. *Adv. Enzymol.* **1**, 99 – 162.

150. Wald G. 1962 Life in the second and third periods; or why phosphorus and sulfur for high-energy bonds? In *Horizons in biochemistry* (eds M Kasha, B Pullman), pp. 127 – 142. New York, NY: Academic Press.

151. Mitchell P. 1961 Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism. *Nature* **191**, 144 – 148. (doi:10.1038/191144a0)

152. Mitchell P. 1979 Keilin's respiratory chain concept and its chemiosmotic consequences. *Science* **206**, 1148 – 1159. (doi:10.1126/science.388618)

153. Mitchell P. 1957 The origin of life and the formation and organising functions of natural membranes. In *Proc. 1st Int. Symp. on the Origin of Life on the Earth* (eds AI Oparin, AG Pasynski, AE Braunstein, TE Pavlovskaya), pp. 229 – 234. Moscow, Russia: USSR Academy of Sciences.

154. Smithers GW, Jahansouz H, Kofron JL, Himes RH, Reed GH. 1987 Substrate activity of synthetic formyl phosphate in the reaction catalyzed by formyltetrahydrofolate synthetase. *Biochemistry* **26**, 3943 – 3948. (doi:10.1021/bi00387a030)

155. Mejillano MR, Jahansouz H, Matsunaga TO, Kenyon GL, Himes RH. 1989 Formation and utilization of formyl phosphate by N10-formyltetrahydrofolate synthetase: evidence for formyl phosphate as an intermediate in the reaction. *Biochemistry* **28**, 5136 – 5145. (doi:10.1021/bi00438a034)

156. Celeste LR, Chai GQ, Bielak M, Minor W, Lovelace LL, Lebioda L. 2012 Mechanism of N^{10}-formyltetrahydrofolate synthetase derived from complexes with intermediates and inhibitors. *Prot. Sci.* **21**, 219 – 228. (doi:10.1002/pro.2005)

157. Wächtershäuser G. 2006 From volcanic origins of chemoautotrophic life to bacteria, archaea and eukarya. *Phil. Trans. R. Soc. B* **361**, 1787 – 1806. (doi:10.1098/rstb.2006.1904)

158. Li F, Hinderberger J, Seedorf H, Zhang J, Buckel W, Thauer RK. 2008 Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from *Clostridium kluyveri*. *J. Bacteriol.* **190**, 843 – 850. (doi:10.1128/JB.01417-07)

159. Hedderich R. 2004 Energy-converting [NiFe] hydrogenases from archaea and extremophiles: ancestors of complex I. *J. Bioenerg. Biomembr.* **36**, 65 – 75. (doi:10.1023/B:JOBB.0000019599.43969.33)

160. Eck RV, Dayhoff MO. 1966 Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science* **152**, 363 – 366. (doi:10.1126/science.152.3720.363)

161. Hall DO, Cammack R, Rao KK. 1971 Role of ferredoxins in the origin of life and biological evolution. *Nature* **233**, 136 – 138. (doi:10.1038/233136a0)

162. Ferry FG. 2010 How to make a living by exhaling methane. *Annu. Rev. Microbiol.* **64**, 453 – 473. (doi:10.1146/annurev.micro.112408.134051)

163. Liu Y, Beer LL, Whitman WB. 2012 Methanogens: a window into ancient sulfur metabolism. *Trends Microbiol.* **20**, 251 – 258. (doi:10.1016/j.tim.2012.02.002)

164. Ueno Y, Yamada K, Yoshida N, Maruyama S, Isozaki Y. 2006 Evidence from fluid inclusions for microbial methanogenesis in the early archaean era. *Nature* **440**, 516 – 519. (doi:10.1038/nature04584)

165. Major TA, Burd H, Whitman WB. 2004 Abundance of 4Fe-4S motifs in the genomes of methanogens and other prokaryotes. *FEMS Microbiol. Lett.* **239**, 117 – 123. (doi:10.1016/j.femsle.2004.08.027)

166. Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012 NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130 – D135. (doi:10.1093/nar/gkr1079)

167. Jansen K, Fuch G, Thauer RK. 1985 Autotrophic CO_2 fixation by *Desulfovibrio baarsii*: demonstration of enzyme activities characteristic for the acetyl-CoA pathway. *FEMS Microbiol. Lett.* **28**, 311 – 315.

168. Brysch K, Schneider C, Fuchs G, Widdel F. 1987 Lithoautotrophic growth of sulfate-reducing bacteria, and description of *Desulfobacterium autotrophicum* Gen-Nov, Sp-Nov. *Arch. Microbiol.* **148**, 264 – 274. (doi:10.1007/BF00456703)

169. Stetter KO. 1992 The genus *Archaeoglobus*. In *The prokaryotes,* vol. I (eds A Balows, HG Trüper, M Dworkin, W Harder, K-H Schleifer), pp. 707 – 711, 2nd edn. New York, NY: Springer.

170. Chivian D *et al*. 2008 Environmental genomics reveals a single-species ecosystem deep within earth. *Science* **322**, 275 – 278. (doi:10.1126/science.1155495)

171. Strittmatter AW *et al*. 2009 Genome sequence of *Desulfobacterium autotrophicum* HRM2, a marine sulfate reducer oxidizing organic carbon completely to carbon dioxide. *Environ. Microbiol.* **11**, 1038 – 1055. (doi:10.1111/j.1462-2920.2008.01825.x)

172. Pereira IAC, Ramos AR, Grein F, Marques MC, da Silva SM, Venceslau SS. 2011 A comparative genomic analysis of energy metabolism in sulfate reducing bacteria and archaea. *Front. Microbiol.* **2**, 69. (doi:10.3389/fmicb.2011.0069)

173. Shen Y, Buick R, Canfield DE. 2001 Isotopic evidence for microbial sulphate reduction in the early Archaean era. *Nature* **410**, 77 – 81. (doi:10.1038/35065071)

174. Philippot P, van Zuilen M, Lepot K, Thomazo C, Farquhar J, van Kranendonk MJ. 2007 Early Archaean microorganisms preferred elemental sulfur, not sulfate. *Science* **317**, 1534–1537. (doi:10.1126/science.1145861)

175. Wacey D, Kilburn MR, Saunders M, Cliff J, Brasier MD. 2011 Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nat. Geosci.* **4**, 698–702. (doi:10.1038/ngeo1238)

176. Milucka J *et al*. 2012 Zero-valent sulphur is a key intermediate in marine methane oxidation. *Nature* **491**, 541–546. (doi:10.1038/nature11656)

177. Canfield DE, Rosing MT, Bjerrum C. 2006 Early anaerobic metabolism. *Phil. Trans. R. Soc. B* **361**, 1819–1834. (doi:10.1098/rstb.2006.1906)

178. Farquhar J, Wu NP, Canfield DE, Oduro H. 2010 Connections between sulfur cycle evolution, sulfur isotopes, sediments, and base metal sulfide deposits. *Econ. Geol.* **105**, 509–533. (doi:10.2113/gsecongeo.105.3.509)

179. Grein F, Ramos AR, Venceslau SS, Pereira IAC. 2013 Unifying concepts in anaerobic respiration: insights from dissimilatory sulfur metabolism. *BBA-Bioenerg.* **1827**, 145–160. (doi:10.1016/j.bbabio.2012.09.001)

180. Molitor M, Dahl C, Molitor I, Schafer U, Speich N, Huber R, Deutzmann R, Truper HG. 1998 A dissimilatory sirohaem-sulfite-reductase-type protein from the hyperthermophilic archaeon *Pyrobaculum islandicum*. *Microbiology* **144**, 529–541. (doi:10.1099/00221287-144-2-529)

181. Dhillon A, Goswami S, Riley M, Teske A, Sogin M. 2005 Domain evolution and functional diversification of sulfite reductases. *Astrobiology* **5**, 18–29. (doi:10.1089/ast.2005.5.18)

182. Crane BR, Siegel LM, Getzoff ED. 1995 Sulfite reductase structure at 1.6 Angstrom: evolution and catalysis for reduction of inorganic anions. *Science* **270**, 59–67. (doi:10.1126/science.270.5233.59)

183. Sousa FL, Shavit-Greivink L, Allen JF, Martin WF. 2013 Chlorophyll biosynthesis gene evolution indicates photosystem gene duplication, not photosystem merger, at the origin of oxygenic photosynthesis. *Genome Biol. Evol.* **5**, 200–216. (doi:10.1093/gbe/evs127)

184. Moser DP *et al*. 2005 *Desulfotomaculum* and *Methanobacterium* spp. dominate a 4- to 5-kilometer-deep fault. *Appl. Environ. Microbiol.* **71**, 8773–8783. (doi:10.1128/AEM.71.12.8773-8783.2005)

185. Takami H *et al*. 2012 A deeply branching thermophilic bacterium with an ancient acetyl-CoA pathway dominates a subsurface ecosystem. *PLoS ONE* **7**, e30559. (doi:10.1371/journal.pone.0030559)

186. Kotelnikova S, Pedersen K. 1998 Distribution and activity of methanogens and homoacetogens in deep granitic aquifers at Äspö Hard Rock Laboratory, Sweden. *FEMS Microbiol. Ecol.* **26**, 121–134.

187. Lever MA, Heuer VB, Morono Y, Masui N, Schmidt F, Alperin MJ, Inagaki F, Hinrichs K-U, Teske A. 2010 Acetogenesis in deep subseafloor sediments of the Juan de Fuca Ridge Flank: a synthesis of geochemical, thermodynamic, and gene-based evidence. *Geomicrobiol. J.* **27**, 183–211. (doi:10.1080/01490450903456681)

188. Chapelle FH, O'Neill K, Bradley PM, Methé BA, Ciufo SA, Knobel LL, Lovley DR. 2002 A hydrogen-based subsurface microbial community dominated by methanogens. *Nature* **415**, 312–315. (doi:10.1038/415312a)

189. Susanti D, Mukhopadhyay B. 2012 An intertwined evolutionary history of methanogenic archaea and sulfate reduction. *PLoS ONE* **7**, e45313. (doi:10.1371/journal.pone.0045313)

190. Wang S, Huang H, Kahnt J, Thauer RK. 2013 A reversible electron-bifurcating ferredoxin- and NAD-dependent [FeFe]-hydrogenase (HydABC) in *Moorella thermoacetica*. *J. Bacteriol.* **195**, 1267–1275. (doi:10.1128/JB.02158-12)

191. Mander GJ, Duin EC, Linder D, Stetter KO, Hedderich R. 2002 Purification and characterization of a membrane-bound enzyme complex from the sulfate-reducing archaeon *Archaeoglobus fulgidus* related to heterodisulfide reductase from methanogenic archaea. *Eur. J. Biochem.* **269**, 1895–1904. (doi:10.1046/j.1432-1033.2002.02839.x)

192. Pires RH, Lourenco AIC, Morais F, Teixeira M, Xavier AV, Saraiva LM, Pereira IA. 2003 A novel membrane-bound respiratory complex from *Desulfovibrio desulfuricans* ATCC 27774. *Biochim. Biophys. Acta* **1605**, 67–82. (doi:10.1016/S0005-2728(03)00065-3)

193. Pires RH, Venceslau SS, Morais F, Teixeira M, Xavier AV, Pereira IA. 2006 Characterization of the *Desulfovibrio desulfuricans* ATCC 27774 DsrMKJOP complex-A membrane-bound redox complex involved in the sulfate respiratory pathway. *Biochemistry* **45**, 249–262. (doi:10.1021/bi0515265)

194. Zane GM, Yen HC, Wall JD. 2010 Effect of the deletion of qmoABC and the promoter-distal gene encoding a hypothetical protein on sulfate reduction in *Desulfovibrio vulgaris* Hildenborough. *Appl. Environ. Microbiol.* **76**, 5500–5509. (doi:10.1128/AEM.00691-10)

195. Oliveira TF, Vonrhein C, Matias PM, Venceslau SS, Pereira IAC, Archer M. 2008 The crystal structure of *Desulfovibrio vulgaris* dissimilatory sulfite reductase bound to DsrC provides novel insights into the mechanism of sulfate respiration. *J. Biol. Chem.* **283**, 34 141–34 149. (doi:10.1074/jbc.M805643200)

196. Schut GJ, Adams MW. 2009 The iron-hydrogenase of *Thermotoga maritima* utilizes ferredoxin and NADH synergistically: a new perspective on anaerobic hydrogen production. *J. Bacteriol.* **191**, 4451–4457. (doi:10.1128/JB.01582-08)

197. Ramos AR, Keller KL, Wall JD, Pereira IAC. 2012 The membrane QmoABC complex interacts directly with the dissimilatory adenosine 5′-phosphosulfate reductase in sulfate reducing bacteria. *Front Microbiol.* **3**, 137. (doi:10.3389/fmicb.2012.00137)

198. Wang S, Huang H, Moll J, Thauer RK. 2010 NADP$^+$ reduction with reduced ferredoxin and NADP$^+$ reduction with NADH are coupled via an electron-bifurcating enzyme complex in *Clostridium kluyveri*. *J. Bacteriol.* **192**, 5115–5123. (doi:10.1128/JB.00612-10)

199. Callaghan AV *et al*. 2012 The genome sequence of *Desulfatibacillum alkenivorans* AK-01: a blueprint for anaerobic alkane oxidation. *Environ Microbiol* **14**, 101–113. (doi:10.1111/j.1462-2920.2011.02516.x)

200. Estelmann S, Ramos-Vera WH, Gad'on N, Huber H, Berg IA, Fuchs G. 2011 Carbon dioxide fixation in 'Archaeoglobus lithotrophicus': are there multiple autotrophic pathways? *FEMS Microbiol. Lett.* **319**, 65–72. (doi:10.1111/j.1574-6968.2011.02268.x)

201. Jeanthon C, L'Haridon S, Cueff V, Banta A, Reysenbach AL, Prieur D. 2002 *Thermodesulfobacterium hydrogeniphilum* sp. nov., a thermophilic, chemolithotrophic sulfate-reducing bacterium isolated from a deep-sea hydrothermal vent at Guaymas Basin and emendation of the genus *Thermodesulfobacterium*. *Int. J. Syst. Evol. Microbiol.* **52**, 765–772. (doi:10.1099/ijs.0.02025-0)

202. Marreiros BC, Batista AP, Duarte AMS, Pereira MM. 2013 A missing link between complex I and group 4 membrane-bound [NiFe]-hydrogenases. *BBA Bioenerg.* **1827**, 198–209. (doi:10.1016/j.bbabio.2012.09.012)

203. Schut GJ, Boyd ES, Peters JW, Adams MWW. 2012 The modular respiratory complexes involved in hydrogen and sulfur metabolism by heterotrophic hyperthermophilic archaea and their evolutionary implications. *FEMS Microbiol. Rev.* **37**, 182–203. (doi:10.1111/j.1574-6976.2012.00346.x)

204. Schlegel K, Leone V, Faraldo-Gómez JD, Müller V. 2012 Promiscuous archaeal ATP synthase concurrently coupled to Na$^+$ and H$^+$ translocation. *Proc. Natl Acad. Sci. USA* **109**, 947–952. (doi:10.1073/pnas.1115796109)

205. McMillan DGG *et al*. 2011 A_1A_o-ATP synthase of *Methanobrevibacter ruminantium* couples sodium ions for ATP synthesis under physiological conditions. *J. Biol. Chem.* **286**, 39 882–39 892. (doi:10.1074/jbc.M111.281675)

206. Batista AP, Pereira MM. 2011 Sodium influence on energy transduction by complexes I from *Escherichia coli* and *Paracoccus denitrificans*. *Biochim. Biophys. Acta* **1807**, 286–292. (doi:10.1016/j.bbabio.2010.12.008)

207. Makarova KS, Yutin N, Bell SD, Koonin EV. 2010 Evolution of diverse division and vesicle formation systems in Archaea. *Nat. Rev. Microbiol.* **8**, 731–741. (doi:10.1038/nrmicro2406)

208. Tocheva EI, Matson EG, Morris DM, Moussavi F, Leadbetter JR, Jensen GJ. 2011 Peptidoglycan remodeling and conversion of an inner membrane into an outer membrane during sporulation. *Cell* **146**, 799–812. (doi:10.1016/j.cell.2011.07.029)

209. Miller DA, Choat JH, Clements KD, Angert ER. 2011 The *spoIIE* homolog of *Epulopiscium* sp. Type B is expressed early in intracellular offspring development. *J. Bacteriol.* **193**, 2642–2646. (doi:10.1128/JB.00105-11)

210. Nitschke W, Russell MJ. 2009 Hydrothermal focusing of chemical and chemiosmotic energy, supported by delivery of catalytic Fe, Ni, Mo, Co, S and Se forced life to emerge. *J. Mol. Evol.* **69**, 481–96. (doi:10.1007/s00239-009-9289-3)

211. Schoepp-Cothenet B *et al*. 2013 On the universal core of bioenergetics. *Biochim. Biophys. Acta* **1827**, 79–93. (doi:10.1016/j.bbabio.2012.09.005)

212. Nitschke W, Russell MJ. 2013 Beating the acetyl coenzyme A-pathway to the origin of life. *Phil. Trans. R. Soc. B* **368**, 20120258. (doi:10.1098/rstb.2012.0258)

213. Heijnen JJ, van Dijken JP. 1992 In search of a thermodynamic description of biomass yields for the chemotrophic growth of microorganisms. *Biotechnol. Bioeng.* **39**, 833–858. (doi:10.1002/bit.260390806)

214. Spatzal T, Aksoyoglu M, Zhang LM, Andrade SLA, Schleicher E, Weber S, Rees DC, Einsle O. 2011 Evidence for interstitial carbon in nitrogenase FeMo cofactor. *Science* **334**, 940. (doi:10.1126/science.1214025)

215. Dermer J, Fuchs G. 2012 Molybdoenzyme that catalyzes the anaerobic hydroxylation of a tertiary carbon atom in the side chain of cholesterol. *J. Biol. Chem.* **287**, 36 905–36 916. (doi:10.1074/jbc.M112.407304)

216. Kloer DP, Hagel C, Heider J, Schulz GE. 2006 Crystal structure of ethylbenzene dehydrogenase from *Aromatoleum aromaticum*. *Structure* **14**, 1377–1388. (doi:10.1016/j.str.2006.07.001)

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

218. Bapteste E, Susko E, Leigh J, Ruiz-Trillo I, Bucknam J, Doolittle WF. 2008 Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. *Mol. Biol. Evol.* **25**, 83–91. (doi:10.1093/molbev/msm229)

219. Springer MS, Cleven GC, Madsen O, de Jong WW, Waddell VG, Amrine HM, Stanhope MJ. 1997 Endemic African mammals shake the phylogenetic tree. *Nature* **388**, 61–64. (doi:10.1038/40386)

220. Whitfield JB, Lockhart PJ. 2007 Deciphering ancient rapid radiations. *Trends Ecol. Evol.* **22**, 258–265. (doi:10.1016/j.tree.2007.01.012)

221. Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011 Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl Acad. Sci. USA* **108**, 13 624–13 629. (doi:10.1073/pnas.1110633108)

222. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008 The archaebacterial origin of eukaryotes. *Proc. Natl Acad. Sci. USA* **105**, 20 356–20 361. (doi:10.1073/pnas.0810647105)

223. Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C. 2010 The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nat. Rev. Microbiol.* **8**, 743–752. (doi:10.1038/nrmicro2426)

224. Williams TA, Foster PG, Nye TMW, Cox CJ, Embley TM. 2012 A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. R. Soc. B* **279**, 4870–4879. (doi:10.1098/rspb.2012.1795)

225. Doolittle WF, Bapteste E. 2007 Pattern pluralism and the tree of life hypothesis. *Proc. Natl Acad. Sci. USA* **104**, 2043–2049. (doi:10.1073/pnas.0610699104)

226. Charlebois RL, Doolittle WF. 2004 Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.* **14**, 2469–2477. (doi:10.1101/gr.3024704)

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

228. Creevey CJ, Fitzpatrick DA, Philip GK, Kinsella RJ, O'Connell MJ, Pentony MM, Travers SA, Wilkinson M, McInerney JO. 2004 Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proc. R. Soc. Lond. B* **271**, 2551–2558. (doi:10.1098/rspb.2004.2864)

229. Doolittle WF. 1999 Phylogenetic classification and the universal tree. *Science* **284**, 2124–2128. (doi:10.1126/science.284.5423.2124)

230. Kelly S, Wickstead B, Gull K. 2011 Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proc. R. Soc. B* **278**, 1009–1018. (doi:10.1098/rspb.2010.1427)

231. Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, Deppenmeier U, Martin WF. 2012 Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl Acad. Sci. USA* **109**, 20 537–20 542. (doi:10.1073/pnas.1209119109)

232. Hansmann S, Martin W. 2000 Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes. *Int. J. Syst. Evol. Microbiol.* **50**, 1655–1663. (doi:10.1099/00207713-50-4-1655)

233. Penn O, Privman E, Landan G, Graur D, Pupko T. 2010 An alignment confidence score capturing robustness to guide tree uncertainty. *Mol. Biol. Evol.* **27**, 1759–1767. (doi:10.1093/molbev/msq066)

234. Thurl S, Witke W, Buhrow I, Schäfer W. 1986 Quinones from archaebacteria, II. Different types of quinones from sulphur-dependent archaebacteria. *Biol. Chem. Hoppe Seyler* **367**, 191–197. (doi:10.1515/bchm3.1986.367.1.191)

235. Hiratsuka T, Furihata K, Ishikawa J, Yamashita H, Itoh N, Seto H, Dairi T. 2008 An alternative menaquinone biosynthetic pathway operating in microorganisms. *Science* **321**, 1670–1673. (doi:10.1126/science.1160446)

236. Nowicka B, Kruk J. 2010 Occurrence, biosynthesis and function of isoprenoid quinones. *Biochim. Biophys. Acta* **1797**, 1587–1605. (doi:10.1016/j.bbabio.2010.06.007)

237. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389)

238. Lindsey JS, Ptaszek M, Taniguchi M. 2009 Simple formation of an abiotic porphyrinogen in aqueous solution. *Orig. Life Evol. Biosph.* **39**, 495–515. (doi:10.1007/s11084-009-9168-3)

239. Pierce E *et al*. 2008 The complete genome sequence of *Moorella thermoacetica* (f. *Clostridium thermoaceticum*). *Environ. Microbiol.* **10**, 2550–2573. (doi:10.1111/j.1462-2920.2008.01679.x)

240. Mitchell P. 1975 The protonmotive Q cycle: a general formulation. *FEBS Lett.* **59**, 137–139. (doi:10.1016/0014-5793(75)80359-0)

241. Yang XH, Trumpower BL. 1988 Protonmotive Q-cycle pathway of electron-transfer and energy transduction in the 3-subunit ubiquinol-cytochrome-*c* oxidoreductase complex of *Paracoccus denitrificans*. *J. Biol. Chem.* **263**, 11 962–11 970.

242. Brandt U, Trumpower B. 1994 The protonmotive Q-cycle in mitochondria and bacteria. *Crit. Rev. Biochem. Mol. Biol.* **29**, 165–197. (doi:10.3109/10409239409086800)

243. Allen JF. 2004 Cytochrome b_6f: structure for signalling and vectorial metabolism. *Trends Plant Sci.* **9**, 130–137. (doi:10.1016/j.tplants.2004.01.009)

244. Baradaran R, Berrisford JM, Minhas GS, Sazanov LA. 2013 Crystal structure of the entire respiratory complex I. *Nature* **494**, 443–448. (doi:10.1038/nature11871)

245. Battley EH. 1987 *Energetics of microbial growth*. New York, NY: John Wiley.

246. Hansen LD, Criddle RS, Battley EH. 2009 Biological calorimetry and the thermodynamics of the origination and evolution of life. *Pure Appl. Chem.* **81**, 1843–1855. (doi:10.1351/PAC-CON-08-09-09)

6 Zusammenfassung der Ergebnisse

Analysen zum Ursprung der Eukaryoten

Es wurden 712 eukaryotische Genfamilien gefunden, welche homologe prokaryotische Sequenzen besitzen. Alle 571 Genfamilien, in deren phylogenetischen Bäumen die Eukaryoten eine monophyletische Gruppe bildeten, wurden näher untersucht. In 370 der rekonstruierten phylogenetischen Bäume wurde eine einzelne prokaryotische Gruppe als nächster Nachbar zu den Eukaryoten gefunden. Die drei am häufigsten vorkommenden Nachbargruppen sind die Euryarchaeota, die Crenarchaeota und die α-Proteobakterien. Alle anderen Gruppen wurden in geringerer Anzahl gezählt. Eine Analyse, die zusätzlich Informationen über die funktionelle Einteilung der entsprechenden Genfamilien berücksichtigt, zeigt, dass in an informationsverarbeitung beteiligte Genfamilien besonders archaebakterielle Gruppen als nächste Nachbarn vorkommen und in Genfamilien die am Stoffwechsel beteiligt sind eher Eubakterien dominieren. Hier stechen die α-Proteobakterien besonders hervor. In Genfamilien deren kodierte Proteine an Aminosäurentransport und -verarbeitung beteiligt sind, wurden sie in übermäßiger Zahl als nächste Verwandte zu den Eukaryoten identifiziert. Eine genauere Betrachtung, welche α-Proteobakterien eine besondere Rolle spielen, lieferte keine genauen Ergebnisse. Die beiden Gruppen der Rhodospirillales und der Rhizobiales traten am häufigsten als nächste Verwandte der Eukaryoten auf, wenn auch nicht in einem signifikanten Maße.

Ursachen widersprüchlicher phylogenetischer Signale

Der Vergleich von Daten aus unterschiedlichen taxonomischen Gruppen lieferte keinen Anhaltspunkt dafür, dass phylogenetische Bäume, welche nah verwandte Spezies enthalten, besser konservierte interne Knoten enthalten, als solche bei denen die Diversität zwischen den Spezies größer ist. Alle phylogenetischen Bäume, welche aus verketteten Alignments erstellt wurden, hatten zwar hohe Bootstrapwerte, aber zeigten wenig Übereinstimmung mit den einzelnen Genbäumen, aus deren Daten sie erstellt wurden. Dabei hat die Sequenzlänge der genutzten Proteine als einziger Parameter einen signifikanten Einfluss auf den Ausgang der phylogenetischen Analysen. Die Ähnlichkeit von Bäumen untereinander in einem Datensatz ist stark mit der Sequenzlänge korreliert. Gleiches gilt für Bootstrapwerte. Die untersuchten Eukaryotendatensätze wiesen einen wesentlich geringeren Anteil an Nicht-Übereinstimmung auf, als die Prokaryotendaten. Simulationen bestätigten den Einfluss der Sequenzlänge auf den Ausgang der phylogenetischen Analyse und lieferten Anhaltspunkte dafür, wonach die Konstruktion des Alignments keinen entscheidenden Schritt darstellt, was die Ähnlichkeit der erstellten phylogenetischen Bäume angeht.

Metalloproteine als alternative phylogenetische Marker

Es wurden 1606 prokaryotische Proteome auf das Vorhandensein von Eisen-Schwefel-Cluster (4Fe-4S) Sequenzmotiven hin untersucht. Eine überdurchschnittliche Anzahl dieser Proteine wurde dabei für folgende Gruppen ermittelt: Archaeoglobales, methanogene Archaebakterien, Coriobacteriales, Dehalococcoidetes, Deferribacteres, Clostridia, Fusobacteria, Deltaproteobacteria und Thermodesulfobacteria. Alle diese Gruppen sind entweder methanogen, acetogen oder sulfat-reduzierend. Eine erhöhte Anzahl an 4Fe-4S Motiven war stets mit dem Vorhandensein entweder der Heterodisulfid Redukatse Unterheinheit (HdrABC) oder der Quinon-interagierenden membrangebundenen Oxidoreduktase Untereinheit (QmoABC) verbunden. Beide Proteine sind an der Sulfatreduktion beteiligt, und ihr vorkommen in den beiden erwähnten beiden Gruppen, könnte für einen gemeinsamen Ursprung sprechen.

Schlussbemerkung

Die vorliegenden Ergebnisse der phylogenetischen Analysen weisen darauf hin, dass komplexe evolutionäre Vorgänge, wie die Entstehung der Eukaryoten, nicht nur mit einfachen phylogenetischen Bäumen zu beschreiben sind. Besonders die häufig genutzten verketteten Sequenzdaten, welche dann als einzelne phylogenetische Bäume dargestellt werden, haben eine geringere Glaubwürdigkeit als gewünscht. Gründliche Analysen sollten daher so viele verschiedene Daten, z.B. eine möglichst große Anzahl an Taxa miteinbeziehen und sich nicht auf eine Gruppe bestimmter Gene beschränken. Denn dies kann dazu führen, dass nur ein kleiner Ausschnitt der tatsächlichen evolutionären Vorgänge betrachtet wird.

7 Literaturverzeichnis

Abhishek A, Bavishi A, Bavishi A, Choudhary M (2011) Bacterial genome chimaerism and the origin of mitochondria. *Canadian Journal of Microbiology,* 57:49–61.

Abby SS, Tannier E, Gouy M, Daubin V (2012) Lateral gene transfer as a support for the tree of life. *Proceedings of the National Academy of Science,* 109:4962–4967.

Allen JF (2003) The function of genomes in bioenergetic organelles. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* 358:19–38.

Allsopp A (1969) Phylogenetic relationships of the procaryota and the origin of the eucaryotic cell. *New Phytologist,* 68:591–612.

Alsmark C, Foster PG, Sicheritz–Ponten T, Nakjang S, Embley TM, Hirt RP (2013) Patterns of prokaryotic lateral gene transfer affecting parasitic microbial eukaryotes. *Genome Biology,* 14:R19.

Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO (2012) Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proceedings of the National Academy of Science,* 110:E1594–E1603.

Andersson SGE, Kurland CG (1998) Reductive evolution of resident genomes. *Trends in Microbiology,* 6:263–268.

Andreini C, Bertini I, Cavallaro G, Holliday GL, Thornton JM (2008) Metal ions in biological catalysis: from enzymes databases to general principles. *Journal of Biological Inorganic Chemistry,* 13:1205–1218.

Andreini C, Bertini I, Rosato A (2009) Metalloproteomes: a bioinformatic approach. *Accounts of Chemical Research,* 42:1471–1479.

Atteia A, Adrait A, Bruhiere S, Tardif M, van Lis R, Deusch O, Dagan T, Kuhn L, Gontero B, Martin W, Garin J, Joyard J, Rolland N (2009) A proteomic survey of *Chlamydomonas reinhardtii* mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the α-proteobacterial mitochondrial ancestor. *Molecular Biology and Evolution,* 26:1533–1548.

Baldauf SL, Roger AJ, Wenk–Siefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science,* 290:972–977.

Bapteste E, Boucher Y, Leigh J, Doolittle WF (2004) Phylogenetic reconstruction and lateral gene transfer. *Trends in Microbiology,* 12:406–411.

Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF (2005) Do orthologous gene phylogenies really support tree-thinking? *BMC Evolutionary Biology*, 5:33.

Bapteste E, Susko E, Leigh J, Ruiz-Trillo I, Buckman J, Doolittle WF (2008) Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. *Molecular biology and Evolution,* 25:83–91.

Barros JA, Hoffman SE (1985) Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Origins of Life and Evolution of the Biosphere,* 15: 327–345.

Bergsten J (2005) A review of long-branch attraction. *Cladistics*, 21:163–193.

Brazhnik P, de la Fuente A, Mendes P (2002) Gene networks: how to put the function in genomics. *Trends in Biotechnology,* 20:467–472.

Brochier-Armanet C, Forterre P, Gribaldo S (2011) Phylogeny and evolution of the Archaea: one hundred genomes later. *Current Opinion in Microbiology*, 14:274–281.

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. *Nature Genetics,* 28:281–285.

Bryant D, Moulton V (2003) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution,* 21:255–265.

Castresana J (2007) Topological variation in single-gene phylogenetic trees. *Genome Biology*, 8:216.

Chen I, Dubnau D (2004) DNA uptake during bacterial transformation. *Nature Reviews Microbiology*, 2:241–249.

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

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM (2008) The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences,* 105:20356–20361.

Creevey CJ, Fitzpatrick DA, Philip GK, Kinsella RJ, O'Connell MJ, Pentony MM, Travers SA, Wilkinson M, McInerney JO (2004) Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 271:2551–2558.

Dagan T, Artzy–Randrup Y, Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences,* 105:10039–10044.

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

Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major P, Gould SB, Goremykin VV, Rippka R, Tandeau de Marsac N, Gugger M, Lockhart PJ, Allen JF, Brune I, Maus I, Pühler A, Martin WF (2012) Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biology and Evolution,* 5:31–44.

Dagan T, Roettger M, Bryant D, Martin W (2010) Genome networks root the tree of life between prokaryotic domains. *Genome Biology and Evolution,* 2:379–392.

Daniel RM, Danson MJ (1995) Did primitive microorganisms use nonheme iron proteins in place of NAD/P. *Journal of Molecular Evolution,* 40:559–563.

Darby AC, Cho NH, Fuxelius HH, Westberg J, Andersson SG (2007) Intracellular pathogens go extreme: genome evolution in the Rickettsiales. *Trends in Genetics,* 23:511–520.

Doolittle WF (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics,* 14:307–311.

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

Feagin JE (2000) Mitochondrial genome diversity in parasites. *International Journal for Parasitology,* 30:371–390.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783–791.

Fortunato S (2010) Community detection in graphs. *Physics Reports*, 486:75–174.

Fuchs G (2011) Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annual Review of Microbiology,* 65:631–658.

Galtier N (2007) A model of horizontal gene transfer and the bacterial phylogeny problem. *Systematic Biology,* 56:633–642.

Galtier N, Daubin V (2008) Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 363:4023–4029.

Gaudet P, Williams JG, Fey P, Chisholm RL (2008) An anatomy ontology to represent biological knowledge in *Dictyostelium discoideum*. *BMC Genomics,* 9:130.

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, Konishi J, Denda K, Yoshida M (1989) Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes. *Proceedings of the National Academy of Sciences,* 86:6661–6665.

Gilbert W (1986) Origin of life: the RNA world. *Nature,* 319:618.

Gray MW (1993) Origin and evolution of organelle genomes. *Current Opinion in Genetics & Development,* 3:884–890.

de Grey ADNJ (2005) Forces maintaining organellar genomics: is any as strong as genetic code disparity or hydrophobicity? *Bio Essays,* 27:436–446.

Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C (2010) The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nature Reviews Microbiology*, 8:743–752.

Hackstein JHP, Tjaden J, Huynen M (2006) Mitochondria, hydrogenosomes and mitosomes: products of evolutionary tinkering! *Current Genetics,* 50:225–245.

Hall DO, Cammack R, Rao KK (1971) Role for ferredoxins in the origin of life and biological evolution. *Nature,* 233:136–138.

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogentic trees of duplicated genes. *Proceedings of the National Academy of Sciences,* 86:9355–9359.

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences,* 96:3801–3806.

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

Joyce GF (2002) The antiquity of RNA-based evolution. *Nature,* 418:214–221.

Kanhere A, Vingron M (2009) Horizontal gene transfer in prokaryotes show differential preferences for metabolic and translational genes. *BMC Evolutionary Biology,* 9:9.

Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics,* 9:605–618.

Lang AS, Zhaxybayeva O, Beatty T (2012) Gene transfer agents: phage-like elements of genetic exchange. *Nature Reviews Microbiology*, 10:472–482.

Lang BF, Burger G, Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Grau MW (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature,* 387:493–497.

Layeghifard M, Peres-Neto PR, Makarenkov V (2013) Inferring explicit weighted consensus networks to represent alternative evolutionary histories. *BMC Evolutionary Biology,* 13:274.

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

Loftus B, Anderson I, Davies R, Alsmark UCM, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ, et al. (2005) The genome of the protist parasite *Entamoeba histolytica*. *Nature,* 433:865–868.

Major TA, Burd H, Whitman WB (2004) Abundance of 4Fe-4S motifs in the genome of methanogens and other prokaryotes. *FEMS Microbiology Letters,* 239:117–123.

Makarenkov V, Legendre P (2004) From a phylogenetic tree to a reticulated network. *Journal of Computational Biology,* 11:195–212.

Margulis L (1970) Origin of eukaryotic cells: evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the precambrian earth. New Haven: Yale university press.

Martin W, Russel MJ (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 (2005) The missing link between hydrogenosomes and mitochondria. *Trends in Microbiology,* 13:457–459.

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

McCutcheon JP, Moran NA (2012) Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10:13–26.

McFadden GI (2014) Origin and evolution of plastids and photosynthesis in eukaryotes. *Cold Spring Harbor Perspectives in Biology,* 6:a016105.

Merhej V, Raoult D (2010) Rickettsial evolution in the light of comparative genomics. *Biological Reviews,* 86:379–405.

Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bacterial symbionts. *Annual Review of Genetics,* 42:165–190.

Mulkidjanian AY (2009) On the origin of life in the zinc world: 1. Photosynthesizing, porous edifices built of hydrothermally precipitated zinc sulfide as cradles of life on earth. *Biology Direct,* 4:26.

Nakamura Y, Itoh T, Matsuda H, Gojobori T (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genetics,* 36:760–766.

Newman MEJ (2003) The structure and function of complex networks. *SIAM Review,* 45:167–256.

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

Philippe H, Douady CJ (2003) Horizontal gene transfer and phylogenetics. *Current Opinion in Microbiology,* 6:498–505.

Popa O, Dagan T (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology,* 14:615–623.

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Research,* 21:599–609.

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

Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences,* 95:6239–6244.

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

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang FB, Philippe H (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. *Systematic Biology*, 56:389–399.

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature,* 425:798–804.

Ruvolo M (1997) Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. *Molecular Biology and Evolution,* 14:248–265.

Salichos L, Rokas A (2013) Inferring ancient divergences requieres genes with strong phylogenetic signals. *Nature,* 497:327–331.

Sicheritz–Ponten T, Kurland CG, Andersson SG (1998) A phylogenetic analysis of the cytochrome b and cytochrome c oxidase I genes supports an origin of mitochondria from within the Rickettsiaceae. *Biochimica et Biophysica Acta (BBA)-Bioenergetics,* 1365:545–551.

Strogatz SH (2001) Exploring complex networks. *Nature*, 410:268–276.

Syvanen M (1985) Cross-species gene transfer; implications for a new theory of evolution. *Journal of Theoretical Biology,* 112:333–343.

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

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research,* 29:22–28.

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. *Current Ospinion in Microbiology,* 11:472–477.

Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics,* 5:123–135.

Vasas V, Fernando C, Santos M, Kauffman S, Szathmáry E (2012) Evolution before genes. *Biology Direct,* 7:1.

Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology,* 21:697–700.

Wächtershäuser G (1998) The case for a hyperthermophilic, chemolithoautotrophic origin of life in an iron–sulfur world, in *Thermophiles: the keys to molecular evolution and the origin of life*? , Taylor & Francis Ltd.

Williams KP, Sobral BW, Dickerman AW (2007) A robust species tree for the alphaproteobacteria. *Journal of Bacteriology,* 189:4578–4586.

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences,* 74:5088–5090.

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria and Eucarya. *Proceedings of the National Academy of Sciences,* 87:4576–4579.

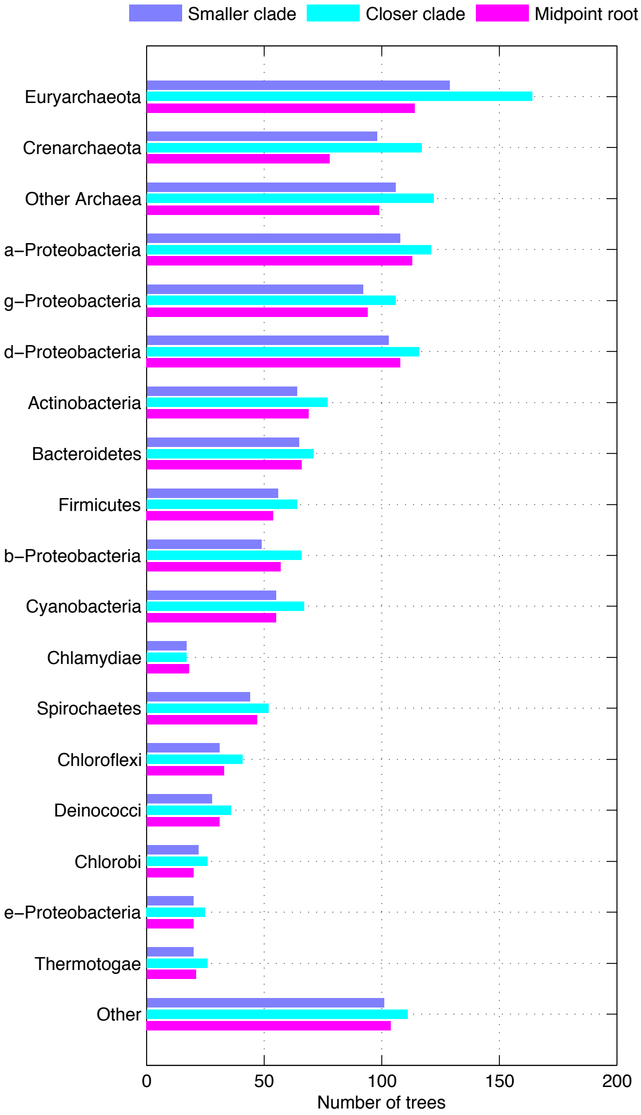
Wozniak RAF, Waldor MK (2010) Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature Reviews Microbiology*, 8:552–563.

Zhaxybayeva O, Lapierre P, Gogarten JP (2004) Genome mosaicism and organismal lineages. *Trends in Genetics,* 20:254–260.

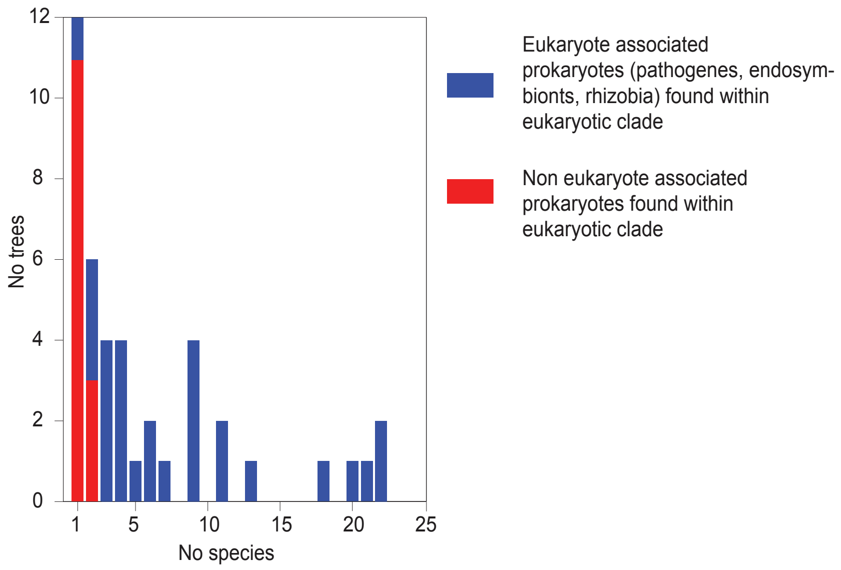
Zinder ND, Lederberg J (1952) Genetic exchange in *Salmonella*. *Journal of Bacteriology*, 64:679–699.

8 Anhang

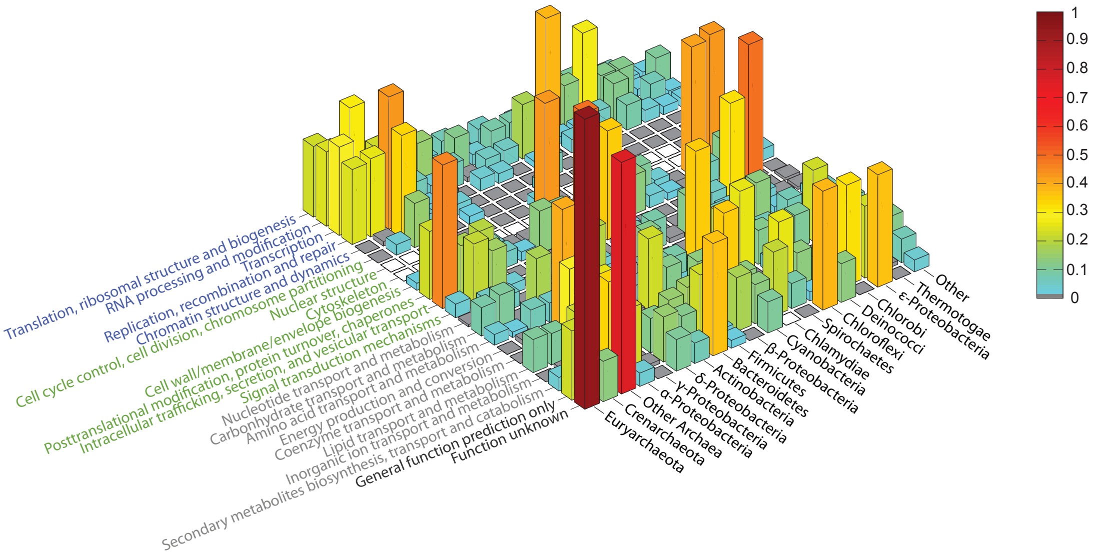
8.1 Anhang zu Thiergart et al. (2012)



Supplementary Figure S1

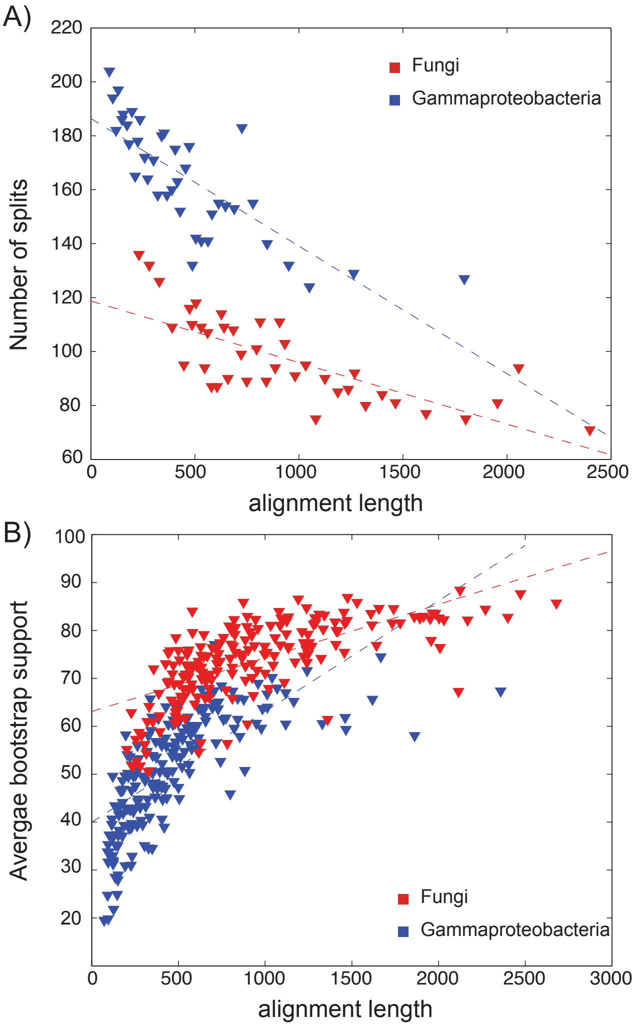


Supplementary Figure S2



Supplementary Figure S3

8.2 Anhang zu Thiergart et al. (2014)



Supplementary Figure S1

VIELEN DANK,

Hiermit möchte ich mich Herzlich bei allen Personen bedanken die mich in den letzten Jahren Unterstützt haben.

Ich danke besonders Prof. Dr. William Martin, Prof. Dr. Tal Dagan und Prof. Dr. Martin Lercher für das Korrigieren dieser Arbeit, für die Betreuung während meiner Promotion, für die Zeit die sie in mich und meine Arbeit investiert haben und für die Möglichkeit überhaupt diesen Weg einzuschlagen.

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der “Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf” erstellt worden ist.

……………………………………………………………………..

(Thorsten Thiergart)