Author:
Moody, Edmund R R

Title:
Protein evolution and the early history of life

Protein Evolution and the Early History of Life

Edmund R. R. Moody

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Life Sciences

School of Biological Sciences
University of Bristol
United Kingdom
February 2022

Word Count: 37200

Abstract

Throughout human history philosophers have tried to understand the inter-relatedness of the vast array of organisms found on Earth. The study of phylogenetics has provided a window into which we can understand how life has evolved on this planet, and analyses have shown that life is composed of two primary domains, Archaea and Bacteria. Traditional phylogenomic analyses have found the evolutionary distance between Archaea and Bacteria to be a significant one. However, a recent analysis suggests that perhaps they were much more closely related than previously thought. Here, I examine an inferred set of universal marker genes suggesting a closer evolutionary distance between the two primary domains. I compare this set of markers with those of previous analyses, and examine the verticality and evolutionary history of the component genes. I also infer a novel set of markers based on these results to infer a new tree of life. I find the distance between Archaea and Bacteria is a great one; however, I find it is susceptible to substitutional saturation and the use of inappropriate models of molecular evolution. I also examine the molecular evolution of multiple proteins found in eukaryotes: a zinc-finger protein involved in blood-cell differentiation, a protein complex involved in protein recycling, and a motor protein which transports cellular cargo. As a whole this work demonstrates how we can use phylogenetic analysis to answer questions about ancient and more recent branches in the tree of life.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Edmund R. R. Moody

June 6, 2022

Acknowledgements

First and foremost to Tom A. Williams, whom without which none of this would have been possible. Tom has of course served as an exemplary supervisor, teacher, and mentor throughout my time at the University of Bristol, but he has also been a dear friend.

I would also like to thank my other friends: whether they be students, staff or just one of the chaps. From the passionate and detailed discussions about less scientific matters during work, to passionate and detailed discussions about science at the pub. Let me also thank all my collaborators, whom without which scientific projects such as these just would not be possible A special thanks to Tara, Anja and Nina, who were instrumental in my first first-author paper; and to Charlie, Kim and Peter for patiently reading through the entire thesis and pointing out mistakes. I am eternally grateful to all of you.

Lastly, I thank my family for their support and the reassurance that I always had someone looking out for me and a place to call home. Mum, always making sure I am happy, healthy and have plenty of food, my mental and physical well-being owe you a great debt. Dad, you always bought me books and patiently answered my endless questions about the universe. Without you, I would not be a scientist. You have had to endure my ramblings about the evolution of monsters (both real and imagined) as a small child, to my ramblings about the evolution of all life today.

And to Kim, who has helped me weather the storm.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

LUCA	Last universal common ancestor
LBCA	Last bacterial common ancestor
LACA	Last archaeal common ancestor
LECA	Last eukaryotic common ancestor
HGT	Horizontal gene transfer
ML	Maximum likelihood
AB	Archaeal-bacterial
ZNF	Zinc-finger protein
SNX	Sorting-nexin protein
KLC	Kinesin-light chain
KHC	Kinesin-heavy chain
ΔLL	Difference in log-likelihood
MAD	Minimum ancestral deviation
PVC	Planctomycetes, Verrucomicrobia, and Chlamydiae
CPR	Candidate phyla raditation
FCB	Fibrobacteres, Chlorobiota, and Bacteroidota
DPANN	Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota
TACK	Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota
LG	Le and Guascuel
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool

AIC	Akaike information criterion
BIC	Bayesian information criterion
AICc	Corrected Akaike information criterion
COG	Clusters of Orthologous Genes

1 INTRODUCTION

Author's contribution

This chapter was written by Edmund. R. R. Moody in its entirety.

1.1 Abstract

Understanding how life began and the relationships of life on Earth is a philosophical question which has intrigued human beings since the beginning. In this introduction, I discuss the history of evolutionary thought and its associated scientific techniques. I introduce Archaea and Bacteria, two of the primary domains of life, and our current and previous understanding of these organisms and how they fit into the tree of life. I also review and discuss the history of the tree of life itself and whether or not such a tree even exists.

1.2 Early understanding of the relatedness of all life

One of the oldest stories we have evidence for relates to the creation of life. On the the first stone tablet of the ancient Mesopotamian creation myth, the Enuma Elish, inorganic liquid separated into two forms of life, which went on to populate the entire Earth[1]. As old as this creation story is, it is perhaps not so far removed from our current understanding[2] of how life first evolved on this planet, and the nature of the relationship between all known living organisms.

The ancient Greek philosophers Aristotle and Plato both helped lay the foundations and attempted the ranking and grouping of various organisms. This was not based on any modern evolutionary understanding, but it is a seed from which the tree of life grew. They placed life within a hierarchy based on their physical characteristics, i.e. humans above animals, animals above plants, plants above minerals. Interestingly, Aristotle also developed an early clustering method for different species based on a combination of morphological and ecological characteristics[3]. These ideas were later merged into the *Scala Naturae* by Christian philosophers known as 'The Great Chain of Being'. In addition to physical organisms and objects, this also placed divine beings such as angels in the chain, above humans, with God at the top. In one sense, an argument could be made that all things were related to one another in-so-much as they were created by God[4].

During the 18th century, Linnaeus's *Systema Naturae*[5] became the system upon which modern taxonomy is based. This system was originally also centered on shared morphological characteristics. Jean Baptiste Lamarck later helped guide biological thought to the idea of progression and change throughout time[6]. Lamarck[6] presented the idea of 'transmutation': animals slowly modify their phenotypic characteristics throughout their lifetime as a response to the environment. Although Lamarck did explore the idea of a progression of organisms, this was more on an individualistic level, e.g. an individual animal that spent all its time in water would develop webbed feet. There is no evidence to suggest Lamarck believed that all life was somehow related, and the transmutation of species hypothesis fit within the strongly religious framework at the time.

The first concept of the biological relatedness of all life is from Charles Darwin[7] (and independently from Alfred Russel Wallace[8]). In the Origin of Species, the formal hypothesis of a common ancestor is proposed:

> "Nevertheless all living things have much in common, in their chemical composition, their germinal vesicles, their cellular structure, and their laws of growth and reproduction. We see this even in so trifling a circumstance as that the same poison often similarly affects plants and animals; or that the poison secreted by the gall-fly produces monstrous growths on the wild rose or oak-tree. Therefore I should infer from analogy that probably all the organic beings which have ever lived on this earth have descended from some one primordial form, into which life was first breathed." Darwin[7]

It is important to note here that Darwin is discussing what are now two distinct modern concepts as one single idea. The origin of life — the 'primordial form' into which life was first breathed — is a distinct entity from the hypothetical Last Universal Common Ancestor (LUCA); which is the last (youngest) ancestor shared between all extant (crown-group) life. It is also interesting and important to note that Darwin had no physical evidence that life was related in this way, and at this point in time the entirety of the field of genetics was hidden in an obscure Austrian botanical journal article written by Gregor Mendel[9].

Assuming a Darwinian mode of evolution, LUCA is the root of the neontological 'tree of life', it is not necessarily the origin of life. Life may have originated and gone extinct several times over without changing our current hypothesis of LUCA. It is also possible that life could have had a successful 3 Ga of evolution with many unknown branches and domains until all but two lineages (Archaea and Bacteria) went extinct. LUCA is the last common ancestor of those two lineages, and therefore it is the last universal common ancestor (of extant life), so anything before LUCA would be classed as stem-group life. Another possibility, albeit unlikely, is that LUCA and the origin of life are actually one and the same thing. However, due to the gradient at which objects transition from non-life to living material, and an arguable necessity for LUCA to be a fully formed cell i.e. the last common ancestor of the two groups of cellular organisms, it is likely that life in some way, shape or form most likely did exist before LUCA, however briefly.

Darwin's non-bifurcating tree diagram (famously, the only illustration in the Origin of Species, Figure 1i) helped synonymise the idea of a 'tree' with evolutionary thought[7]. If a single tree of life exists, then so does LUCA. Haeckel[10] visualised various trees of life, one of his more famous examples being 'The Pedigree of Man'. At the base of Haeckel's tree is *Monera* where bacteria were initially placed, with *Homo sapiens* at the top, suggesting the evolutionary progression of life from single-celled ancestors to the most complex 'higher' organism. This was developed in more detail in his earlier work where he devised the term 'phylogeny'[11], and divided life into three major groups — plants, protists and animals — based on morphological data[12]. The common ancestor of all three groups being the primordial *Moneres* which evolved into all other organisms (Figure 1ii).

1.3 Bacteria

The main constituent of Monera, Bacteria, had been discovered several hundred years earlier in 1676 by Antony van Leeuwenhoek[13] who had managed to create the first light microscopes effective enough to observe organisms that small. Monera (prokaryotes) was later divided[14] into several groups dependent on the gram staining of the organisms Gracilicutes (gram-negative), Firmicutes (gram-positive) and Mollicutes (no cell wall). They also included a group of organisms which possessed a cell wall, but did not contain the petptidoglycan polymer and tentatively placed them into Mendocutes, which would later become known as Archaea[14].

Extant Bacteria inhabit a staggering range of environments on Earth: a large range of temperatures, radiation & salt levels, pH, and pressure[15]. Bacteria also live on and within other organisms and are vital for their host's regular function[16]. They exhibit a wide-range of morphologies: usually spherical cocci or rod-like bacilli as well an array of of irregular and interesting shapes[17]. They also can also cluster as chains, botryoidal patterns or as biofilms and microbial mats[18,19]. Bacterial metabolism is just as varied, with a range of heterotrophic and autotrophic lifestyles, including photosynthetic Cyanobacteria[20], nitrate-reducing Nitrospirae[21] and parasitic Chlamydiae[22].

Today, molecular phylogenetics has helped place most Bacteria into one of two major clades. These are the Terrabacteria, comprising Actinobacteria, Firmicutes,

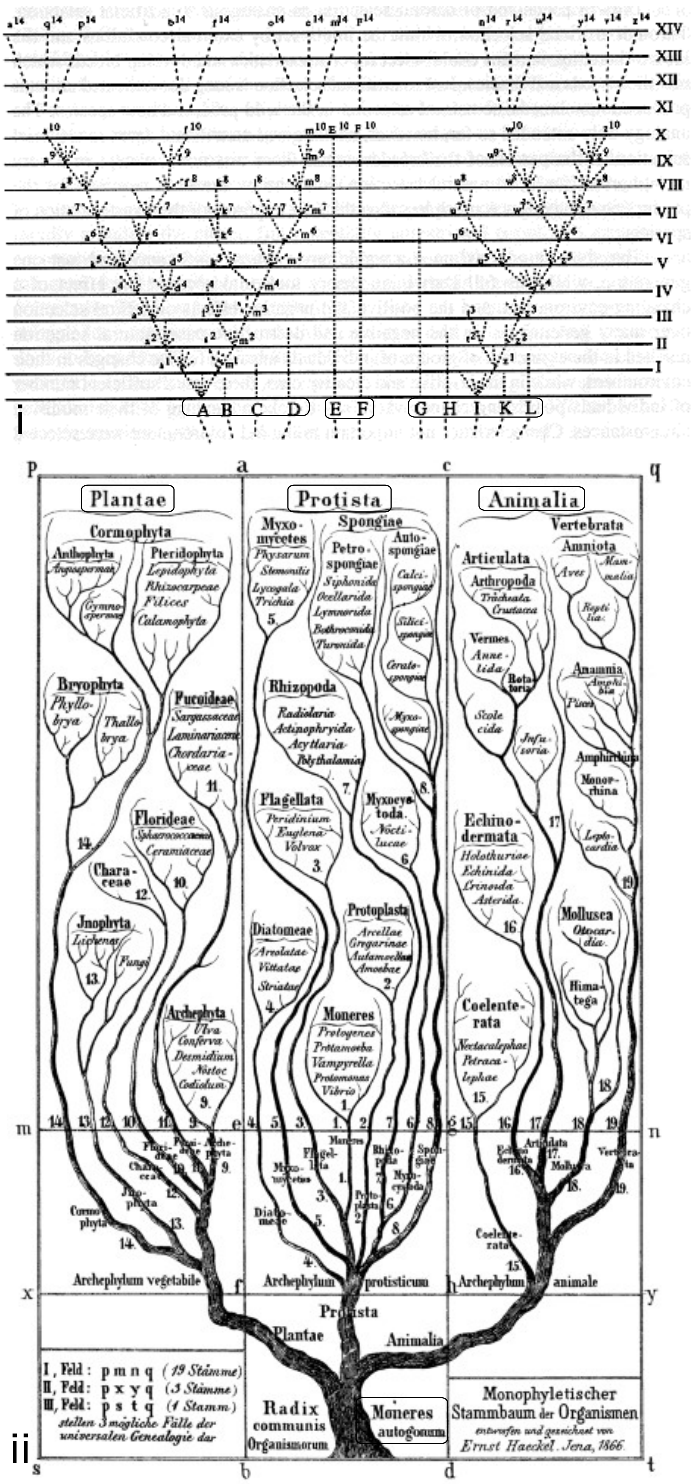

Figure 1 – i: Darwin's[7] tree figure from the origin of species, letters A to L represent different species belonging to the same genus, with three distinct groups all sharing a closer common ancestor with each other. Group one: A,B,C & D. Group two: E & F. Group three: G,H,I,K & L. The intervals between the dotted lines represent a number of generations, Darwin[7] suggests one thousand. Continued on next page.

Figure 1 – Continued. After one thousand generations, species A undergoes a divergence, with two distinct groups emerging, a1 and m1, which then go onto diverge with varying success until we get to the 14,000th generation, where only A, F and I are shown to have survived with varying success. ii: Haeckel's[12] original phylogeny dividing life up into three groups: Plantae, Protista and Animalia. Each lineage is numbered. Crown-groups are titled 'Archephylums'. All extant life can be traced back to a common ancestry with 'Moneres' or prokaryotes. i is adapted from Darwin[7], ii is adapted from Haeckel[12].

Cyanobacteria, Chloroflexi, and Deinococcus-Thermus[23,24] and the Gracilicutes, including Proteobacteria, PVC (Planctomycetes, Verrucomicrobia, and Chlamydiae), FCB (Fibrobacteres, Chlorobiota, and Bacteroidota), Spirochaetae and Acidobacteria[23,24]. Recently, a large radiation of bacterial phyla was discovered from metagenomic and single-cell sequencing known informally as the CPR (candidate phyla radiation) group[25,26] (see Figure 2).

Topical questions surrounding the bacterial tree include the gene content, metabolism and ecology of the last common bacterial ancestor (LBCA). Was it diderm (gram-negative) or monoderm (gram-positive)? Should CPR be placed within the Terrabacteria as some have suggested[24] or are they sister to all other bacterial groups[27,28,29,30]? Strongly related to this is the question of the bacterial root, does it lie between Terrabacteria and Gracilicutes as previously suggested[24], or within another bacterial clade as others have inferred[27,28,29,30,31,32]?

Mitochondria found in eukaryotes were hypothesised to share a common ancestor with other bacteria[33]. This was later confirmed through molecular phylogenetic analyses to fall within Alphaproteobacteria[34], though the exact placement of the ancestral mitochondrion or protomitochondrion remains a topic of active research[34,35,36].

1.4 Archaea

Archaea were first identified as being separated from Bacteria by Woese and Fox[37] using an early distanced-based clustering phylogenetic method using ribosomal RNA (rRNA) sequences. At the time the only sampled Archaea were methanogens and the name Archaebacteria was given. It was proposed that Bacteria was divided into

Eubacteria and *Archaeabacteria*[37].

Although they were first known mainly as extremophiles[38], representatives from Archaea can now be found in many different environments such as oceans[39], soil[40] and as an important constituent of the human gut microbiome[41]. Surprisingly, as of yet there are no known pathogenic Archaea[42].

Aside from on an evolutionary level[43], we know Archaea are different to Bacteria in various and important biochemical ways. Archaea possess ether-linked lipid membranes as opposed to bacterial ester-linked lipids, and as a result a whole host of different biochemical machinery associated with synthesising and utilizing such molecules[44]. The archaeal cell wall never contains peptidoglycan, unlike many bacterial cell-walls, although Archaea can contain a functional homologue pseudomurein[37,45]. The structure of RNA polymerase (an essential component of gene transcription) in Archaea is more similar to eukaryotic RNA polymerases than the bacterial orthologue[46].

Archaea were initially divided into two major groups: Crenarchaeota and Euryarchaeota[43]. Increased sampling and genomic sequencing has also resulted in the discovery of additional new groups. The Euryarchaeota clade has remained relatively intact but Crenarchaeota is now contained within the TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota) which is sister to the Asgard superphylum[39] (see Figure 2).

More recently, a new group of Archaea has been proposed: DPANN[47,48], composed of Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota. It is hard to shake the parallels to the discovery of CPR and this begs the question — how many more times will large unknown groups of taxa be discovered which need to be grafted onto the tree of life? The placement of DPANN, similar to CPR, is also an area of active research, with phylogenetic analyses placing DPANN within Euryarchaeota[49,50,51] and as sister to the rest of Archaea[48,52,53].

Several analyses place either Asgard archaea or a member of Asgard archaea as the closest relatives to eukaryotes (in terms of eukaryotic nuclear DNA)[39,54,55]. Excitingly, after an experiment spanning over a decade, a member of the Asgard clade was cultured *Candidatus* Prometheoarchaeum syntrophicum, the unique morphology of which was used to propose a hypothetical scenario for the eukaryogenesis event[56].

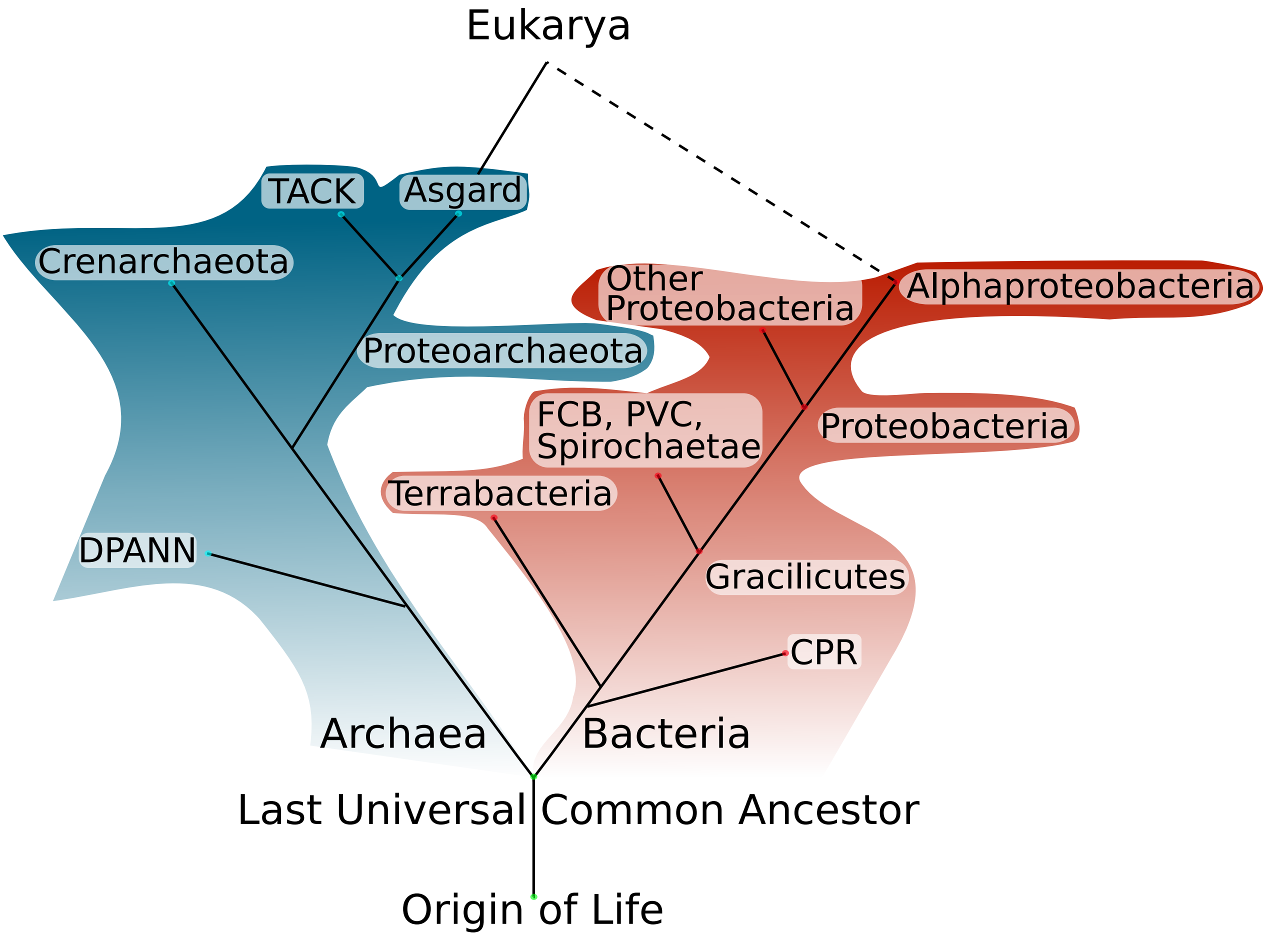

Figure 2 – A cartoon visualising the potential branch between the origin of life and the last universal common ancestor, the division of the two primary domains of life, and a member of the Asgard as a potential host for the common ancestor of eukaryotes, with the alphaproteobacterial mitochondrial ancestor. It should be noted that there are disagreements about the placements of some of these groups, such as DPANN and CPR.

1.5 Phylogenetics

Mendel's discovery of heritable traits through 'factors' (genes) in the 1860s[9], the subsequent discovery of DNA[57] as a molecule responsible for determining genetic inheritance, and finally the understanding of the central dogma of DNA to RNA to proteins[58,59,60,61] provided us with a mechanism for evolution. By building upon and integrating this knowledge, the until-then mainly morphology-based systematic approach and the new field of cladistics[62] begat an expansion to incorporate the ever growing wealth of molecular data[63].

In order to infer a phylogenetic tree, a multiple sequence alignment is required. This is arguably one of the most important steps in the whole tree-building process because if this alignment is incorrect, then anything later derived from it will also be incorrect[64]. A multiple sequence alignment is a hypothesis of the homology of each nucleotide base or amino acid residue for a given gene or protein. Distance-based methods of phylogeny estimation such as unweighted pair group method with arithmetic mean (UPGMA)[65] or neighbour-joining (NJ)[66], use a model of substitution to calculate an evolutionary distance matrix. These models range greatly in their complexity. The simplest model: Jukes-Cantor (JC)[67] assumes all rates of substitution between two nucleotides are equal. The most complex: the General-time-reversible (GTR) model allows for a different instantaneous rate of substitution between each nucleotide as well as different base frequencies[68]. Similar substitution models such as the Dayhoff[69], JTT (Jones-Taylor-Thornton[70]) and LG (Le and Gascuel[71]) matrices exist for protein datasets.

After distance calculation, a clustering algorithm is deployed to group sequences with the least distance between them. These methods are computationally efficient and modern software like IQ-Tree[72] often uses them for a quickly-generated starting tree before more complex analyses. Most modern molecular phylogenetics uses a probabilistic framework; either through a maximum likelihood (ML)[73] or a Bayesian inference[74] implementation with variations of the previously mentioned substitution models. The goal of this is to calculate the probability of the tree with fixed parameters (ML) or a distribution of parameters and a posterior probability distribution of trees (Bayesian inference).

Genome sequencing technology has also advanced, we can now quickly and (relatively) inexpensively sequence the genome of entire organisms. The availability of genomes, transcriptomes, and proteomes have increased massively over the last few decades[75]. For species-tree inference, phylogenetic analyses have transitioned from a focus on individual genes to phylogenomics: using multiple genes from each organism.

The increase in computing power has enabled analyses to make use of more complex models that better capture our understanding of the evolutionary process. Some of these are models that allow for unequal rates across sites through the inclusion of a discretised γ-distribution of substitution rates[76]. In addition to substitution rate variation, we also know that the substitution process itself is not homogeneous across the whole sequence, this is especially noticeable if using a concatenation of multiple proteins in a supermatrix analysis. Certain residues and nucleotides can be functionally constrained for structural or evolutionary reasons. One solution to this is the CAT-GTR model, which allows for a theoretically infinite variation of substitution matrices to be selected from[77] as a parameter of the model. These issues are reviewed in depth in Williams et al.[78].

CAT-GTR would seem to be the gold standard for phylogenomic inference, but it is not (as is the case with all models) without its limitations. On current computer hardware, analyses of more than a few hundred taxa[79] are intractable. Depending on the length of the alignment, the convergence time may be on a scale of months, rather than days or hours. For site- and taxon-rich alignments, one solution is a maximum-likelihood implementation of the CAT model. Although this is limited in the maximum number of replacement profiles, it is still an approximation which performs better in terms of model fit, and in overall likelihood for saturated data (sites which have undergone many substitutions over their evolution) than other maximum-likelihood models[80]. In theory, this makes these types of models more appropriate for deep-time questions such as the tree of life[54]. Other solutions could be to reduce the number of taxa in the alignment[39]. Or to re-code the alignment reducing the number of estimated parameters[81] (or also as a method to avoid composition bias[82]).

Although our models have improved over time, there are still many fundamental problems with tree inference. The major limitation to phylogenetic modelling is that unless we are dealing with organisms for which we have witnessed or know the true relationships, we can never measure the true accuracy of a given tree hypothesis. This can be mitigated by a number of different support methods, such as randomly resampling the data with replacement, known as the 'bootstrap'[83], or through topological tests[84,85,86]. A suite of model tests also allow us some further confidence behind the models we use[87,88,89,90,91].

Horizontal gene transfer (HGT) poses another issue for phylogeneticists. Originally shown through experimentation[59,92,93,94], it took several decades before it was proposed that HGT (also called lateral gene transfer) could be an important evolutionary force across the tree of life[95] and subsequently proved to be so[96,97,98,99]. There are various mechanisms by which prokaryotes can transfer genes laterally: through transformation (the uptake and expression of genetic material through the cell membrane), transduction (virally), or conjugation (a non-sexual method of transferring material directly from one organism to another)[100]. On a practical level, this can result in conflicting evolutionary histories for different genes. This is a problem for concatenation analysis, where multiple genes contribute to branch length and topology inferences estimated under the same model.

The detection of paralogous genes can also generate issues for species tree estimation, for example: if the paralogues are erroneously included as orthologues then this will most likely lead to the inference of an inaccurate species tree[101]. Further complications arise from gene duplication events and subsequent losses. A loss of the gene removes our ability to infer the evolution of said gene, but if these losses occur after a gene duplication, then rather than just a lack of signal, we may infer a phylogeny which is in direct conflict with that of the species tree. This is known as hidden paralogy.

Taxon sampling is another issue; different sets of representative sequences can lead to the inference of incompatible species trees, and taxon sampling has been shown to be one source of conflicting results[102,103,104]. Sequence saturation can also cause a decay in the phylogenetic signal of sequences[105,106], which in turn can cause the common issue of 'long-branch attraction': long branches with high substitution rates grouping together due to convergence, but several studies suggest that this is less of a problem for trees inferred under ML or Bayesian inference[107,108].

1.6 The tree of life

The purpose of this section is to place my work into the historical context of research on the tree of life and the evolutionary relationships among different groups. In the interest of brevity, I use only the current names for each of the groups, although many of their names have changed through time as ideas about early evolution were proposed and subsequently modified or abandoned. The taxonomic history is not the focus of this work and I suggest that synonymy lists for prokaryotes are probably worthy of their own theses in and of themselves.

1.6.1 Tripartite tree vs eocyte tree

Just before the the discovery of Archaea, Woese and Fox[109] introduced the idea of a 'progenote'. This is a hypothetical organism, fulfilling the same role as LUCA with an additional caveat, which is that the progenote was not a fully functioning cellular organism of the type we know today. It had not evolved the full suite of ribosomal and protein machinery which links phenotype (an organism's observable traits) to genotype (an organism's genes). The suggestion was this progenote underwent rapid divergent evolution into two lineages, the eukaryotes and the prokaryotes. The difference between the level of complexity between the progenote and these prokaryotic Bacteria would be similar to the structural differences we see between prokaryotes and eukaryotes.

The discovery of an additional main lineage of life[37] necessitated a re-imagining of the status quo, which was this division between prokaryotes and eukaryotes. Evidence from small-subunit rRNA showed that this new group, Archaea (introduced originally as *Archaebacteria*[37]), were just as distantly related to Bacteria as they were to eukaryotes. Therefore the tree of life was split into three (Figure 3 i), with each lineage independently evolving from this hypothetical non-prokaryotic progenitor. The endosymbiosis event leading to to the formation of eukaryotes[110] and their subsequent organelles was thought to be independent events, i.e. the nuclear genome of eukaryotes has its own lineage separate to that of the organelles. Woese and Fox[37] acknowledged the distance-based phylogeny method used did not account for multiple substitutions and as such did not give any real indication of time, and thus it is impossible to say if all three clades branched at the same time but hypothesised:

> "One of the three may represent a far earlier bifurcation than the other two, making there in effect only two urkingdoms." Woese and Fox[37]

Lake et al.[111] built upon this using the structure of the ribosome to find two (rather than three) distinct groups: eukaryotes and eocytes (now known as Crenarchaeota) as one group, and Archaebacteria and Eubacteria as another. Lake[112] built upon this later using small-subunit rRNA[112] with a new method dubbed 'evolutionary parsimony', developed prior[113]. Evolutionary parsimony considered only consistent substitutions rather than all of them, in an effort to overcome the effect of long-branch attraction, similar to the LogDet method developed several years later[114]. This alternative to regular maximum-parsimony[115,116] or distance-based trees[37] favoured a tree which split Archaea into three groups: methanogens, Halobacteria (which are both Euryarchaeota) and eocytes (now known as Crenarchaeota)[112]. In the tree of Lake[112], methanogens and Halobacteria are contained within the 'Parkaryotes' group, and eocytes share a closer common ancestor with eukaryotes, known as the 'Karyotes'. The inferred rooting suggested that Archaea were a polyphyletic group, and that Eukaryotes and Eubacteria were both independently monophyletic (Figure 2). Incidentally, Lake's[111] hypothesis was consistent with Woese's[37] insomuch as the divide between prokaryotes and eukaryotes did not represent the deepest split in the tree of life. However, Lake's eocyte tree differed from Woese's by removing the hypothetical last archaeal common ancestor (LACA), as it was synonymous with all extant life. Lake's parsimonious estimation was that LUCA was a thermophilic, sulfur reducing prokaryote[112].

After these ideas had been published, other avenues of molecular phylogenetic evidence followed. Gouy and Li[117] used neighbour-joining and maximum-parsimony methods on rRNA sequences from both subunits of the ribosome. In both cases (and additionally in the case of using Lake's[112] evolutionary parsimony method on the large subunit rRNA) a tree favouring the archaebacterial tree of Woese was inferred[117]. Iwabe et al.[118] rooted and inferred a neighbour-joining archaebacterial tree using two pairs of paralogues, elongation factors G & Tu as well as the subunits of ATPase: α and β. For both sets of paralogues, thought to have duplicated in or before LUCA, the inferred trees placed Archaea and eukaryotes as sister-groups, With the root between Bacteria (with sequences from chloroplasts and mitochondria) and the Archaea +

Eukaryota group. Iwabe et al.[118] noted that that the sequence similarity of the ribosomal subunits did not reflect the tree inferred from the paralogous proteins, which could be indicative of a faster rate of substitutions between eukaryotes and their shared common ancestor with Archaea than between LUCA & LECA and LUCA & LBCA.

Further evidence consistent with this work came from Pühler et al.[119] who used RNA polymerase sequences (using both a distance matrix, and maximum-parsimony) and inferred a tree also dividing life into three, with Archaea sharing a closer common ancestor than with Bacteria. Gogarten et al.[120] examined the paralogous ATPases and the inferred tree topology placed Archaea (represented only by *Sulfolobus acidocaldarius*) with various eukaryote sequences. Interestingly, as *S. acidocaldarius* was a crenarchaeote (eocyte), it did not refute the eocyte tree of Lake et al.[111].

Woese et al.[43] formalised the argument with the introduction of a new taxonomic rank: 'Domain'. As well as naming *Archaea*, *Bacteria* and *Eukarya* as the primary domains of life, they provided the formal names for the archaeal kingdoms: *Crenarchaeota* and *Euryarchaeota*. The tree was rooted according to the work of Iwabe et al.[118] and Gogarten et al.[120], with branch lengths and topology inferred under a distance-matrix from Woese[121]. The tree presented[43] became the working hypothesis for the tree of life for the next couple of decades (Figure 3).

Forterre et al.[122] cast doubt on the rooting of the tree of life based on paralogous genes, and proposed that LUCA was not a progenote as suggested by Woese and Fox[109] but a cell containing multiple DNA polymerases. They suggested that the ATPase genes do not sample Bacteria, and that the rooted trees[120] are the result of grouping paralogous bacterial F-type ATPase and orthologous eukaryotic and archaeal V-type ATPase genes together. The bacterial V-types (hypothesised to be discovered later), would agree with the tripartite tree, rather than the eukaryote and archaeal sister-grouping[122]. In contrast to this, Hilario et al.[96] showed how the ATPase conflict is a result of problematic HGT events, and that the grouping of thermophilic bacteria with Archaea was due to this rather than hidden paralogy being the problem, they suggested that using a single-gene duplication to root the tree of life was not enough. Doolittle and Brown[123] perhaps best summarised the

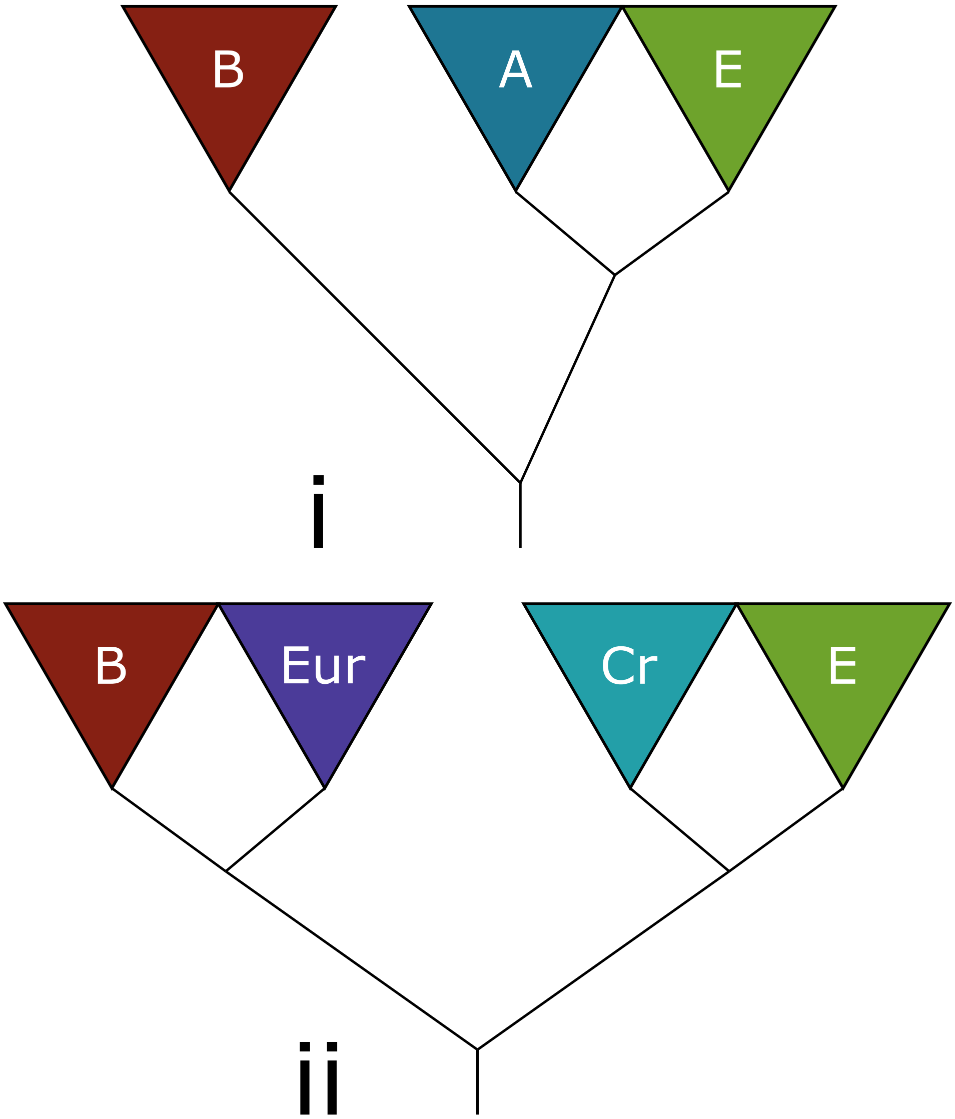

Figure 3 – A cartoon figure comparing the Woesean[43] tripartite tree of life (i), to the Eocyte tree of Lake et al.[111] (ii). In both cases these trees are rooted according to how they were presented in the original publication. A = Archaea, B = Bacteria, E = Eukaryota, Eur = Euryarchaoeta, Cr = Crenarchaeota (described as Eocytes by Lake et al.[111]

work above on the tree of life up until this point.

In an attempt to argue for the taxonomic classification of prokaryotes and eukaryotes, instead of Archaea, Bacteria and Eukaryota, Mayr[124] wrote a perspective summing up some of the philosophical differences between the approaches of differentiating using shared common ancestry versus the traditional classification based on 'grades' of evolution in regards to the tree of life. Mayr[124] suggested that the two 'Empires' (prokaryotes and eukaryotes) are a more consistent way of describing the tree of life. The obvious morphological diversity apparent in the phenotypes of 'higher' eukaryotes compared to the shared simplicity of Archaea and Bacteria justified this system[124]. In terms of practicality, the argument of Mayr[124] had merit. Sometimes it is more appropriate to use the classification-based approach, for example in medicine: Bacteria are classified based on their shape, ecology or their pathogenicity. There is certainly utility in being able to refer to prokaryotes as a group as they do differ spectacularly from eukaryotes in many important ways. However, in evolutionary biology, organisms are placed into groups based on their evolutionary history and shared ancestry. Prokaryotes are not a monophyletic clade, and there is a wealth of molecular evidence contradicting Mayr[124] showing how disparate these domains of life truly are, as summarised concisely by Pace[125].

Gupta[126] provided an alternative view of the tree of life reinforcing the idea of a prokaryote-eukaryote divide. However, the main focus of the paper was a re-imagining of prokaryote taxonomy based on novel molecular evidence[126]. They suggested that the earliest split in the tree of life was between monoderm (gram-positive) prokaryotes and diderm (gram-negative) bacteria. This division is evidenced by the presence of a singular lipoprotein membrane and the shared absence of a large insertion within a heat shock protein (Hsp 70). The proposed groups were *Monodermata*, consisting of Archaea (divided into Euryarchaeota and Crenarchaeota) and Gram-positive bacteria (further divided based on GC content) and *Didermata* (gram-negative bacteria). In retrospect, neither of these arguments are without problems, for a start there are only five sampled archaeal Hsp70 proteins in the work[126]. One insertion in the diderm sequences of one protein does not seem enough to overturn the phylogenetic evidence from other proteins. Sequence similarity in and of itself is not enough to suggest closeness of evolutionary relationships[127,128,129]. The monoderm-diderm suggestion on the other hand is interesting, and the question of when diderm cells evolved is a current matter of debate[24], however the differences between the archaeal cell wall and bacterial cell wall (peptidoglyclan vs pseudopeptidoglycan, ether- vs ester-based lipids) as well as fundamental differences in their respective biochemistry (known at the time)[130] suggested that the grouping of Archaea with monoderm bacteria was not justifiable.

The additional focus of the paper was on the divide between prokarytoes and eukaryotes[126]. Seemingly echoing the previous argument[124], and again based on the philosophical difference between 'cladists' and 'Darwinian taxonomists', i.e. that the phenotypic differences between eukaryotes and prokaryotes are so great that this should be the very reason for the domain divide, and not based on shared ancestry. This seemed to be at odds with the cladistics based approaches used for the prokaryote groups presented within the same paper[126], but could be appropriate considering the unique nature of the event leading the formation of the eukaryotic cell from Archaea and Bacteria.

The opinion of Cavalier-Smith[131,132] changed during this period: from a classical taxonomy based view favouring the prokaryote-eukaryote divide, to a re-envisaged tree of life split between Bacteria and Archaea + eukaryotes, dubbed *Neomura*[133] (although still based on taxonomic, rather than phylogenetic evidence). In each case[131,132,133], the root lay within a paraphyletic Bacteria, and LUCA and LBCA were one and the same: a diderm (*Negibacteria*). Monoderms (*Posibacteria*) contained a monophyletic Archaea[132]. The later revision[133] placed Archaea as a monophyletic sister-group to eukaryotes. These classifications were based on shared-traits (apomorphies) and novel-traits (autoapomoprhies) rather than evidence from inferred phylogenies. Nonetheless, the approach of Cavalier-Smith[133] tied together multiple independent forms of evidence. The proposed root within Bacteria[133] agreed with the conclusion of Gupta[126] that Archaea evolved from a monoderm bacterial ancestor. One challenge in comparing and distinguishing the hypotheses in some of these key papers[131,132,133] is that a largely original nomenclatural scheme is used. Phylogenetic bracketing would make the LUCA/LBCA of Cavalier-Smith[133] a mesophilic, cellular organism with fully evolved informational processing machinery. This was in stark contrast to the progenote of Woese and Fox[109].

An analysis by Ribeiro and Golding[134] on a newly sequenced *Methanococcus jannaschii* genome showed how individual gene phylogenies supported different tree of life topologies. In that analysis, the majority of genes either favoured the Archaea and Eukaryota sister-group, or eukaryotes grouping with gram-negative bacteria. Ribeiro and Golding[134] performed ML analysis on every protein-coding gene from the *M. jannaschii* genome, using BLAST and a Dayhoff et al.[69] substitution model in PROTML[135]. The results showed that the genes favouring the Woesean Archaeea + eukaryotes monophyly were involved in translation, e.g. ribosomal proteins. The authors came to the conclusion that Bacteria + Eukaryota grouping was probably the result of HGT between both groups early on in eukaryote evolution.

A definitive rebuttal of the tree of Woese et al.[43] was shown through phylogenomic analyses through a collection of papers[136,137,138] (see the *Universal Marker Genes* section below). These works suggested that the tree of life is more similar to that of Lake[111], however the two domains in this 'Two-domain tree' are Bacteria and Archaea.

1.6.2 The ring of life

Doolittle[139] philosophised on previous work into the effect of HGT on the tree of life. They suggested that the mounting evidence of HGT across all domains of life required attention, as it had the potential to undermine the phylogenetic classification itself. How can there be a tree of life if different genes give contradictory evolutionary histories, and by extension — could a single LUCA possibly exist? Although rRNA and other core markers involved in translation are unlikely to have undergone HGT, Doolittle[139] suggested that the existence of conflicting gene histories undermined the phylogenetic classification as envisaged by Darwin[7], Zuckerkandl and Pauling[63] or Woese and Fox[37]. The striking figure of a tree with a complicated web of transfer events[139] foreshadowed the idea of a network- rather than a tree of life[140].

Woese[141] reiterated earlier points about the progenote, although under the guise of a 'Darwinian threshold'. This is the idea that LUCA was not a single 'species', but a collection of different organisms undergoing rampant HGT. It was not until the last common ancestors of Archaea, Bacteria and Eukaryota when the central dogma of DNA to RNA to protein synthesis finished evolving[141]. The evidence for this theory came from the differences between the translational and transcription machinery, tRNA and RNA polymerase. If this was the case, then this would mean that the common ancestor of Archaea and Eukaryota evolved into a cell later, which begs the question, why did HGT from Bacteria not resolve the genotype/phenotype issue in LACA and LECA? It would also seem somewhat contradictory to have a tree with a single root if the belief was that there was so many HGTs occurring, a single LUCA would not have existed. Although the evolution of the cell may have been the result of HGT, this could also have occurred before the root of the tree of life. The similarities we see across each domain suggests that the *central dogma* of the cell was present in LUCA. However, Woese[141] was correct in so much as the differences between cell membranes both within and between domains were questions that needed to be addressed.

In a bid to try and understand some of the biases and factors behind HGT, Jain et al.[142] used an early method of gene-tree species-tree reconciliation[143] to measure the number of 'transfer steps' a gene tree is away from the species tree. Using transfers taken between an organism preferring one particular environmental condition to another, a score was given using regression between the transfer score and the maximum-parsimony score. From their analysis[142] they found that both oxygen-tolerance and temperature were significantly associated with HGT (i.e. a gene in thermophilic environment to thermophilic environment happened more frequently than a gene found in a mesophilic environment to thermophilic environment), non-proximity based adaptations were also clear indicators of likely HGT. These were features such as genome size, GC content and carbon utilization (which can all vary widely in the same proximal environment). Other proximal features like salinity, pressure or pH were only shown to have weak effects on the rate of HGT[142]. However, the authors concluded that some environmental constraints apply more pressure on HGT than others[142]. This provided some backing for the idea of using a smaller set of universal marker genes rather than all genes to infer species trees as they are less likely to have been transferred[144,145], however there was evidence to suggest that ribosomal proteins (a traditional choice for universal markers) are not immune to the effects of HGT[146].

Rivera and Lake[147] suggested that because using whole genome data leads to contradictory trees, the tree of life was not a simple bifurcating tree at all, and is more ‘ring’-like culminating in the eukaryote fusion event between Archaea and Bacteria[147]. This was another premonition of modern network-based approaches. The argument is because eukaryotic genomes are composed of genes acquired from both Archaea and Bacteria, when they are represented in a species tree, these multiple origin points for their genomes should be displayed.

One issue with this was the implication that, in the ring-tree, Crenarchaeota are equidistant to Proteobacteria than Euryarchaeota which contradicted the actual phylogenetic analyses at the time[37,111,120,118]. The analysis of Rivera and Lake[147] reaffirmed the idea that whole-genome data are not ideal for inferring the species tree. Although eukaryogenesis itself appears to be a unique event, conflicting gene histories are present in more than just this instance. Following the logic, the true species-tree would be akin to the mess of HGTs drawn by Doolittle[139] (Figure 4). The argument against this would be that there is a vertical signal which represents Darwinian-vertical descent, but it does raise a good question — what genes are appropriate for inferring the species tree of life?

It would seem that Rivera and Lake[147] would agree that Elongation factor TU and ATPase are ‘good’ genes as they based their rooting of the ‘ring’ on these genes, but a ring of life would not be inferred by using these markers. Choosing to present the tree of life as a ring is an aesthetic choice, and one could simply choose to present multiple trees. The view from Rivera and Lake[147] was reinforced by Dagan and Martin[148] who concluded that trees are no longer fit-for-purpose in displaying the ever-growing mass of conflicting data. They saw a dichotomy: either evolution is not treelike and therefore a network is the most appropriate way of displaying evolutionary relationships, or one takes the ‘positivist’ approach and uses a knowingly incorrect species tree. Ultimately, this matter comes to a preference for aesthetics, as there was no argument over the presence of HGT, but simply if such transfers should be included in the species tree or not.

There is evidence for certain groups of genes being less likely to be transferred than others[149], where a combination of laboratory methods on *Escherichia coli* combined with simulations showed that proteins involved in translation machinery were potentially toxic when transferred. This suggested that there are marker genes which would follow the species tree in most cases, casting some doubt on Dagan and

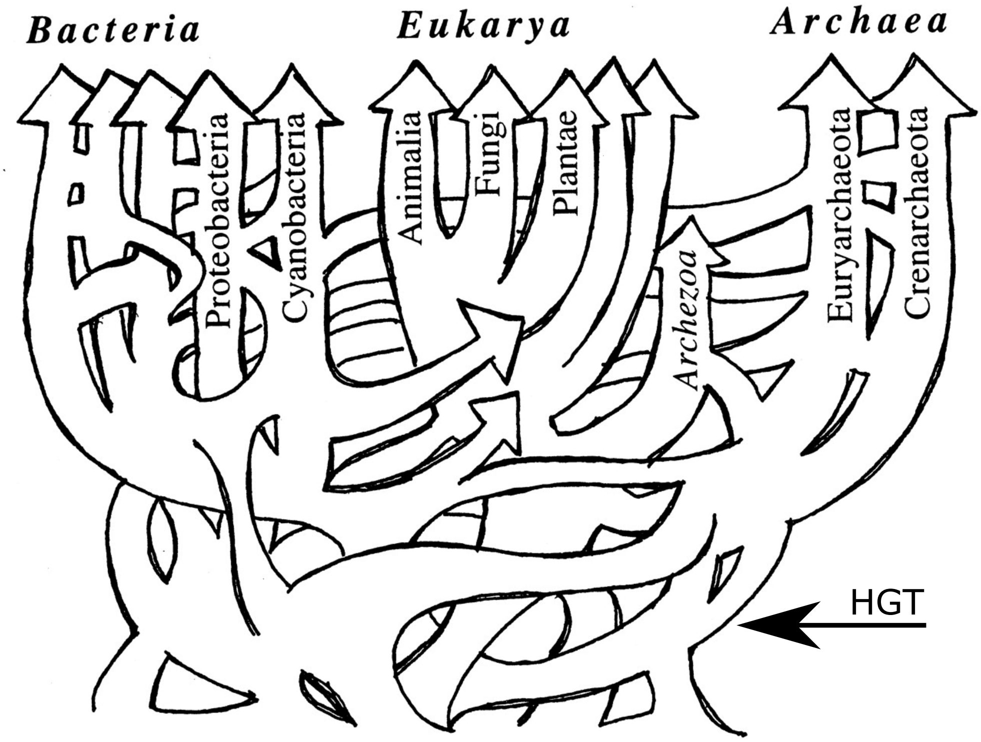

Figure 4 – The reticulated tree from Doolittle[139], showing the potential reality of species evolution in the light of HGT (Horizontal gene transfer), undermining the vertical line of descent described by Darwin[7]. Adapted from Doolittle[139].

Martin's argument[148].

Cox et al.[136] took a novel phylogenomic approach concatenating 45 proteins into a supermatrix. They also tested an expanded dataset of ribosomal large- and small-subunit RNA, and applied more recent ML and Bayesian inference models. They found that for both datasets (rRNA and protein), a tree favouring a topology similar to that of Lake et al.[111] was inferred (assuming a root on Bacteria). Cox et al.[136] suggested that by accounting for compositional heterogeneity with the CAT model[77,150] the result was less likely to be affected by compositional biases. They also employed Dayhoff et al.[69] recoding to alleviate substitutional saturation, which may have affected earlier phylogenetic analyses. Through this work, they showed that when using methods which do not account for compositional heterogeneity, the three-domain tree was still inferred for the rRNA trees, and the results from individual gene trees also became obfuscated.

To address the issue of species-tree rooting, Lake et al.[32] proposed 'indel rooting', a method by which shared indel (insertion or deletion) events are used as evidence for the location of the root. Specifically: shared indels across three paralogous genes PyrD, HisA, and HisF. These genes have shared indels between Archaea and Firmicutes in HisA, and between Actinobacteria and diderm bacteria within PyrD[32]. The conclusion of Lake et al.[32] was that a network is a better representation of the tree of life, as it allows for the multiple ways in which trees can be rooted[151].

Another alternative to tree of life inference is through the use of supertrees. Pisani et al.[152] took this approach, where individual gene trees are computed, in this case using maximum-parsimony and NJ based methods. The topological information from these trees are then combined to make a 'supertree'. For these analyses the authors used whole genome data to infer hundreds of individual gene trees[152]. They found that the majority (83%) of (gene trees from) eukaryote genomes supported a placement of eukaryotes to be within a bacterial clade, either from within Cyanobacteria or Alphaproteobacteria, with the strongest support suggesting a cyanobacterial origin of these genes (exactly where and when in the cyanobacterial tree the plastid was derived from would be discovered later[153]). Pisani et al.[152] applied a 'phylogenetic signal stripping' approach in order to distinguish between the conflicting signals in the data. Firstly, by removing all the gene-trees which supported a cyanobacterial origin to infer a tree with the next strongest signal (an alphaproteobacterial origin of eukrayote genes). These gene-trees were then removed and surprisingly a eukaryote-thermoplasmatales link was found, which was in contrast to the usual placement of eukaryotes as sister-group to Crenarchaeota. Thermoplasmatales is a euryarchaeote, so this was different to both the Woese and Fox[37] and Lake et al.[111] hypothesis. The conclusion the authors made seems sensible, i.e. there are two primary divisions of life, Archaea and Bacteria, and eukaryotes arose from a symbiosis-event between the two primary domains. They conclude a new paradigm is needed to address the fusion event, tentatively agreeing that the 'ring of life' model from Rivera and Lake[147] maybe the best way in which this is achieved. The thermoplasmatales-eukaryote relationship was later investigated by Williams et al.[154] with better fitting models using both single-gene trees and a supermatrix approach, the relationship was found to be poorly (21% bootstrap) supported in the few cases where it was recovered.

A different approach at understanding prokaryote evolutionary history[155] used amino acid identity in lieu of traditional marker genes. Although amino acid similarity is an important factor, it has been shown that not only is it not necessarily an indicator of evolutionary relatedness[127,128,129] but is also susceptible to convergence adaptation[156] or even just from HGT[142]. On a whole-genome scale these factors are likely to be compounded and as such, caution must be advised when using genome similarity as a basis for taxonomy (if we agree with the premise that taxonomy should be based on evolutionary relationships). Curiously, Konstantinidis and Tiedje[155] also showed that the archaeal-bacterial branch length was liable to shortening when using different markers to the traditional 16s rRNA. Their analysis also showed that the archaeal-bacterial branch length was consistently longer when using 16s marker genes compared to using amino-acid identity or other gene sets[155]. These results highlight some of the potential pitfalls when using only a single gene dataset.

Dagan et al.[157] utilised a novel method of rooting the tree. Similar to Rivera and Lake[147], they posited that using the classic bifurcating dendrogram was an inferior way of displaying the evolutionary relationships at the deepest part of the tree. They suggested that using a network of 'splits' better visually represented the data. Ultimately a splits network looks similar to a tree, but instead of a bifurcating tree, there are multiple interconnected branches which can be connected to by other multiple branches. In that paper[157], they used sequence-similarity as the metric by which to create the splits, i.e. certain ribosomal proteins over a specific sequence-similarity threshold create a binary split between (generally speaking) Archaea and Bacteria. This was done for a range of proteins and then used to produce a network of relationships. One of the issues with this was that similarity is not indicative of molecular evolution[128,129] and does not take into account convergence. Another issue was that using this similarity-based approach did not take into account rate of evolution and assumed a homogeneous rate of evolution all over the tree.

Theobald[129] provided a formal test of LUCA, using maximum likelihood and Bayesian inference with model selecting using the AIC, (but with a penalisation for increased model complexity), log likelihood ratio tests and log Bayes factors[158,159]. Their results showed that when performing model selection tests on a range of hypothesis of separated trees, i.e. Bacteria as a separate tree to an archaeal and eukaryotic tree compared to a tree in which all three domains are present. An unrooted tree similar to that of Woese et al.[43] was strongly preferred[129]. They also tested additional models allowing for a network rather than a bifurfacting tree, so allowing free HGT, and with these tests the three-domain was still strongly selected for; Theobald[129] used this as evidence that LUCA is more likely than multiple independent ancestors of domains, which could be argued also rejects the 'progenote' hypothesis of Woese[141].

1.6.3 Universal marker genes

Harris et al.[160] used the clusters of orthologous genes (COG) database and found 80 genes present across all domains of life (a selection of 34 bacterial, archaeal and eukaryotic genomes). They used NJ and maximum-parsimony to infer individual gene trees, the topology of these trees and their function was then used to classify the different genes into groups[160]. The majority of the genes (50/80) preserved the three domain topology outlined by Woese and Fox[37], and none of the gene families shown supported the eocyte topology[111]. This paper is important because it highlights the synonymy of ribosomal proteins and other protein-coding genes which track the vertical signal of evolution[7], as 37 of the 50 three domain trees were from genes coding for proteins physically associated with the ribosome[160]. Unfortunately the trees which do not reflect the three-domain topology were not included in the publication.

Another phylogenomic analysis found 31 orthologues present across 191 species representing Archaea, Bacteria and Eukaryota[161] after the removal of five genes contaminated by HGT. All 31 were involved in translational processes. The inferred tree is a classical three-domains tree. Within the bacterial domain, Acidobacteria are monophyletic with Proteobacteria, the terrabacterial clade is non-existent, with the deepest split in the bacterial tree being between Firmicutes and all other sampled bacteria (albeit with admittedly poor bootstrap support). Within the archaeal domain, Nanoarchaea is placed as sister to Crenarchaeota, which in turn are sister to a monophyletic Euryarchaeota. As an aside, the ecdysozoan monophyly is also broken up in this tree, with arthropods being placed as sister to chordates, and nematodes being placed outside this clade. This is acknowledged in the paper as potentially being a result of accelerated sequence evolution in arthropods and nematodes potentially affecting the results. One potential issue with the supermatrix approach taken in this analysis is the use of separate multiple sequence alignments for each COG, for each domain. These were then concatenated into the overall supermatrix from which the maximum likelihood tree was inferred. An argument could be made that having three independent alignments for each gene would predispose this tree towards a three-domain tree. Is it possible that an eocyte-like tree would be more likely if Archaea and Eukaryota had been aligned under the same model?

A study from Yutin et al.[162] recovered multiple origins for the majority of a selection of eukaryotic genes, either archaeal, bacterial or unresolved. 136 gene trees display the monophyly of Bacteria, Crenarchaeota, Euryarchaeota and Eukaryota. Surprisingly though, while the origin for the majority were apparently from a 'deep', as of yet, undiscovered and possibly extinct clade of Archaea, the other markers of archaeal origin were likely to have come from Crenarcheota and there was minimal support for the small number of markers from a euryarchaeal origin. The results of that study reinforced the idea of a close relationship between some Archaea and Eukaryotes in contrast to the classical three-domains tree[43], and indirectly supported the ideas of eocyte-like tree[111].

Foster et al.[137] used an an updated set of 41 protein-coding genes (from Cox et al.[136] and ribosomal RNA genes to test the affect on resultant tree topology, they also introduced a new model node-discrete rate matrix heterogeneity (NDRH)[163] which allows for heterogenous substitution rates across the tree. They showed that using ribosomal RNA genes with less better-fitting models supports a Woesean tree, whereas using the universal marker genes with any probabilistic method (i.e. excluding maximum parsimony), or using better fitting models (such as the CAT model) on the ribosomal RNA genes, an eocyte-like[111] tree is inferred; albeit with the root on the branch leading to the bacteria, rather than Bacteria and Euryarchaeota as first suggested[111].

The archaeal domain underwent a substantial expansion[154], with the dawn of a a new super-phylum named TACK[164] which includes Crenarchaeota, Thaumarchaeota, Aigararchaeota and Korarchaota. Due to an increased genomic sampling, these more recently discovered groups were shown to be monophyletic through a range of different phylogenetic models on protein concatenations (including maximum likelihood[165] and Bayesian inference accounting for compositional heterogeneity through the use of the CAT model[77,150]). According to this analysis Archaea were not monophyletic, in contrast to the tree of Woese et al.[43], and the resulting tree inferred from the concatenation is similar to the tree of Lake et al.[111]. According to that analysis, Eukaryota and TACK are sister-groups, and regardless of rooting position the three-domains tree is not recovered[154]. The sister relationship of TACK rather than a clade within Euryarchaeota is in conflict with the results of Pisani et al.[152].

Beiko et al.[166] used simulations on 1000 genomes to examine the effect of gene duplication, gene transfer, and loss on the inferred phylogenies. They found that deeper relationships in particular can have misleadingly high support as result of compositional bias, even when we know *a priori* that those branches are incorrect[166]. They also found that HGT across closely related genomes effectively decreases inferred branch length as habitat based transfer will infer trees supporting these horizontal relationships. When this data is combined with vertical signal the resulting tree is a compromise: neither a reflection of horizontal nor vertical relationships[166]. Beiko et al.[166] noted that supertree methods provide an easier solution to removing poorly supported trees (as one can remove individual offending trees). However, it should be noted that based on their simulation results[166] this still requires manual inspection of the component gene trees. One of the conclusions from that paper was that using whole-genome information for concatenation based approaches are likely to be strongly affected by HGT[166].

Williams et al.[138] re-affirmed that there are only two primary domains of life: Bacteria and Archaea, and that in terms of nuclear DNA, eukaryotes share a common ancestor closer to TACK than to Euryarchaeota. Although in principle this idea is similar to that of Lake et al.[111], the neomuran clade put forward by Cavalier-Smith[133] was also similar in that Archaea + eukaryotes is a monophyletic clade — so perhaps the term Neomura should be used today? Although this is a review article, it essentially echoes the results of Williams et al.[154] amongst an increasing wealth of other molecular evidence[164,167].

Petitjean et al.[168] used three supermatrices, one from a concatenation of 32 ribosomal proteins, one using a concatenation of 38 non-ribosomal proteins and the other composed of both aforementioned datasets. In each case, orthologous proteins from Bacteria are used as an outgroup. Not only did their ML analyses reinforce the idea of a monophyletic TACK (formally dubbed Proteoarchaeota in this analysis, a parallel to the bacterial proteobacterial lineage), they found the paraphyly of Euryarchaeota when using solely ribosomal proteins as markers[168]. They[168] suggested that this was a reason to be sceptical of analysis such as Cox et al.[136] where results are dependent upon supermatrices composed of only ribosomal proteins. However, it should be noted that the supermatrices of Cox et al.[136] were not composed solely of ribosomal proteins, and actually only 40% of the markers used were ribosomal. Petitjean et al.[168] presented the set of 38 non-ribosomal markers, and when using both this and a combined set of 70 proteins, monophyletic Euryarchaeota was inferred. Another interesting result from that paper concerns the archaeal-bacterial branch length. They found that the branch length for the tree inferred from the ribosomal markers was three times longer than the non-ribosomal marker tree. In the combined supermatrix analysis, the inferred archaeal-bacterial branch length was much closer to that of the non-ribosomal markers. This suggested that the non-ribosomal markers are much slower at evolving than the ribosomal ones. Could this suggest an accelerated rate of evolution in ribosomal markers or perhaps a much slower rate of evolution in non-ribosomal markers? Other potential explanations could be the effect of substitutional saturation not captured by the model, leading to long branch attraction, or compositional biases in the non-ribosomal set artificially reducing the archaeal-bacterial branch length. This analysis challenged the traditional use of ribosomal markers in supermatrix analyses.

The discovery of Lokiarchaeota assembled from metagenomes found in the Arctic Mid-Ocean Ridge (near Loki's Castle)[169], changes the tree of life significantly. A ML and Bayesian phylogeny of eukaryotes and Archaea show Lokiarchaeota to be sister to Eukaryota (albeit with only 80% bootstrap support but maximum posterior probability with CAT-GTR[77,150]). These were the first Asgard to be discovered, which would later be expanded by further analysis[39], showing that other Asgard groups such as the Heimdallarchaeota could be even more closely related to Eukaryota. These results suggest that the two-domain tree is correct, but specifically the TACK sister-group hypothesis is the result of under-sampling archaeal diversity. The 'deep' archaeal origin[162] of many of eukaryote genes makes sense in light of this new group of archaea[169] and may explain the difficulties in previously inferred topologies.

1.6.4 Fine details

Perhaps one of the most up-to-date and holistic approaches to the entire tree of life (both primary domains and eukaryotes) came from Hug et al.[27]. This analysis used a concatenation of 16 ribosomal marker genes for over 3000 taxa. All taxa involved had to pass a basic test for genome quality: a threshold for completeness and the presence of domain-specific orthologues. After trimming, the number of sites in the supermatrix was only 2596. Unlike when using maximum-parsimony where using fewer sites than taxa is intrinsically problematic, sacrifices are being made in terms of model-choice and support methods. Hug et al.[27] used the simple LG+G model in a ML framework, as well as only 156 rapid-bootstrap replicates (with 100/156 sampled for support values). It could be suggested that using longer sequences and reducing the number of taxa[170] or recoding[81] may have been worthwhile alternatives or additions, both for a more accurate, better supported and overall more efficient analysis as the tree inference of the analysis took 3840 computational hours.

Nevertheless, the ambitious work of Hug et al.[27] reaffirmed the two-domains tree, with Eukaryota sister to the Asgard archaea. However, the tree failed to realise a monophyletic DPANN, but they are generally placed closer to the archaeal-bacterial branch. Both TACK and Euryarchaeota were monophyletic, however bootstrap support is relatively low all over the tree. The earliest split in the bacterial domain is between the newly discovered CPR[25] and all other bacteria, neither Gracilicutes nor Terrabacteria were monophyletic in the bacterial portion of the tree. The authors also included a ribosomal small-subunit analysis, which still recovered the three-domain tree under a ML implementation of the GTR model accounting for across-site rate heterogeneity (incidentally this is called GTRCAT[171], but it is a different model to CAT-GTR[79], which accounts for across-site compositional heterogeneity). Ultimately this tree served as a good checkpoint in tree of life analysis up to this point, as it incorporated all domains it allows us to make quick comparisons for where the major groups are placed, but was by no means a complete or final tree of life (see Figure 5 for a simplistic tree based incorporating the results of Hug et al. [27]).

The expansion of the archaeal domain with the discovery of the DPANN clade of Archaea[47,48] and the even more recently discovered Asgard archaea[169,39] allowed Williams et al. [52] to infer a clearer picture of the evolutionary history of Archaea and go further in using that information to understand how LACA may have lived. That analysis[52] combined a supertree approach from 3242 single-copy gene families, a supermatrix approach (a concatenation of 45 mainly ribosomal proteins) and a new method of incorporating evidence from homologous genes using gene-tree species-tree reconciliation[172], allowing gene duplications and HGT to inform rather than obscure. From their analyses (using Bayesian inference, with variations of the CAT model[77] and Dayhoff[69] recoding on the supermatrix) they found that the root of Archaea was between TACK and Euryarchaeota, but when including DPANN, the root actually was between DPANN and other archaea[52].

Recently, using a massive dataset of over 10,500 genomes of Bacteria and Archaea Zhu et al. [30] found a much bigger list of universal marker genes than others up until this point. Using an automated pipeline[30] to find 381 markers (derived from a larger set[173]) which were present in the majority of the genomes they tested (and all genomes used had at least 100 of these markers present). They used both a supertree and a supermatrix approach to infer a tree, but the main focus of this paper was the supertree, with branch lengths estimated using a maximum-likelihood estimation on taxa form subsampled supermatrix. Their results agreed with those of Petitjean

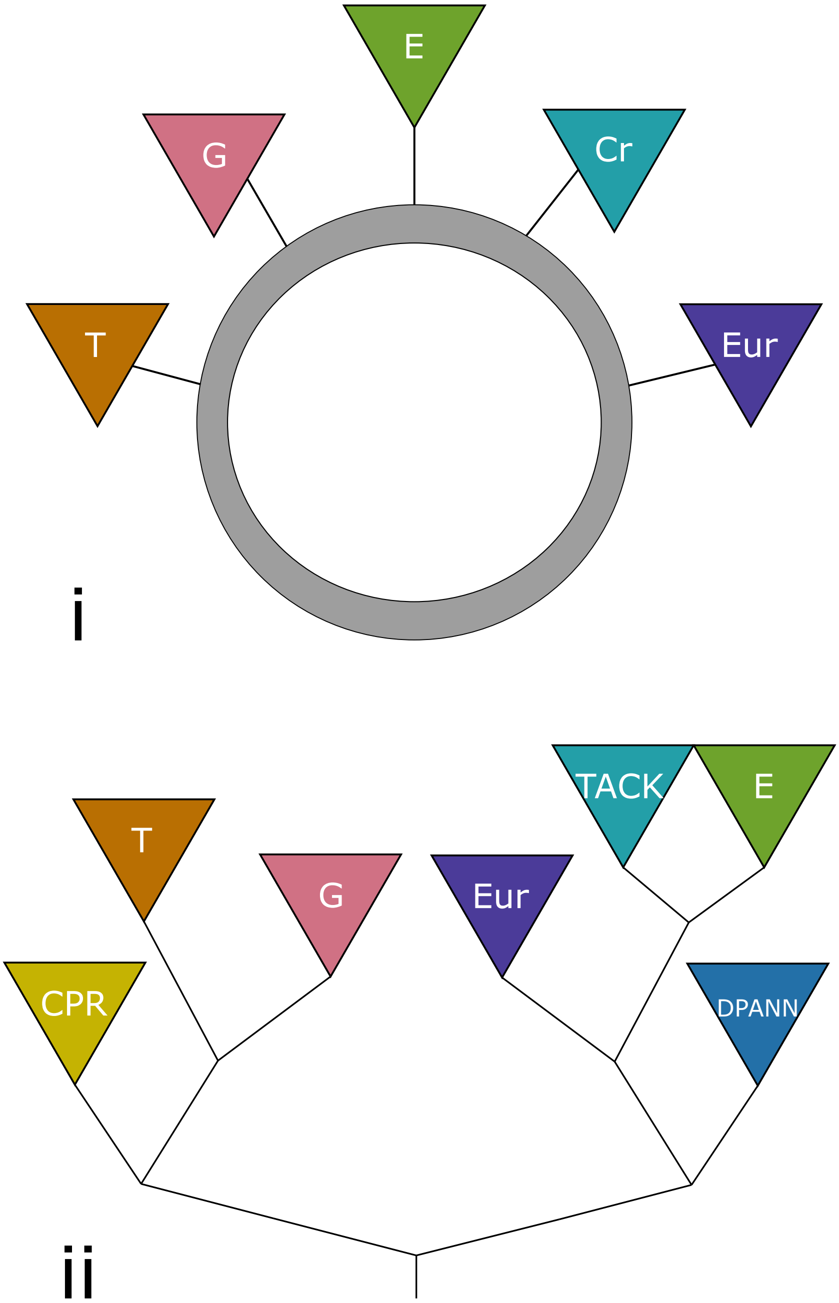

Figure 5 – A cartoon figure showing the ring of life[147] (i), and a two-domains tree (ii). The two domains-tree is loosely based on the one inferred by Hug et al.[27] but incorporating the split between Gracilicutes and Terrabacteria as defined by Coleman et al.[24], with the exception of CPR which both Zhu et al.[30] and Hug et al.[27] find to be sister to other bacteria. A = Archaea, E = Eukaryota, Eur = Euryarchaoeta, Cr = Crenarchaeota (later expanded to TACK), T = Terrabacteria, G = Gracilicutes.

et al.[168] in that there was a stark difference between a tree inferred from ribosomal markers in comparison to one derived from additional non-ribosomal proteins.

The resulting topology of the supertree is somewhat similar to the Hug et al.[27] tree but differed in several important ways. The CPR bacterial group are still found to be monophyletic and sister to the rest of bacteria consistent with the analysis of Hug et al.[27], but Terrabacteria were found to be monophyletic in that tree (unlike Gracilicutes). The archaeal domain had monophyletic TACK, but found a paraphyletic euryarchaeota (with Thermococci grouping closer to TACK than Thermoplasmata, Methobacteria, Halobacteria and Archaeoglobus). Curiously, the archaeal-bacterial branch length inferred from the 381 marker set was smaller than multiple intra-bacterial branches. This has implications for the root of the tree of life, as a longer branch length could be argued as a more likely location for the root placement. Through additional analyses, Zhu et al.[30] showed ribosomal genes appear to be evolving much more quickly than the majority of other genes (except for a small group of outliers) and suggested the traditional long branch between Archaea and Bacteria is an artefact of this ribosomal bias. They also performed a divergence time analysis, where a significantly younger age for LUCA was inferred (4.5 Ga), from the 381 marker set, as opposed to the ribosomal set from which a divergence time of (7+ Ga) is derived. However, it should be noted that these divergence times were estimated using a strict clock, which assumes the same substitution rate across the entire tree.

Williams et al.[54] used a modest taxon selection (92 genomes), with a gene selection consisting of 35 (mainly ribosomal) proteins[27,30]. The limited taxa selection here allowed the use of more complex models than previous analyses, which account for across-site compositional heterogeneity (using both ML and Bayesian inference). The results of that analysis suggested that the three-domains tree is a result of long-branch attraction, and that when re-analysing the marker sets previously used to infer such a tree, with these better fitting models, a two-domains tree was inferred. In addition to the supermatrix-based approach, they also used a supertree method which corroborates these results. The tree of life inferred from these analyses showed the paraphyly of Archaea: eukaryotes were found nested within the Asgard archaea and the deepest split within the archaeal domain was between Euryarchaeota and

Proteoarchaeota. Interestingly, the inclusion of Bacteria deconstructed the otherwise monophyletic Euryarchaeota. The resolution of the bacterial domain itself was also less clear, however with such a limited sampling of bacteria this was not a surprise.

A much more thorough investigation of the bacterial domain was reported in Coleman et al.[24], using a range of probabilistic methods (maximum likelihood[80] implementation of the CAT model, as well as the CAT-GTR Bayesian implementation[77,150] on a recoded dataset to account for compositional bias and limit computational time) on a set of 62 orthologous proteins to infer a species-tree. They also performed gene-tree species-tree reconciliation through amalgamated likelihood estimation[172] for inferences about LBCA, and to establish where the root of the bacterial tree is (this is the main focus of the paper). They also used outgroup rooting using a smaller selection of proteins found in both Archaea and Bacteria[24]. At the time of writing this is probably the most thorough and up-to-date hypotheses for the tree of Bacteria. They found monophyletic Terrabacteria and Gracilicutes, with the root somewhere between these two clades with Fusobacteria falling on either side of this divide (Terrabacteria consistently had a much smaller clade comprised of Deinococcota, Synergistota, and Thermotogota as a sister-group). In contrast to the earlier work[27,30], the CPR were placed within the terrabacterial clade, as sister to Chloroflexi and Dormibacterota. The inferred LBCA from gene-tree species-tree reconciliation was a diderm, rod-shaped cell with flagella and a developed CRISPR-Cas system[24].

An alternative hypothesis was proposed by Xavier et al.[174] which used a larger dataset (146 proteins with a roughly even split between information-processing genes and genes involved in metabolism) but opted to use automated inspection of individual protein trees, which are then rooted using minimal ancestral deviation (MAD)[175] — an improved version of midpoint rooting which seeks to use deviations from the strict clock to root a tree, so the resulting rooted tree is one in which the minimal deviation from a strict clock is inferred. The authors suggested that the ancestral bacterium was similar to a Clostridales (a Terrabacteria) based on the short root-to-tip branch length[174].

A recent analysis from Aouad et al.[55] used a supermatrix approach on 72 protein families which did not resolve a monophyletic Euryarchaeota. They proposed that the archaeal tree was split into two major clades: *Ouranosarchaea* (containing a monophyletic Proteoarchaeota with Asgard as a sister-group, with the Methanomada, Archeronita and Stygia clades also within the new cluster) and *Gaiarchaea* (with remaining clades of Euryarchaeota). They also resolved a monophyletic DPANN (along with Altiarchaea), although they suggested this could be a result of a lack of signal or other potential biases brought on by using a bacterial outgroup. They suggested there are two alternative positions for the root, either with DPANN as sister to both Ouranosarchaea and Gaiarchaea, or between Gaiarchaea and other archaea, with DPANN as the sister to Ouranosarchaea. This result was sensitive to the methods used to infer the tree. Most recently Muñoz-Gómez et al.[34] used a model accounting for branchwise compositional heterogeneity to infer the ancestor of mitochondria was nestled comfortably within Alphaproteobacteria.

In chapter two, I will look into the larger set of markers from Zhu et al.[30], which suggested a much shorter evolutionary distance between the primary domains of life. I use the non-ribosomal markers of Petitjean et al.[168], the ribosomal markers of Williams et al.[54] and the bacteria focused markers of Coleman et al.[24] as comparisons. In chapter three, I will expand upon the archaeal and bacterial tree by inferring a novel prokaryotic tree of life from a new consensus set of vertically-evolving marker genes in addition to investigating the potential sources of bias and artefacts in this marker set. During the course of my PhD, I have also been involved in a number of other projects looking at the evolution of specific proteins and protein complexes, which are included as chapter four. Chapter five is a set of concluding remarks and future considerations.

2 A SHORT BRANCH BETWEEN THE PRIMARY DOMAINS IS AN ARTEFACT OF MARKER SELECTION AND OTHER BIASES

Author's contribution

Parts of this chapter form part of a publication[176]:

Moody, E. R. R., Mahendrarajah, T. A., Dombrowski, N., Clark, J. W., Petitjean, C., Offre, P., Szöllősi, G. J., Spang, A., and Williams, T. A. (2022). An estimate of the deepest branches of the tree of life from ancient vertically-evolving genes. *eLife*, 11:e66695

The project was conceived by Tom A. Williams, Anja Spang and Edmund R.R. Moody. Marker sets were collated by Edmund R. R. Moody. Individual gene trees, tree processing, concatenation, manual curation and individual gene tree inspection, statistical analyses, tree rooting, rate inference, archaeal-bacterial branch length estimation, tree length estimation, and relative archaeal-bacterial distance were carried out by Edmund R.R. Moody. Annotations and split scores were carried out by Tara Mahendrarajah and Nina Dombrowski on gene trees generated by Edmund R. R. Moody. Tara Mahendrarajah manually performed taxonomy counts and manually inspected the 381 individual gene trees from the expanded set before sub-sampling. Edmund R. R. Moody wrote the paper with comments and suggestions from other co-authors.

2.1 Abstract

Traditional estimates of the tree of life have divided the tree into two primary domains, Archaea and Bacteria. Previously, these domains were thought to be very distantly related. Recently, an analysis[30] has suggested a minimal branch length between these two domains and therefore a much closer proximal relationship than ever before. Here, we examine the dataset used to infer this minimal distance, amongst other previous datasets used to infer the tree of life. We find that conflicting evolutionary histories, poor model fit and substitutional saturation all serve to artificially reduce evolutionary distance between Archaea and Bacteria.

2.2 Introduction

Much remains unknown about the earliest period of cellular evolution and the deepest divergences in the tree of life. Phylogenies encompassing both Archaea and Bacteria have been inferred from a "universal core" set of 16-57 genes encoding proteins involved in translation and other aspects of the genetic information processing machinery[161,177,160,27,178,179,168,180,181,129,54]. While representing a small fraction of the total genome of any organism[148], these genes are thought to predominantly evolve vertically and are thus best-suited for reconstructing the tree of life[161,182,183,180,129]. In these analyses, the branch separating Archaea from Bacteria (hereafter, the AB branch) is often the longest internal branch in the tree[136,120,27,118,119,54].

Recently, Zhu et al.[30] inferred a phylogeny from 381 genes distributed across Archaea and Bacteria using the supertree method ASTRAL[184]. These markers increase the total number of genes compared to other universal marker sets and comprise not only proteins involved in information processing but also proteins affiliated with most other functional COG categories, including metabolic processes (Supplementary Information Table S1, see Appendix). The genetic distance (AB branch length) between the domains[30] was estimated from a concatenation of the same marker genes, resulting in a much shorter AB branch length than observed with the core universal markers[27,54]. These analyses were consistent with the hypothesis[168,30] that the apparent deep divergence of Archaea and Bacteria might be the result of an accelerated evolutionary rate of genes encoding translational and in particular ribosomal proteins along the AB branch as compared to other genes. Interestingly, the same observation was made previously using a smaller set of 38 non-ribosomal marker proteins[168], although the difference in AB branch length between ribosomal and non-ribosomal markers in that analysis was reported to be substantially lower (roughly two-fold, compared to roughly ten-fold for the 381 protein set[168,30].

Alternatively, differences in the inferred AB branch length might result from varying rates or patterns of evolution between the traditional core genes[169,54] and the expanded set[30]. Substitutional saturation (multiple substitutions at the same site[185]) and across-site compositional heterogeneity can both impact the inference of tree topologies and branch lengths[163,150,77,80,186,78]. These difficulties are particularly significant for ancient divergences[187]. Failure to model site-specific amino acid preferences has previously been shown to lead to under-estimation of the AB branch length due to a failure to detect convergent changes[188,54], although the published analysis of the 381 marker set did not find evidence of a substantial impact of these features on the tree as a whole[30]. Those analyses also identified phylogenetic incongruence among the 381 markers, but did not determine the underlying cause[30].

This recent work[30] raises two important issues regarding the inference of the universal tree: first, that estimates of the genetic distance between Archaea and Bacteria from classic "core genes" may not be representative of ancient genomes as a whole, and second, that there may be many more suitable genes to investigate early evolutionary history than generally recognized, providing an opportunity to improve the precision and accuracy of deep phylogenies. Here, we investigate these issues in order to determine how different methodologies and marker sets affect estimates of the evolutionary distance between Archaea and Bacteria. First, we examine the evolutionary history of the 381 gene marker set (hereafter, the expanded marker gene set) and identify several features of these genes, including instances of inter-domain gene transfer and mixed paralogy, that may contribute to the inference of a shorter AB branch length in concatenation analyses. Then, we reevaluate the marker gene sets used in a range of previous analyses to determine how these and other factors, including substitutional saturation and model fit, contribute to interdomain branch length estimation and the shape of the universal tree.

2.3 Methods and Materials

2.3.1 Data

We downloaded the individual alignments from[30] `https://github.com/biocore/wol/tree/master/data/`, along with the genome metadata and the individual newick files. We checked each published tree for domain monophyly, and also performed approximately unbiased (AU)[86] tests to assess support for domain monophyly on the underlying sequence alignments using IQ-TREE 2.0.6[72]. The phylogenetic analyses were carried out using the 'reduced' subset of 1000 taxa outlined by the authors[30], for computational tractability. These markers were trimmed according to the protocol in the original paper[30], i.e. sites with >90% gaps were removed, followed by removal of sequences with >66% gaps. We also downloaded the Williams et al.[54] ("core"), Petitjean et al.[168] ("non-ribosomal") and Coleman et al.[24] ("bacterial") datasets from their original publications.

2.3.2 Annotations

Proteins used for phylogenetic analyses by Zhu et al.[30]), were annotated to investigate the selection of sequences comprising each of the marker gene families. To this end, we downloaded the protein sequences provided by the authors from the following repository: https://github.com/biocore/wol/tree/master/data/alignments/genes. To obtain reliable annotations, we analysed all sequences per gene family using several published databases, including the arCOGs (version from 2014)[189], KOs from the KEGG Automatic Annotation Server (KAAS; downloaded April 2019)[190], the Pfam database (Release 31.0)[191], the TIGRFAM database (Release 15.0)[192], the Carbohydrate-Active enZymes (CAZy) database (downloaded from dbCAN2 in September 2019)[193], the MEROPs database (Release 12.0)[194,195], the hydrogenase database (HydDB; downloaded in November 2018)[196], the NCBI- non-redundant (nr) database (downloaded in November 2018), and the NCBI COGs database (version from 2020). Additionally, all proteins were scanned for protein domains using InterProScan (v5.29-68.0; settings: –iprlookup –goterms)[197].

Individual database searches were conducted as follows: arCOGs were assigned using PSI-BLAST v2.7.1+ (settings: -evalue 1e-4 -show_gis -outfmt 6 -max_target_seqs 1000 -dbsize 100000000 -comp_based_stats F -seg no)[198]. KOs (settings: -E 1e-5), PFAMs (settings: -E 1e-10), TIGRFAMs (settings: -E 1e-20) and CAZymes (settings: -E 1e-20) were identified in all archaeal genomes using hmmsearch v3.1b2[199]. The MEROPs and HydDB databases were searchfed using BLASTp v2.7.1 (settings: -outfmt 6, -evalue 1e-20). Protein sequences were searched against the NCBI_nr database using DIAMOND v0.9.22.123 (settings: –more-sensitive –e-value 1e-5 –seq 100 –no-self-hits –taxonmap prot.accession2taxid.gz)[200]. For all database searches the best hit for each protein was selected based on the highest e-value and bitscore and all results are summarized in Supplementary Information Table S1 and full results are in the Data Supplement: Expanded_Bacterial_Core_Nonribosomal_analyses/Annotation_Tables/0_Annotation_tables_full/All_Zhu_marker_annotations_16-12-2020.tsv.zip. For InterProScan we report multiple hits corresponding to the individual domains of a protein using a custom script (parse_IPRdomains_vs2_GO_2.py).

Assigned sequence annotations were summarized and all distinct KOs and Pfams were collected and counted for each marker gene. KOs and Pfams with their corresponding descriptions were mapped to the marker gene file downloaded from the repository: https://github.com/biocore/wol/blob/master/data/markers/metadata.xlsx and used in the summarization of the 381 marker gene protein trees (Supplementary Information Table S1, see Appendix).

For manual inspection of single marker gene trees, KO and Pfam annotations were mapped to the tips of the published marker protein trees, downloaded from the repository: https://github.com/biocore/wol/tree/master/data/trees/genes. Briefly, the Genome ID, Pfam, Pfam description, KO, KO description, and NCBI Taxonomy string were collected from each marker gene annotation table and were used to generate mapping files unique to each marker gene phylogeny, which links the Genome ID to the annotation information (GenomeID—Domain—Pfam—Pfam Description—KO—KO Description). An in-house perl script replace_tree_names.pl (https://github.com/ndombrowski/Phylogeny_tutorial/tree/main/Input_files/5_required_Scripts) was used to append the summarized protein annotations to the corresponding tips in each marker gene tree. Annotated marker gene phylogenies were manually inspected using the following criteria including: 1) retention of reciprocal domain monophyly (Archaea and Bacteria) and 2) for the presence or absence of potential paralogous families. Paralogous groups and misannotated families present in the gene trees were highlighted and violations of search criteria were recorded in Supplementary Information Table S1, see Appendix.

2.3.3 Phylogenetic Analyses

Constraint analysis

We performed a maximum likelihood free topology search using IQ-TREE 2.0.6[72] under the LG+G4+F model, with 1000 ultrafast bootstrap replicates[201] on each of the markers from the expanded, bacterial, core and non-ribosomal sets. We also performed a constrained analysis with the same model, in order to find the maximum likelihood tree in which Archaea and Bacteria were reciprocally monophyletic. We then compared both trees using the approximately unbiased (AU)[86] test in IQ-TREE 2.0.6[72] with 10,000 RELL[202] bootstrap replicates. To evaluate the relationship between marker gene verticality and AB branch length, we calculated the difference in log-likelihood between the constrained and unconstrained trees to rank the genes from the expanded marker set. We then concatenated the top 20 markers (with the lowest difference in log-likelihood between the constrained and unconstrained trees, Appendix: Table 7) and iteratively added five markers with the next smallest difference in log-likelihood to the concatenate, this was repeated until we had concatenates up to 100 markers (with the lowest difference in log-likelihood) we inferred trees under LG+C10+G4+F in IQ-TREE 2.0.6, with 1000 ultrafast bootstrap replicates and calculated AB length.

Site and gene evolutionary rates

Site and gene evolutionary rates We inferred rates using the –rate option in IQ-TREE 2.0.6[72] for both the 381 marker concatenation from Zhu[30] and the top 5% of marker genes based on the results of the difference in log-likelihood between the constrained tree and free-tree search in the constraint analysis (above). We built concatenates for sites in the slowest and fastest rate categories, and inferred branch lengths from each of these concatenates using the tree inferred from the dataset as a fixed topology.

Split score analysis for expanded set markers

We used the previously described split score ranking procedure to quantify the number of taxonomic splits in the 381 marker gene phylogenies generated using the 1000-taxa subsample defined by Zhu et al.[30]. Taxonomic clusters were assigned using the Genome Taxonomy Database (GTDB) taxonomic ranks downloaded from the repository: `https://github.com/biocore/wol/tree/master/data/taxonomy/gtdb/`. Lineage-level monophyly was defined at the class level for all archaea (Arc1) and the phylum level for all bacteria (Bac0) (Supplementary Information Table S1, see Appendix). Of the original 10,575 genomes, 843 lacked corresponding GTDB assignments. For complete taxonomic coverage of the dataset, we used the GTDB Toolkit (GTDB-Tk) v0.3.2[203] to classify these genomes based on GTDB release 202. One of the 843 unclassified taxa (gid: G000715975) failed the GTDB-Tk quality control check resulting in no assignment, therefore we manually assigned this taxon to the Actinobacteriota based on the corresponding affiliation to the Actinobacteria in the NCBI taxonomic ranks provided in the genomic metadata downloaded from the repository:

https://github.com/biocore/wol/blob/master/data/genomes/. Additionally, two archaeal taxa within the Poseidoniia_A (gids: G001629155, G001629165) were manually assigned to the archaeal class MGII (Supplementary Information Table S1, see Appendix).

Plotting

Split score statistical analyses were performed using R 3.6.3[204]. All other statistical analyses were performed using R 4.0.4[205], and data were plotted with ggplot2[206].

2.4 Results and Discussion

Genes from the expanded marker set are not widely distributed in Archaea

The 381 gene set was derived from a larger set of 400 genes used to estimate the phylogenetic placement of new lineages as part of the PhyloPhlAn method[173] and applied to a taxonomic selection that included 669 Archaea and 9906 Bacteria[30]. Perhaps reflecting the focus on Bacteria in the original application, the phylogenetic distribution of the 381 marker genes in the expanded set varies substantially (Figure 6) (and Supplementary Information Table S1, see Appendix), with many being poorly represented in Archaea. Specifically, 41% of the published gene trees (https://biocore.github.io/wol/,[30]) contain less than 25% of the sampled archaea, with 14 and 68 of these trees including zero or ≤ 10 archaeal homologues, respectively. Across all of the gene trees, archaeal homologues comprise 0-14.8% of the dataset (Supplementary Information Table S1, see Appendix). Manual inspection of subsampled versions of these gene trees suggested that 317/381 did not possess an unambiguous branch separating the archaeal and bacterial domains (Supplementary Table S1, see Appendix). These distributions suggest that many of these genes are not broadly present in both domains, and that some might be specific to Bacteria.

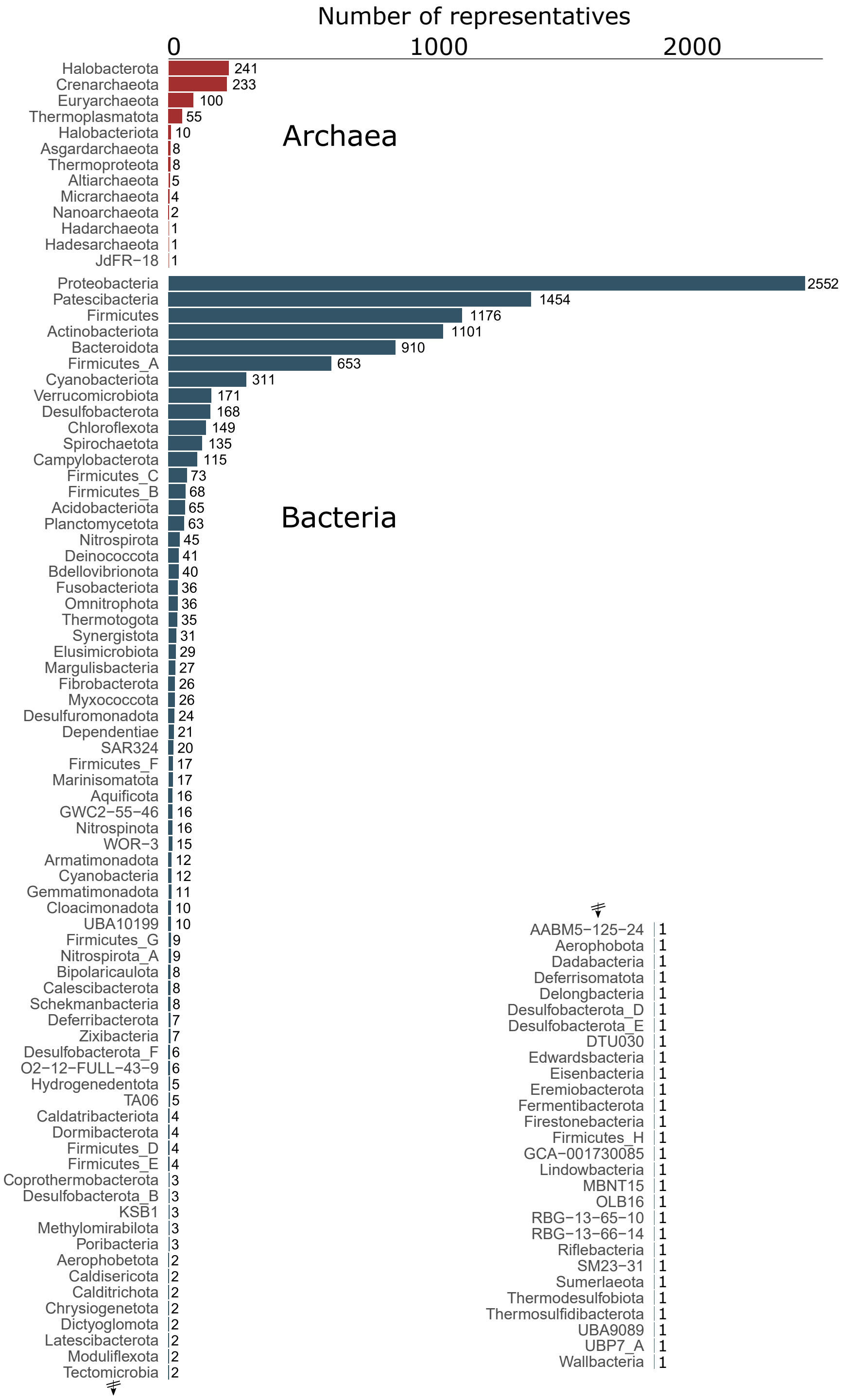

Figure 6 – Count of phyla from expanded marker set, GTDB-defined phyla for 10,575 archaeal and bacterial genomes in the expanded marker set analysis adapted from Moody et al.[176].

Conflicting evolutionary histories of individual marker genes and the inferred species tree In the published analysis of the 381 gene set[30], the tree topology was inferred using the supertree method ASTRAL[184], with branch lengths inferred on this fixed tree from a marker gene concatenation[30]. The topology inferred from this expanded marker set[30] is similar to previous trees[29,27] and recovers Archaea and Bacteria as reciprocally monophyletic domains, albeit with a shorter AB branch than in earlier analyses. However, the individual gene trees[30] differ regarding domain monophyly: Archaea and Bacteria are recovered as reciprocally monophyletic groups in only 22 of the 381 published[30] maximum likelihood (ML) gene trees of the expanded marker set (see Appendix: Table 7).

Since single gene trees often fail to strongly resolve ancient relationships, we used approximately-unbiased (AU) tests[86] to evaluate whether the failure to recover domain monophyly in the published ML trees is statistically supported. For computational tractability, we performed these analyses on a 1000-species subsample of the full 10,575-species dataset that was compiled in the original study[30]. For 79 of the 381 genes, we could not perform the test because the gene family did not contain any archaeal homologues (56 genes), or contained only one archaeal homologue (23 genes); in total, the 1000-species sample included 74 archaeal genomes. For the remaining 302 genes, domain monophyly was rejected at the 5% significance level (with Bonferroni correction, $p < 0.0001656$) for 151 out of 302 (50%) genes (Appendix: Table 7). As a comparison, we performed the same test on several smaller marker sets used previously to infer a tree of life[24,168,54]; none of the markers in those sets rejected reciprocal domain monophyly ($p < 0.05$ for all genes, with Bonferroni correction: Coleman: > 0.001724, Petitjean: > 0.001316, Williams: > 0.00102: Figure 7).

In what follows, we refer to four published marker gene sets as: i) the Expanded set (381 genes[30], ii) the Core set (49 genes[54], encoding ribosomal proteins and other conserved information-processing functions; itself a consensus set of several earlier studies[207,169,154]), iii) the Non-ribosomal set (38 genes, broadly distributed and explicitly selected to avoid genes encoding ribosomal proteins[168]), and iv) the Bacterial set (29 genes used in a recent analysis of bacterial phylogeny[24]) (Appendix: Table 8).

To investigate why 151 of the marker genes rejected the reciprocal monophyly of

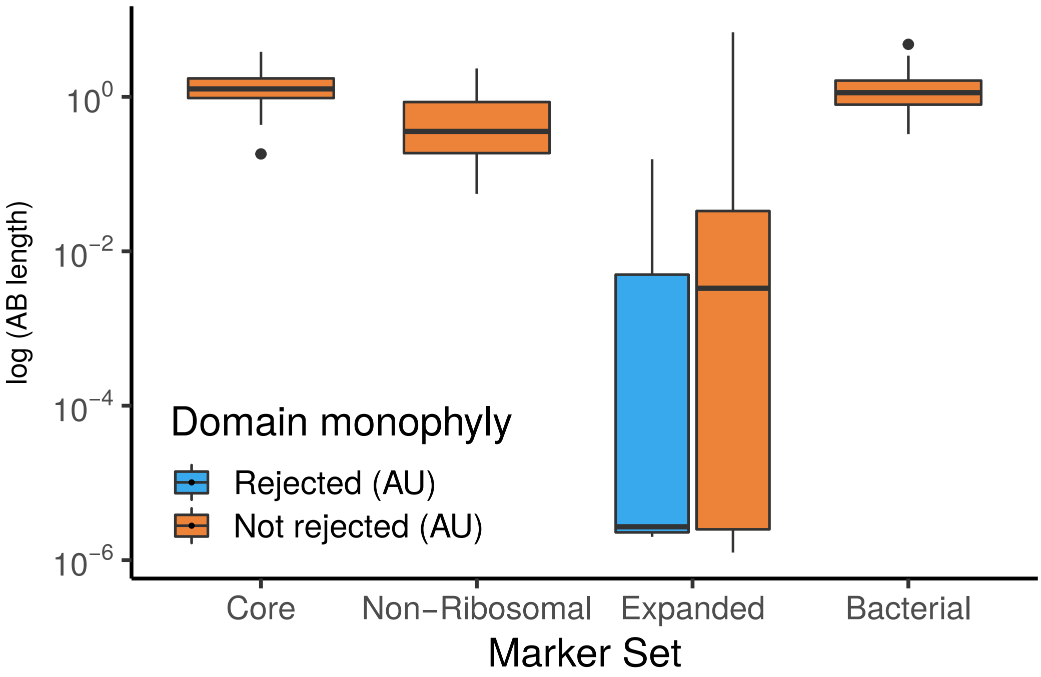

Figure 7 – A box-plot comparing marker gene monophyly to AB branch length. Expanded set genes that reject domain monophyly (p $<$ 0.05, AU test, with Bonferroni correction (see main text) support significantly shorter AB branch lengths when constrained to follow a domain monophyletic tree ($p = 3.653 \times 10^{-6}$, Wilcoxon rank-sum test). None of the marker genes from several other published analyses significantly reject domain monophyly (Bonferroni-corrected p $<$ 0.05, AU test) for all genes tested, consistent with vertical inheritance from LUCA to the last common ancestors of Archaea and Bacteria.

Archaea and Bacteria, we returned to the full dataset[30], annotated each sequence in each marker gene family by assigning proteins to KOs, Pfams, and Interpro domains, among others (Supplementary Information Table S1, see Methods for details) and manually inspected the tree topologies (Appendix: Table 7). This revealed that the major cause of domain polyphyly observed in gene trees was inter-domain gene transfer (in 359 out of 381 gene trees (94.2%)) and mixing of sequences from distinct paralogous families (in 246 out of 381 gene trees (64.6%)) (for a summary see Table7). For instance, marker genes encoding ABC-type transporters (p0131, p0151, p0159, p0174, p0181, p0287, p0306, p0364), tRNA synthetases (i.e. p0000, p0011, p0020, p0091, p0094, p0202), aminotransferases and dehydratases (i.e. p0073/4-aminobutyrate aminotransferase; p0093/3-isopropylmalate dehydratase) often comprised a mixture of paralogues.

Together, these analyses indicate that the evolutionary histories of the individual markers of the expanded set differ from each other and from the species tree. The original study investigated and acknowledged[30] the varying levels of congruence between the marker phylogenies and the species tree, but did not investigate the underlying causes. Our analyses establish the basis for these disagreements in terms of gene transfers and the mixing of orthologues and paralogues within and between domains. The estimation of genetic distance based on concatenation relies on the assumption that all of the genes in the supermatrix evolve on the same underlying tree; genes with different gene tree topologies violate this assumption and should not be concatenated because the topological differences among sites are not modelled, and so the impact on inferred branch lengths is difficult to predict. In practice, it is often difficult to be certain that all of the markers in a concatenate share the same gene tree topology, and the analysis proceeds on the hypothesis that a small proportion of discordant genes are not expected to seriously impact the inferred tree. However, the concatenated tree inferred from the expanded marker set differs from previous trees in that the genetic distance between Bacteria and Archaea is greatly reduced, such that the AB branch length appears comparable to distances among bacterial phyla[30]. Because an accurate estimate of the AB branch length has a major bearing on unanswered questions regarding the root of the universal tree[187], we next evaluated the impact of the conflicting gene histories within the expanded marker set on inferred AB branch length.

The inferred branch length between Archaea and Bacteria is shortened by inter-domain gene transfer and hidden paralogy

To investigate the impact of gene transfer and mixed paralogy on the AB branch length inferred by gene concatenations[30], we compared branch lengths estimated from markers on the basis of whether or not they rejected domain monophyly in the expanded marker set (Figure 7). To estimate AB branch lengths for genes in which the domains were not monophyletic in the ML tree, we first performed a constrained ML search to find the best gene tree that was consistent with domain monophyly for each family under the LG+G4+F model in IQ-TREE 2[72]. While it may seem strained to estimate the length of a branch that does not appear in the ML tree, we reasoned that this approach would provide insight into the contribution of these genes to the AB branch length in the concatenation, in which they conflict with the overall topology. AB branch lengths were significantly ($p = 3.653 \times 10^{-6}$, Wilcoxon rank sum test) shorter for markers that rejected domain monophyly (Bonferroni-corrected $p < 0.0001656$; Figure 7): mean AB branch length was 0.00668 substitutions/site for markers that significantly rejected domain monophyly, and 0.287 substitutions/site for markers that did not reject domain monophyly). This behaviour might result from marker gene transfer reducing the number of fixed differences between the domains, so that the AB branch length in a tree in which Archaea and Bacteria are constrained to be reciprocally monophyletic will tend towards zero as the number of transfers increases.

Interestingly, we found that a number estimated AB lengths for the constrained expanded set were close to zero (<0.00001), we suggest that this is the minimum possible tree length (a branch length of zero cannot exist) inferred when using a constrained topology. In this instance, the constraint of domain monophyly is so greatly at odds with the true phylogenetic relationship of this gene (assuming no paralogous or misannotated sequences) that the minimum possible branch length is inferred. We suggest that gene-trees which so greatly disagree with domain monophyly should be excluded as universal marker genes.

Substitutional saturation affect branch lengths and compositionally heterogeneous data

Substitutional saturation has previously been suggested as a factor in obscuring phylogenetic signal, sometimes leading to LBA attraction and other undesirable consequences[105,106,107,108]. In order to test the effect of substitutional saturation on the expanded marker set, we concatenated the fastest and slowest evolving sites for the whole supermartix (381 markers) and the top 5% of markers (see Table 1 and Figure 8). Initially, maximum likelihood trees were inferred from the corresponding supermatrix (one at each sampling levels, i.e. a 381-marker supermatrix and a top 5% supermatrix). Rates for each site in each supermatrix were calculated. The 25% fastest evolving sites and 25% slowest evolving (invariable sites were excluded) from each supermatrix were selected and concatenated separately and branch lengths were inferred on the previously inferred maximum likelihood topology. We found that the fastest evolving sites actually inferred branch lengths relatively shorter than the slowest evolving sites, suggesting substitutional saturation not captured by the model could be artificially reducing the AB branch length for the expanded dataset.

We also examined the average rate of the component gene trees for a smaller selection of datasets: the core, non-ribosomal, and expanded marker sets (see Figure 9). These results suggest that there is a much higher variation in the average rate of the expanded set in comparison with the core and non-ribosomal marker sets. As part of this test, we examined the effect of better fitting models which account for across-site compositional heterogeneity. Interestingly, the results were different for each marker set. For the core set, there was no real difference between the two models on the mean rate, however, for the non-ribosomal dataset, we found that using a better fitting model inferred significantly faster rates of substitution. The opposite was the case for the expanded set, using the model accounting for compositional heterogeneity actually inferred a reduced mean rate. This suggests that the simpler model for this dataset is actually underestimating the mean rate.

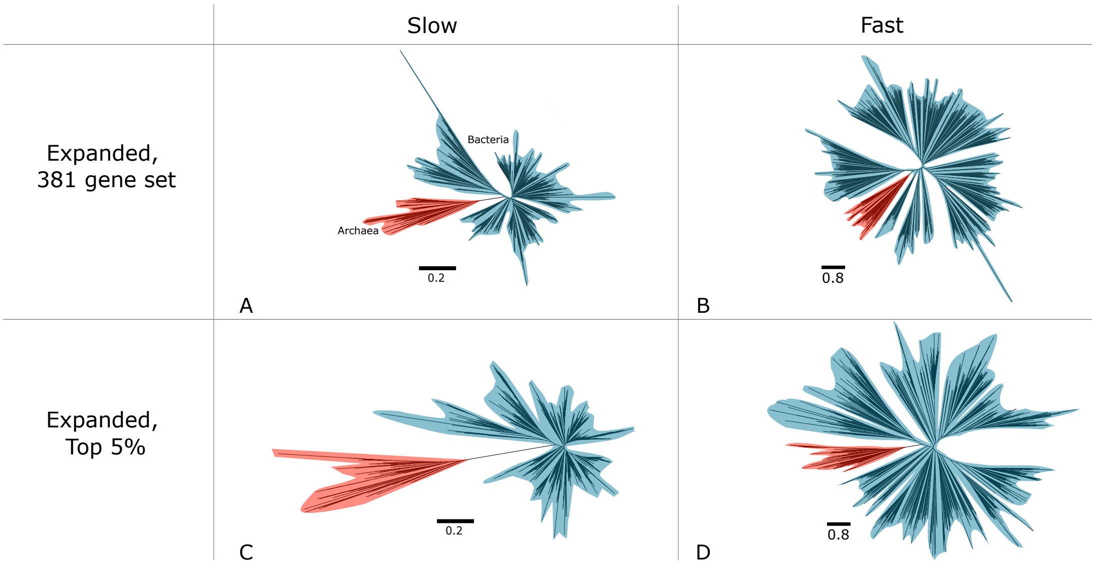

Figure 8 – Vertically-evolving genes and slow-evolving sites support a longer relative AB branch length. We estimated site-specific evolutionary rates for all marker genes in the expanded dataset (A-B), as well as for the 20 genes with the smallest ΔLL (top 5%) in that dataset (C-D). Concatenations based on the 25% slowest sites (A,C) and on the top 5% vertical genes (C,D) support a longer AB branch. This suggests that the inference of a short AB branch is impacted by both substitutional saturation and unmodelled inter-domain transfer of marker genes. Phylogenies were inferred under the LG+G4+F model. Branch lengths are the expected number of substitutions per site, as indicated by the scale bars. Alignment lengths in amino acids: A: 36797, B: 67274, C: 2736, D: 3884.

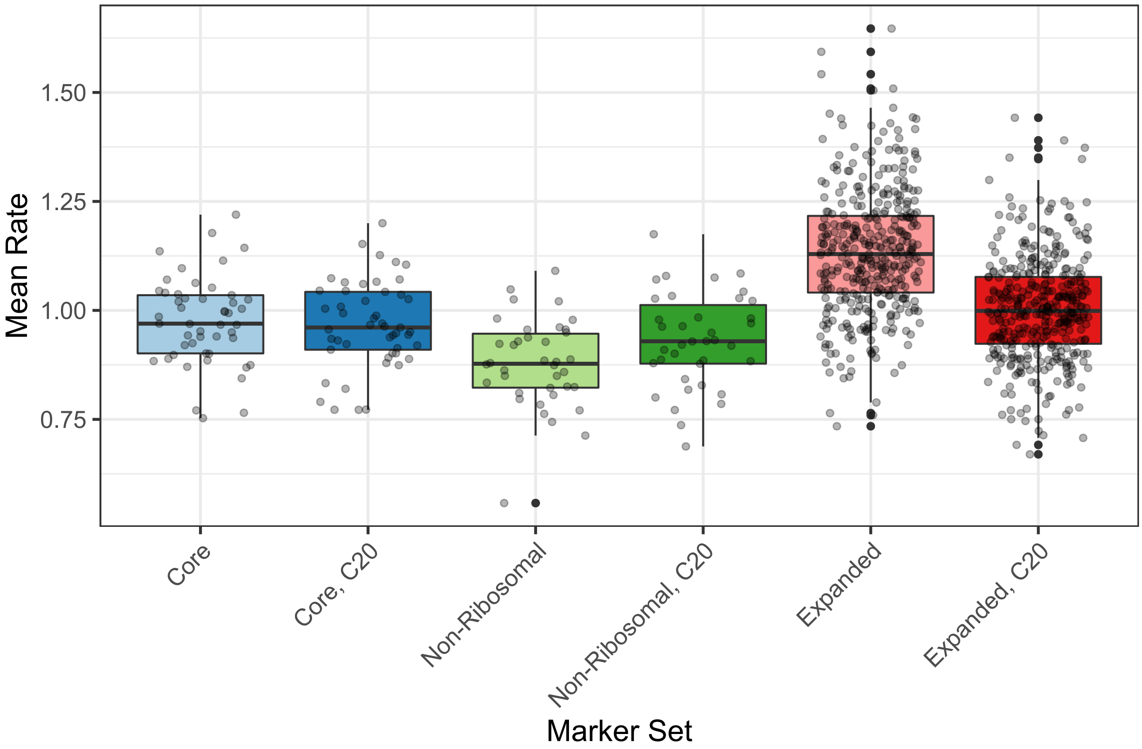

Figure 9 – The mean evolutionary rate of gene trees for the core, non-ribosomal and expanded marker-sets for simple models and models accounting for across-site compositional heterogeneity. Maximum likelihood trees inferred under LG+G+F or LG+C20+G+F with 1000 ultrafast bootstrap replicates. No significant difference was observed between the two models for the core dataset ($p = 0.7772$), however there was a significant difference between them for the non-ribosomal ($p = 0.03137$), and expanded ($p < 2.2 \times 10^{-16}$) marker sets. There was no significant difference between the core and expanded set under the LG+C20+G+F model ($p = 0.05382$) and the non-ribosomal and core set ($p = 0.1187$), but there was a significant difference between the non-ribosomal and expanded ($p = 0.0004726$) set under the same model. For the the trees inferred under LG+G+F, there was a significant difference between each dataset, core and non-ribosomal ($p = 1.842 \times 10^{-5}$, core and expanded ($p = 1.542 \times 10^{-13}$) and non-ribosomal and expanded ($p < 2.2 \times 10^{-16}$).

ID	**Gene**
p0023	infB
p0313	argS
p0076	miaB
p0004	tuf
p0383	serC
p0389	rpsG
p0072	pheT
p0010	rpoB
p0229	argJ
p0268	pyrB
p0027	leuS
p0182	gltB_1
p0109	mutS
p0346	deoA
p0162	era
p0030	pcrA1
p0138	lon
p0241	proB
p0071	polA
p0046	ruvB

Table 1 – IDs and gene names for the top 20 genes from the expanded marker set used for the fastest and slowest evolving sites supermatrix comparison (see Figure 8), and the molecular clock analysis (see Figure 29.)

Testing marker gene verticality

To test the hypothesis that phylogenetic incongruence among markers might reduce the inferred Archaea-Bacteria distance, we evaluated the relationship between AB distance and two complementary metrics of marker gene verticality: ΔLL , the difference in log likelihood between the constrained ML tree and the ML gene tree (a proxy for the extent to which a marker gene rejects the reciprocal monophyly of Bacteria and Archaea) and the "split score"[53], which measures the extent to which marker genes recover established relationships for defined taxonomic levels of interest (for example, at the level of domain, phylum or order), averaging over bootstrap distributions of gene trees to account for phylogenetic uncertainty (see Methods). We evaluated split scores at both the between-domain and within-domain levels (Figures 10-12) .

ΔLL and between-domain split score were positively correlated with each other (Figure 13) and negatively correlated with both AB stem length (Figures 10, 14) and relative AB distance (Figure 15), an alternative metric[30] that compares the average tip-to-tip distances within and between domains Figure 16. Interestingly, between-domain and within-domain split scores were strongly positively correlated (Figure 15), and the same relationships between within-domain split score, AB branch length and relative AB distance were observed (Figures 11, 12).

Overall, these results suggest that genes that recover the reciprocal monophyly of Archaea and Bacteria also evolve more vertically within each domain, and that these vertically evolving marker genes support a longer AB branch and a greater AB distance. Consistent with this inference, AB branch lengths estimated using concatenation decreased as increasing numbers of low-verticality markers (that is, markers with higher ΔLL) were added to the concatenate (Figure 17). These results suggest that inter-domain gene transfers reduce the AB branch length when included in a concatenation.

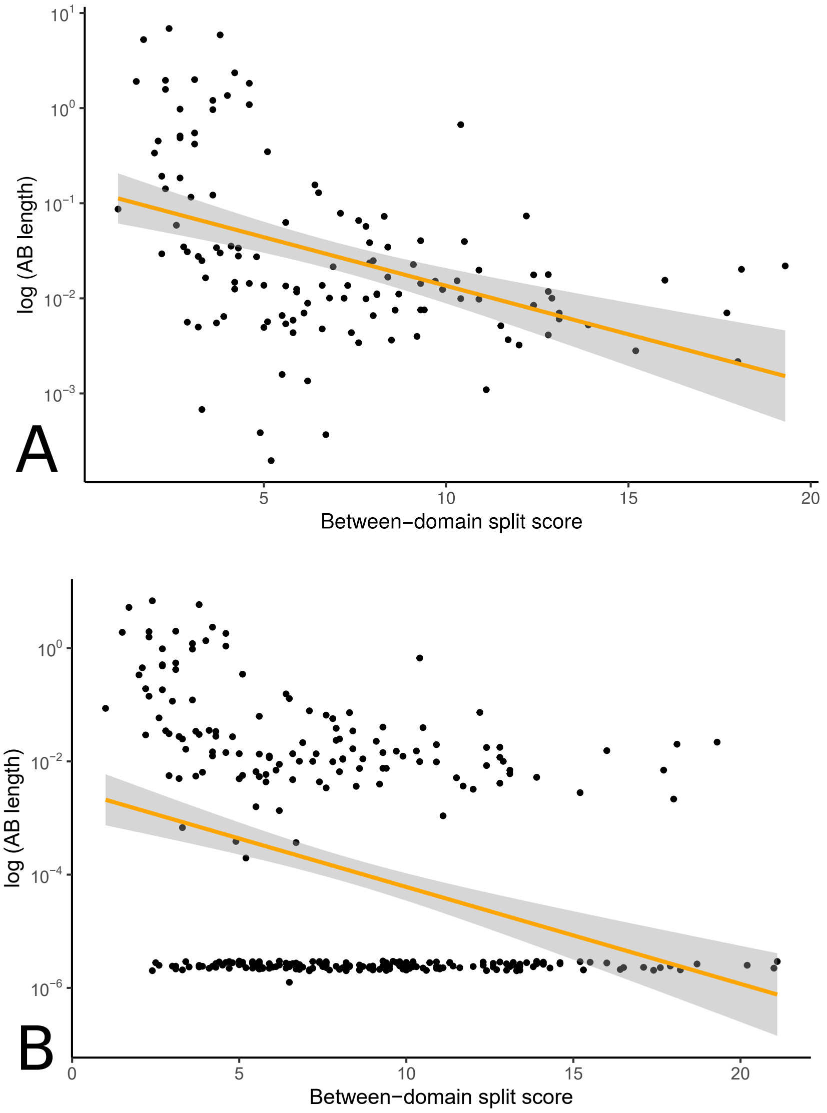

Figure 10 – Between-domain split score against AB branch length. Between-domain split score quantifies the extent to which marker genes recover monophyletic Archaea and Bacteria; a higher split score (see Methods) indicates the splitting of domains into multiple gene tree clades due to gene transfer, reciprocal sorting-out of paralogues or lack of phylogenetic resolution. Marker gene trees with AB length <0.00001 excluded (A): $p = 0.0005304$, R = -0.3043537, or included (B): $p = 1.111 \times 10^5$, R = -0.2498829, Pearson's correlation. The grey zones represent the standard error of the regression line.

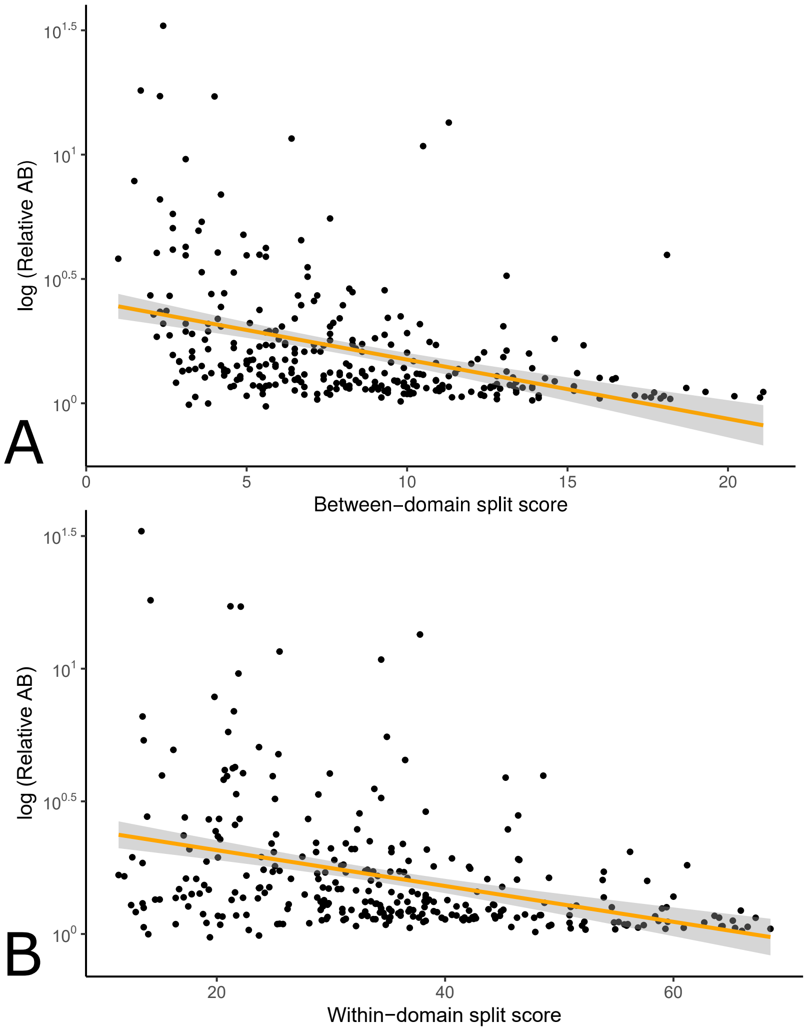

Figure 11 – Low-verticality marker genes have shorter relative AB distances. A higher split score denotes lower verticality. Between-domain split score (A), within-domain split score (B). A: $p = 2.572 \times 10^{-6}$, R = -0.2667739, B: $p = 5.685 \times 10^{-6}$, R = -0.257762. Pearson's correlation. The grey zones represent the standard error of the regression line.

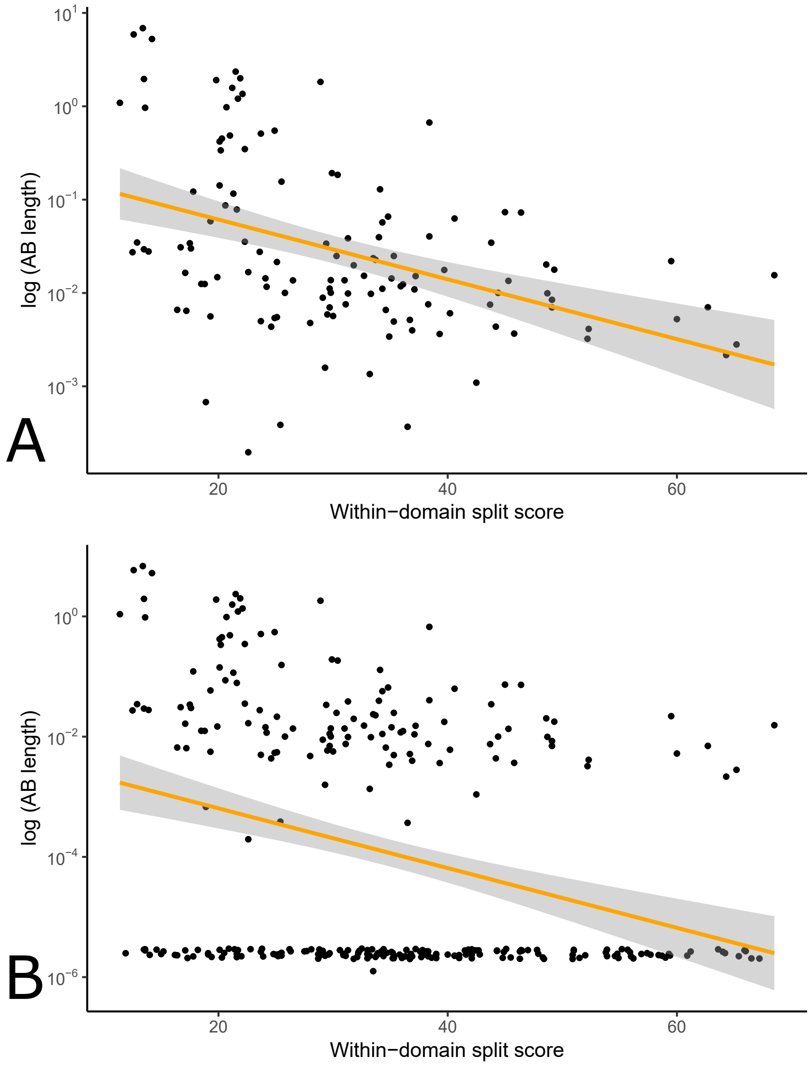

Figure 12 – Low-verticality marker genes have shorter AB branch lengths. Low-verticality marker genes (measured as within-domain split score) have shorter AB branch lengths. A higher split score denotes lower verticality. Marker gene trees with AB length <0.00001 excluded (A) or included (B). (A: p = 0.0001467, R = -0.3318924, B: $p = 7.498 \times 10^{-6}$, R = -0.2545369, Pearson's correlation.) The grey zones represent the standard error of the regression line.

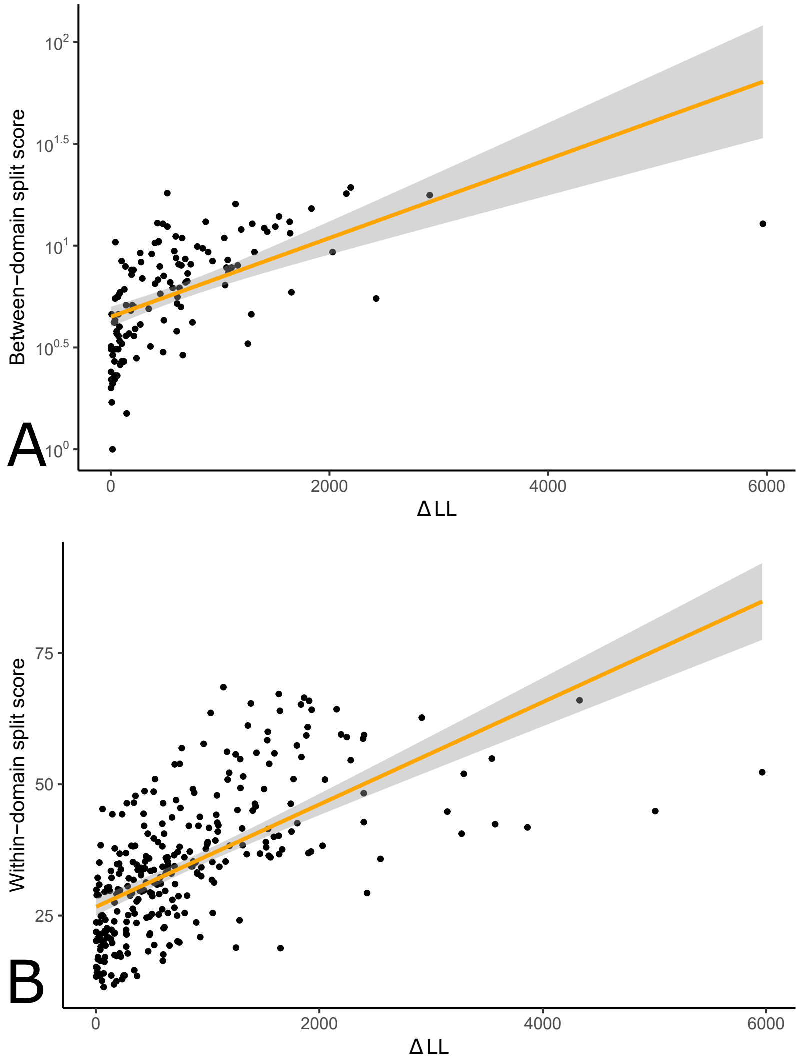

Figure 13 – Split score is strongly correlated with marker gene verticality. Both Between-domain split score (A) and within-domain split score (B) are strongly correlated with ΔLL, suggesting that both proxies capture a common signal of marker gene verticality. (A: $p < 2.2 \times 10^{-16}$, R = 0.6201967, B: $p < 2.2 \times 10^{-16}$, R = 0.836679. Pearson's correlation.) The grey zones represent the standard error of the regression line.

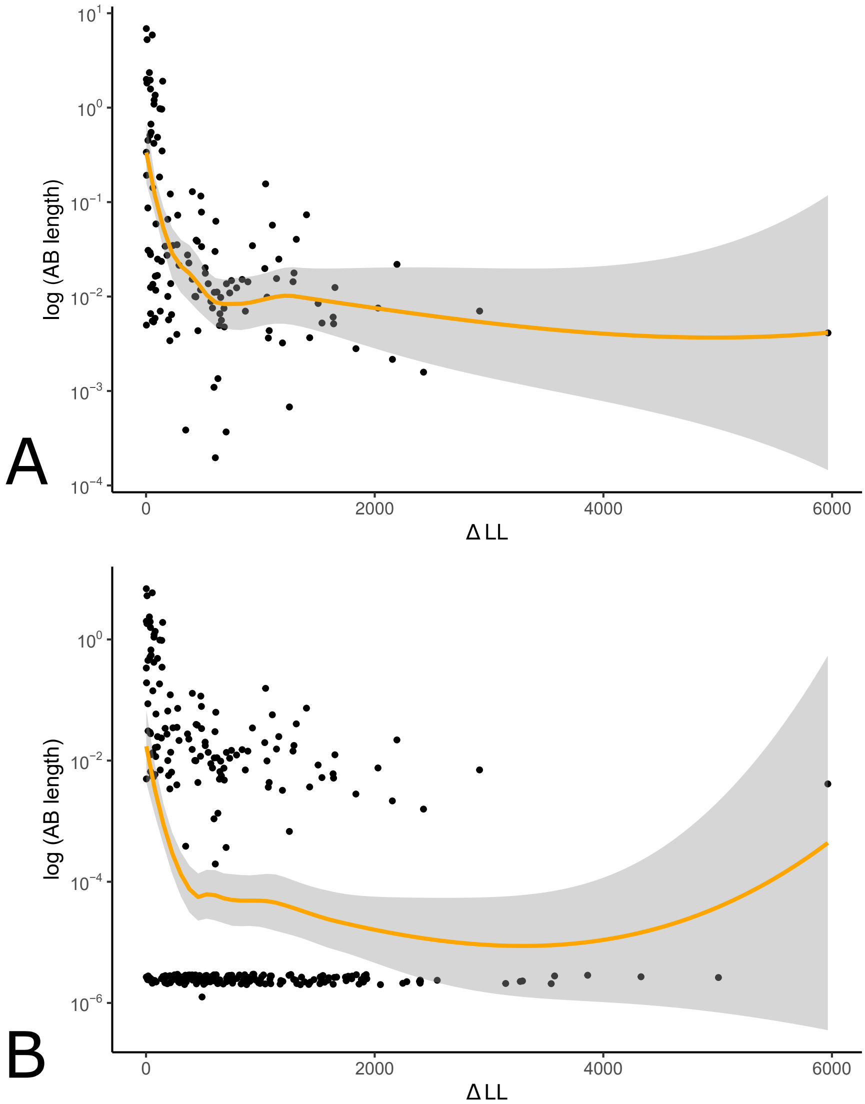

Figure 14 – ΔLL against AB branch length, which reflects the degree to which marker genes reject domain monophyly. Marker gene trees with AB length <0.00001 excluded (A) or included (B). A: p = 0.009013, R = -0.2317894, B: p = 0.00145, R = -0.1824596. We used a LOESS regression as the trendline here as the relationship varies across markers of differing verticality. Pearson's correlation. The grey zones represent the standard error of the regression line.

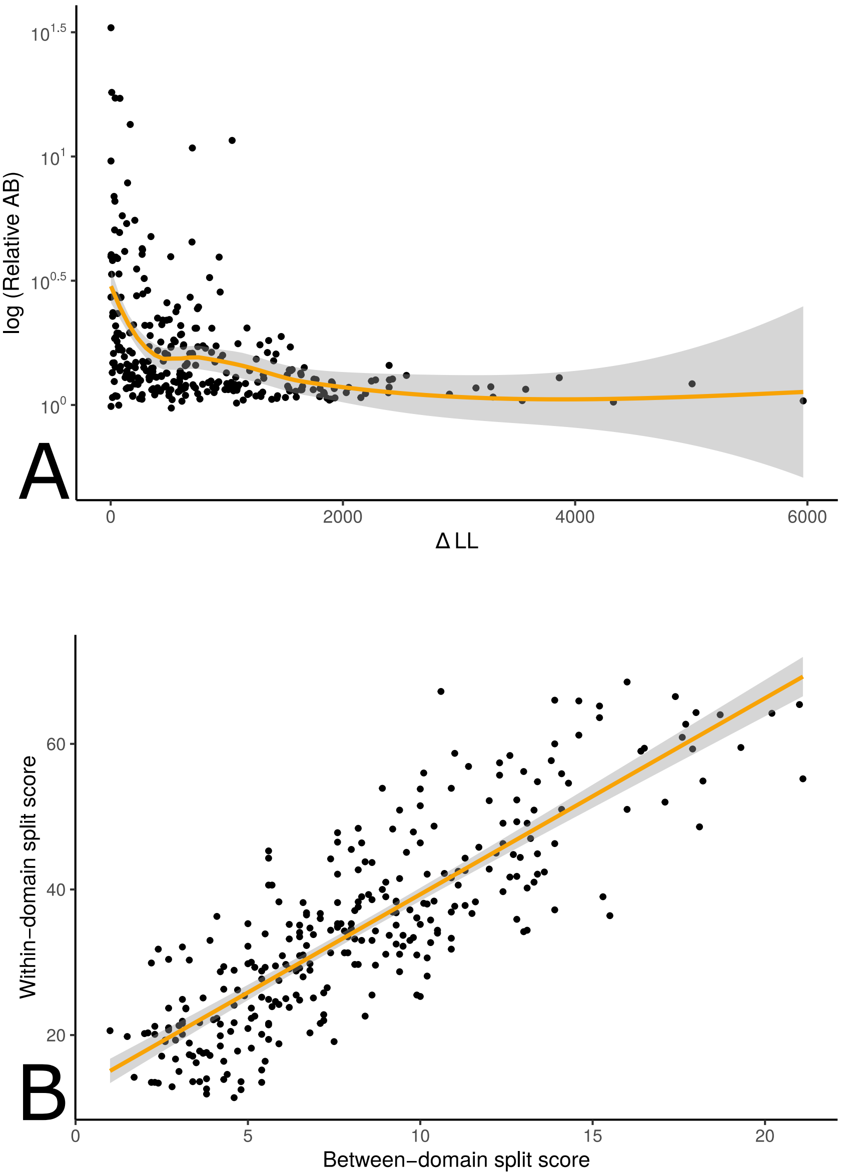

Figure 15 – All proxies for marker verticality are correlated. A: ΔLL against relative AB length, which reflects the degree to which marker genes reject domain monophyly (p = 0.0001051, R = -0.2213292). B: Between- and within-domain split scores are positively correlated (R = 0.836679, $p < 2.2 \times 10^{-16}$, Pearson's correlation), indicating that markers which recover Archaea and Bacteria as monophyletic also tend to recover established within-domain relationships. The grey zones represent the standard error of the regression line.

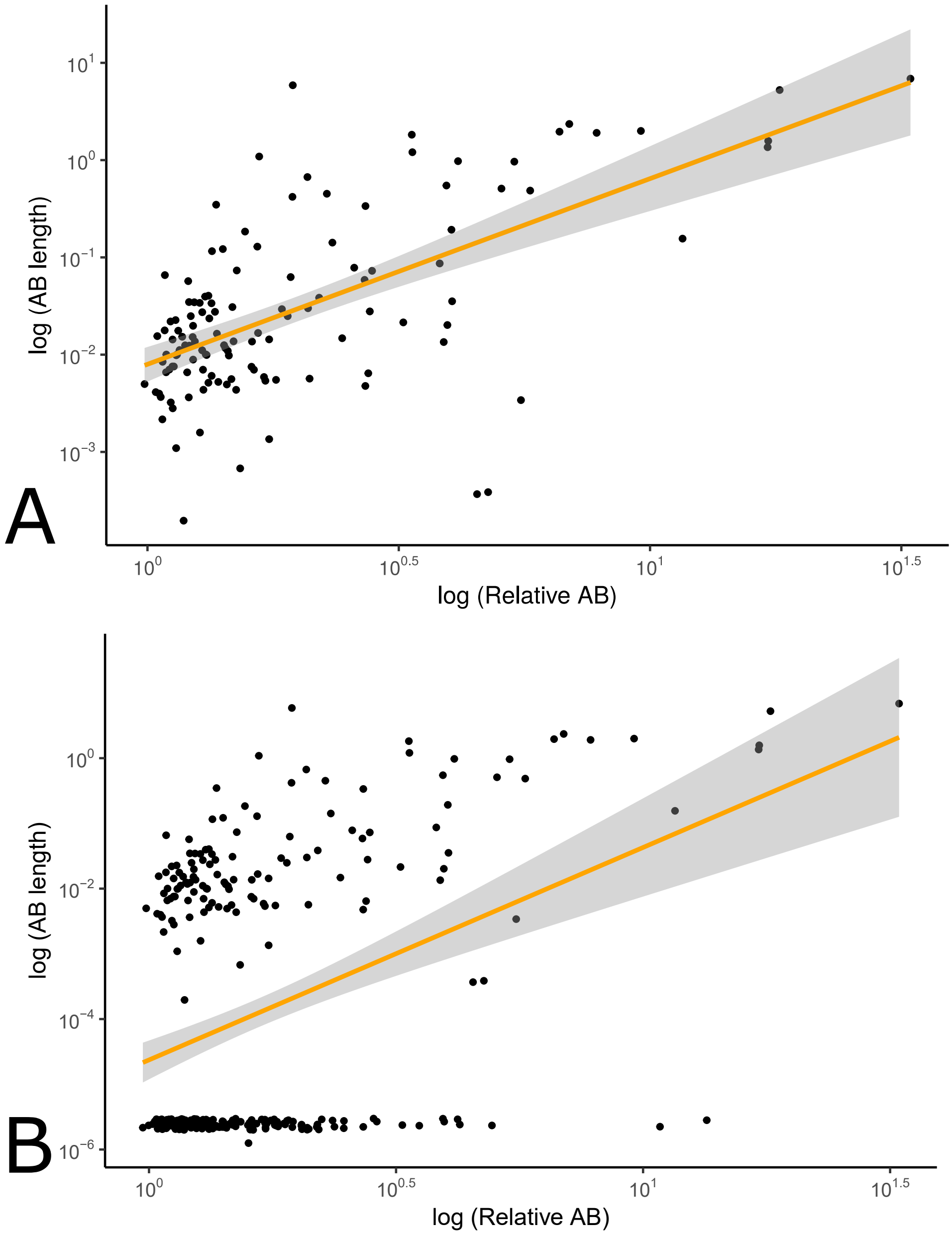

Figure 16 – Two measures of evolutionary proximity, AB branch length and relative AB distance, are positively correlated. Marker gene trees with AB length <0.00001 excluded (A) or included (B). A: R = 0.7426499, $p < 2.2 \times 10^{-16}$). B: $p < 2.2 \times 10^{-16}$, R = 0.706099. Pearson's correlation. The grey zones represent the standard error of the regression line.

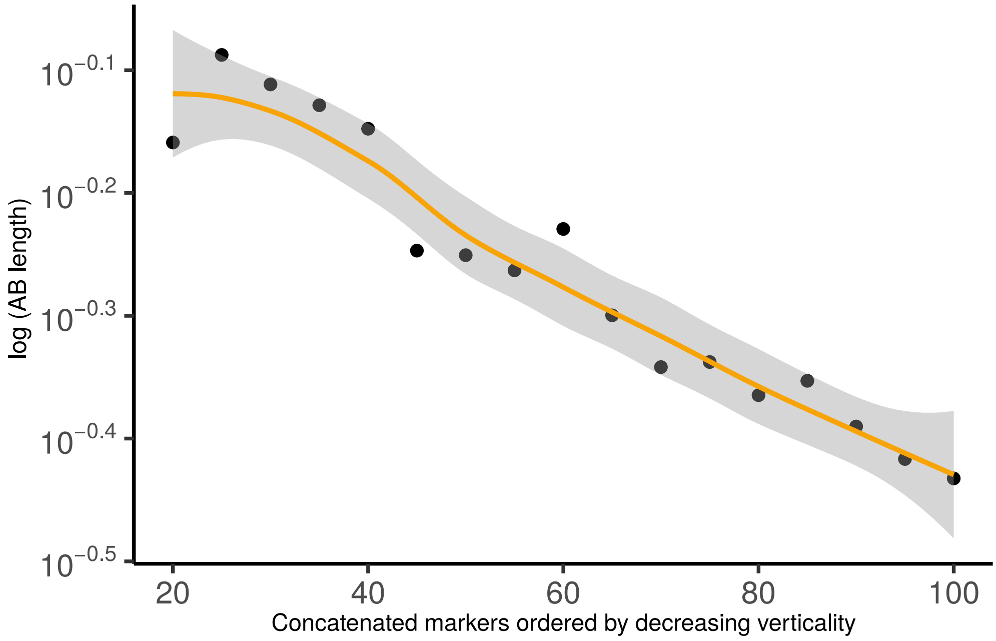

Figure 17 – Inferred AB length decreases as marker genes of lower verticality (larger ΔLL) are added to the concatenate. Marker genes were sorted by ΔLL, the difference in log-likelihood between the maximum likelihood gene family tree under a free topology search and the log-likelihood of the best tree constrained to obey domain monophyly. The grey zones represent the standard error of the regression line.

Do vertically evolving genes experience higher rates of sequence evolution?

An alternative explanation for the positive relationship between marker gene verticality and AB branch length could be that vertically-evolving genes experience higher rates of sequence evolution. For a set of genes that originate at the same point on the species tree, the mean root-to-tip distance (measured in substitutions per site, for gene trees rooted using the MAD method[175]) provides a proxy of evolutionary rate. Mean root-to-tip distances were significantly positively correlated with ΔLL, between-domain split score and relative AB distance (see Figure 18), indicating that vertically-evolving genes evolve relatively more slowly (note that higher values of ΔLL and split score denote lower verticality). Thus, the longer AB branches (Figure 19) of vertically-evolving genes do not appear to result from a faster evolutionary rate for these genes. Taken together, these results indicate that the inclusion of genes that do not support the reciprocal monophyly of Archaea and Bacteria, or their constituent taxonomic ranks, in the universal concatenate explain the reduced estimated AB branch length.

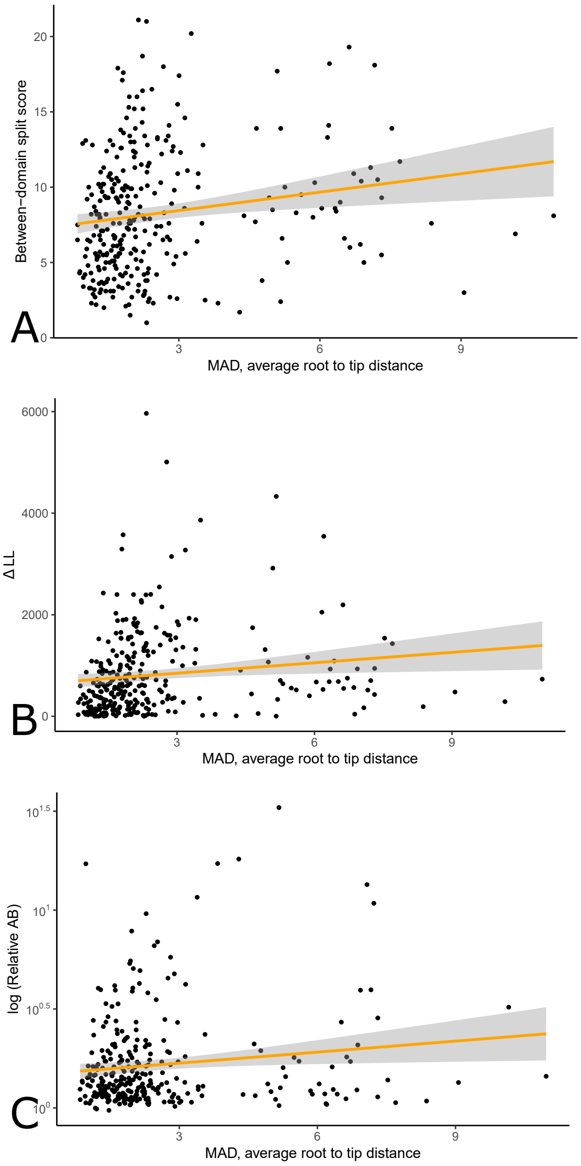

Figure 18 – Lower verticality is correlated with a higher evolutionary rate. Low-verticality genes as measured by between-domain split score (A), ΔLL (B) and relative AB distance (C) have a higher evolutionary rate (as measured by mean root-to-tip distance on MAD-rooted gene trees). Less vertically-evolving marker genes evolve faster as do markers with a higher relative AB distance although the effect is moderate. A: p = 0.002947, R = 0.1705415. B: p = 0.01506, R = 0.1397803. C: p = 0.007435, R = 0.1537479, Pearson's correlation. The grey zones represent the standard error of the regression line.

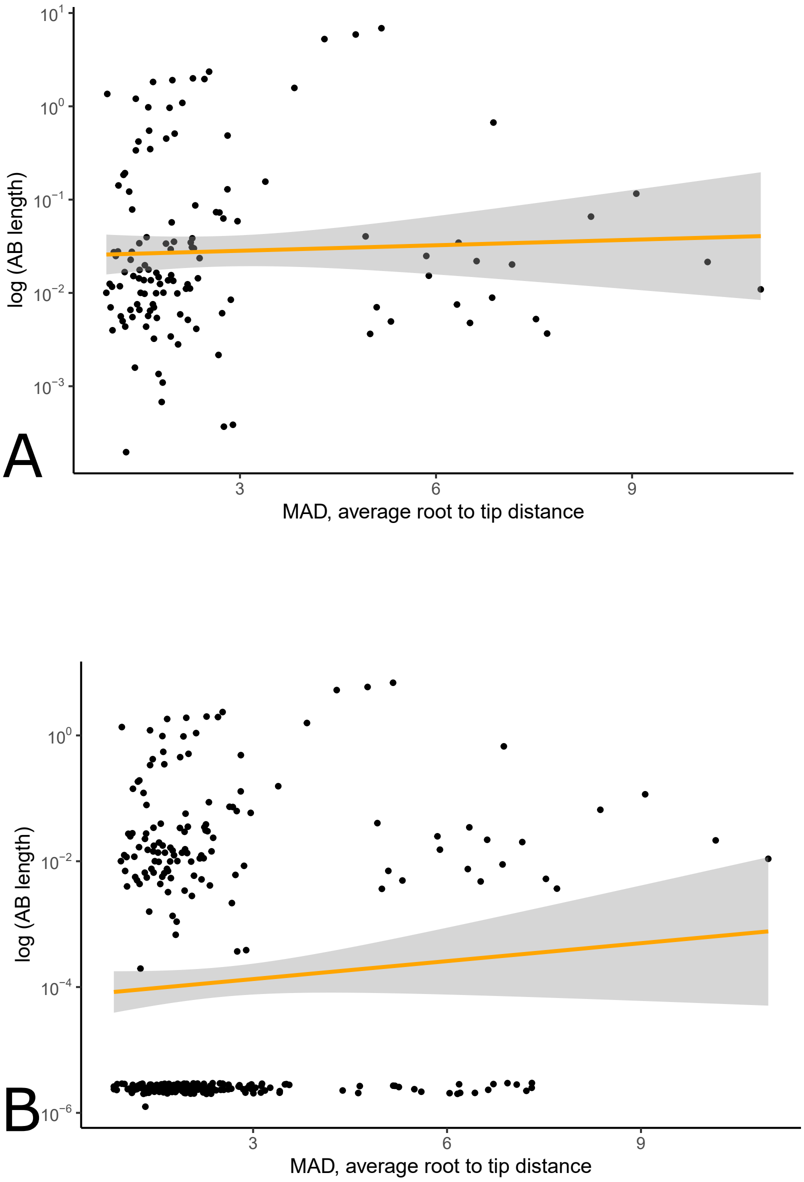

Figure 19 – No evidence for a relationship between AB branch length and gene evolutionary rate (average MAD root-to-tip distance). We did not detect a significant correlation between AB length and rate (average MAD-root to tip distance) when excluding (A) or including (B) markers with AB length <0.00001. A: p = 0.2025, R = 0.1143076. B: p = 0.0761, R = 0.102226. Pearson's correlation. The grey zones represent the standard error of the regression line.

2.5 Conclusion

Our analysis of a range of published marker gene datasets[168,169,54,30] indicates that the choice of markers and models is important for inference of deep phylogeny from concatenations, in agreement with an existing body of literature[208,209,78]. Phylogenies inferred from "core" genes involved in translation and other conserved cellular processes have provided one of the few available windows into the earliest period of archaeal and bacterial evolution. However, core genes comprise only a small proportion of prokaryotic genomes and have sometimes been viewed as outliers[30] in the sense that they are unusually vertical among prokaryotic gene families.

This means that they are among the few prokaryotic gene families amenable to concatenation methods, which are useful for pooling signal from individual weakly-resolved gene trees but which make the assumption that all sites evolve on the same underlying tree. If other gene families are included in concatenations, the results can be difficult to predict because differences in topology across sites are not modelled. Our analyses of the 381 gene expanded set suggest that this incongruence can lead to under-estimation of the evolutionary distance between Archaea and Bacteria, in the sense of the branch length separating the archaeal and bacterial domains. We note that alternative conceptions of evolutionary distance are possible; for example, in a phenetic sense of overall genome similarity, extensive HGT will increase the evolutionary proximity[30] of the domains so that Archaea and Bacteria may become intermixed at the single gene level. While such data can encode an important evolutionary signal, it is not amenable to concatenation analysis.

At the same time, it is clearly unsatisfactory to base our view of early evolution on a relatively small set of genes that appear to experience selective pressures rather distinct from the forces at play more broadly in prokaryotic genome evolution. These limitations are particularly unfortunate given the wealth of genome data now available to test hypotheses about early evolution. Exploring the evolutionary signal in more of the genome than hitherto is clearly a worthwhile endeavour. New methods, including more realistic models of gene duplication, transfer and loss[210,172], and extensions to supertree methods to model paralogy[211] and gene transfer, promise to enable genome-wide inference of prokaryotic history and evolutionary processes using methods that can account for the varying evolutionary histories of individual gene families.

Our analyses show that when removing HGTs and paralogous sites, removing poorly aligned sites, and using the best-fitting substitution model, a long branch between Archaea and Bacteria is retained (Figure 20). This is consistent with previous estimates of the tree of life[27,136,78] and refutes the short inter-domain branch length results of Zhu et al.[30] which we show to be a result of multiple compounding factors. We find that the core[54], non-ribosomal[168] and bacterial[24] marker sets do not intrinsically reject domain monophyly for any of the individual markers used. In the next chapter, we use the unique markers from a combination of all three of these datasets to infer a new prokaryotic tree, and assess the potential biases in functional classification for particular markers, in addition to the effects of using different models.

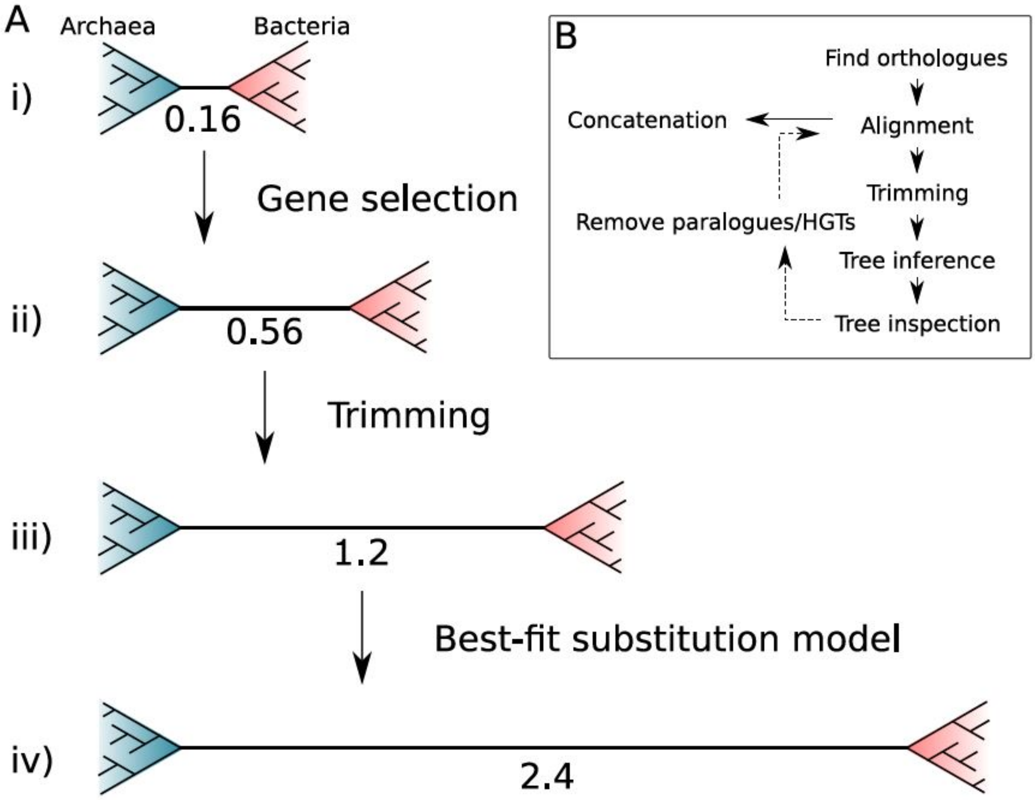

Figure 20 – The impact of marker gene choice, phylogenetic congruence, alignment trimming, and substitution model fit on estimates of the Archaea-Bacteria branch length. (A) Analysis using a site-homogeneous model (LG+G4+F) on the complete 381 gene expanded set (i) results in an AB branch substantially shorter than previous estimates. Removing the genes most seriously affected by inter-domain gene transfer (ii), trimming poorly-aligned sites (iii) using BMGE[212] in the original alignments (see below), and using the best-fitting site-heterogeneous model (iv) (LG+C60+G4+F) substantially increase the estimated AB length, such that it is comparable with published estimates from the "core" set: 3.3[54] and the consensus set of 27 markers identified in the present study: 2.5. Branch lengths measured in expected number of substitutions/site. (B) Workflow for iterative manual curation of marker gene families for concatenation analysis. After inference and inspection of initial orthologue trees, several rounds of manual inspection and removal of HGTs and distant paralogues were carried out. These sequences were removed from the initial set of orthologues before alignment and trimming. For a detailed discussion of some of these issues, and practical guidelines on phylogenomic analysis of multi-gene datasets, see Kapli et al.[208] for a useful review.

3 AN ESTIMATE OF THE DEEPEST BRANCHES OF THE TREE OF LIFE FROM ANCIENT VERTICALLY-EVOLVING GENES

Author's contribution

Parts of this chapter form part of a publication[176]:

Moody, E. R. R., Mahendrarajah, T. A., Dombrowski, N., Clark, J. W., Petitjean, C., Offre, P., Szöllősi, G. J., Spang, A., and Williams, T. A. (2022). An estimate of the deepest branches of the tree of life from ancient vertically-evolving genes. *eLife*, 11:e66695

The project was conceived by Tom A. Williams, Anja Spang and Edmund R. R. Moody. Individual gene trees, tree processing, concatenation, manual curation and individual gene tree inspection, statistical analyses, tree rooting, rate inference, archaeal-bacterial branch length estimation, tree length estimation, and relative archaeal-bacterial distance calculations were carried out by Edmund R.R. Moody. Split scores were carried out by Tara Mahendrarajah and Nina Dombrowski on gene trees generated by Edmund R. R. Moody. Tara Mahendrarajah manually performed taxonomy counts and taxon sampling. James W. Clark performed the dating analysis. Edmund R.R. Moody wrote the paper with comments and suggestions from other co-authors.

3.1 Abstract

Estimations of the tree of life rely on selections of universal marker genes thought to reflect the true underlying signal of vertical inheritance first described by Darwin[7]. Here, we infer a novel set of universal marker genes and determine the effect of model selection, functional classification and substitutional saturation have on the subsequent topology and branch lengths. We also infer a tree of life using a selection of vertically evolving universal marker genes, and find results consistent with previous single-domain analyses.

3.2 Introduction

One of the ways we can infer a tree of life is through phylogenomic analyses, although some have suggested there are issues with such the concept of a single species tree of life[148], the alternatives, such as using whole-genome data[147] or supertree[152,30] methods present their own problems such as HGT, paralogy and conflicting phylogenetic signal (see chapter two). Traditionally the way such trees have been inferred is through the concatenation of a handful of single-copy genes thought to represent the underlying species tree[160,161,136,27,54]. However, almost no set of genes will be free from duplications, transfers and losses[172], which could obscure the results, and as such manual curation and selection of such markers is an important stage in inferring the tree of life.

Previous analyses[168,30] have suggested that the traditional universal marker sets comprised mainly of ribosomal proteins could be biased. The long branches inferred from concatenations of these proteins could be a result of an artefact or a genuine acceleration of evolution in ribosomal proteins for the branch between the two primary domains of life. This might be the result of an accumulation of compensatory substitutions at the interaction surfaces among the protein subunits of the ribosome[168,213], or as a compensatory response to the addition or removal of ribosomal subunits early in evolution[168]. In this chapter, we test for evidence of an acceleration in ribosomal genes in comparison to other functional classes of proteins, using evidence from trees inferred from supermatrices and individual genes.

Other potential sources of bias in tree inference can arise from a failure of models to account for compositional heterogeneity[77,214,186,54] or substitutional saturation[105,106,107] in the data. To test these effects, as well as reduce any potential biases in our results, we perform several analyses testing different models and examine how the inferred archaeal-bacterial branch length varies between slow- and fast-evolving sites.

In molecular phylogenetics, branch lengths are usually measured in expected numbers of substitutions per site, with a longer branch corresponding to a greater degree of genetic change. Long branches can therefore result from high evolutionary rates, long periods of absolute time, or a combination of the two. If a sufficient number of fossils are available for calibration, molecular clock models can, in principle, disentangle the contributions of these effects. However, limited fossil data[215] is currently available to calibrate early divergences in the tree of life[216,217,218,219], and as a result, the age and evolutionary rates of the deepest branches of the tree remain highly uncertain.

Using several existing marker sets[168,54,24] as a starting point, we identified single-copy orthologues of these markers in a representative set of archaeal and bacterial genomes. We performed manual inspection and curation of these orthologues resulting in a set of 27 universal marker genes, chosen for their verticality. We concatenate these markers into a supermatrix and infer an updated tree of prokaryte and estimate the divergence times of the last common ancestors of Archaea, Bacteria and all extant life.

3.3 Methods and Materials

3.3.1 Phylogenetic Analyses

COG assignment for the Core, Non-Ribosomal, and Bacterial marker genes

First, all gene sequences in the three published marker sets (core, non-ribosomal, and bacterial) were annotated using the NCBI COGs database (version from 2020). Sequences were assigned a COG family using hmmsearch v3.3.2[199] (settings: -E 1e-5) and the best hit for each protein sequence was selected based on the highest e-value and bit score. To assign the appropriate COG family for each marker gene, we quantified the percentage distribution of all unique COGs per gene and selected the family representing the majority of sequences in each marker gene. Accounting for overlap, this resulted in 95 unique COG families (Appendix: Table 9) from the original 119 total marker genes across all three published datasets (Supplementary Information Table S2, see appendix). Orthologues corresponding to these 95 COG families were identified in the 700 genomes (350 Archaea, 350 Bacteria, Supplementary Information Table S3, see appendix) using hmmsearch v3.3.2 (settings: -E 1e-5). The reported BinID and protein accession were used to extract the sequences from the 700 genomes, which were used for subsequent phylogenetic analyses.

Marker gene inspection and analysis

We aligned these 95 marker gene sequence sets using MAFFT L-INS-i[220] and removed poorly-aligned positions with BMGE[212]. We inferred initial maximum likelihood trees (LG+G4+F) for all 95 markers and mapped the KO and Pfam domains and descriptions, inferred from the annotation of the 700 genomes, to the corresponding tips (see above). Manual inspection took into consideration monophyly of Archaea and Bacteria and the presence of paralogues and other signs of contamination (HGT, LBA). Accordingly, single gene trees that failed to meet reciprocal domain monophyly were excluded, and any instances of HGT, paralogous sequences, and LBA artefacts were manually removed from the remaining trees, resulting in 54 markers across the three published datasets that were subject to subsequent phylogenetic analysis (LG+C20+G4+F) and further refinement (see below).

Ranking markers based on split score

We applied an automated marker gene ranking procedure devised previously (the split score[53]) to rank each of the 54 markers that satisfied reciprocal monophyly based on the extent to which they recovered established phylum-, class- or order-level relationships within the archaeal and bacterial domains (Supplementary Information Table S4, see appendix). The script quantifies the number of splits, or occurrences where a taxon fails to cluster within its expected taxonomic lineage across all gene phylogenies. Briefly, we assessed monophyletic clustering using phylum-, class-, and order-level clades within Archaea (Cluster1) in combination with Cluster0 (phylum) or Cluster3 (i.e. on class-level if defined and otherwise on phylum-level; Supplementary Information Table S4, see appendix) for Bacteria. We then ranked the marker genes using the following split score criteria: the number of splits per taxon and the splits normalized to the species count. The percentage of split phylogenetic groups was used to determine the highest ranking (top 50%) markers.

Concatenation

Based on the split score ranking of the 54 marker genes (above), the top 50% (27 markers, Supplementary Information Table S4, see appendix) marker genes were manually inspected using the criteria as defined above, and contaminating sequences were manually removed from the individual sequence files. Following inspection, marker protein sequences were aligned using MAFFT-L-INS-i[221] and trimmed using BMGE (version 1.12, under default settings)[212]. We concatenated the 27 markers into a supermatrix, which was used to infer a maximum-likelihood tree (Figure 27, under

LG+C60+G4+F), evolutionary rates (see below), and rate-category supermatrices as well as to perform model performance tests (see below). We also concatenated the non-ribosomal and ribosomal (COG category J) markers from the 27 and 54 marker sets (Appendix: Table 10) into four more supermatrices and inferred maximum likelihood trees under (LG+C60+G4+F) (Table 1). Two additional supermatrices were constructed from the 54 markers, one before manual removal of apparent HGTs and one after the removal, with both sets of markers aligned and trimmed in the same way as the other datasets (see above). We also inferred a maximum likelihood tree under LG+C60+G4+F from a supermatrix consisting of a concatenation of 25 marker genes, after removing COG0480 and COG5257 as these have been previously implicated to be unsuitable for universal markers[222].

Evolutionary rates of sites and genes

We inferred rates using the –rate option in IQ-TREE 2.0.6[72] to explore the differences in rates for the 27 marker set. We built concatenates for sites in the slowest and fastest rate categories, and inferred branch lengths from each of these concatenates using the tree inferred from the corresponding dataset as a fixed topology.

Molecular clock analyses

Molecular clock analyses were devised to test the effect of genetic distance on the inferred age of LUCA. Following the approach of Zhu et al.[30], we subsampled the alignment to 100 species. Five alternative alignments were analysed, representing conserved sites across the entire alignment, randomly selected sites across the entire alignment (data from Zhu et al.[30], only ribosomal marker genes, the top 5% of marker genes according to ΔLL and the top 5% of marker genes further trimmed under default settings in BMGE 1.12[212] (these alignments can be found be found in the dating directory, within the 'Vertically_Evolving_Marker_Analyses directory at: https://doi.org/10.6084/m9.figshare.13395470. Divergence time analyses were performed in MCMCTree[223] under a strict clock model. We used the normal approximation approach, with branch lengths estimated in codeml[223] under the LG+G4 model.

In each case, a fixed tree topology was used alongside a single calibration on the Cyanobacteria-Melainabacteria split. The calibration was modelled as a uniform prior distribution between 2.5 and 2.6 Ga, with a 2.5% probability that either bound could be exceeded. For each alignment, four independent MCMC chains were run for 2,000,000 generations to achieve convergence. We repeated the clock analyses under a relaxed (independent rates drawn from a lognormal distribution) clock model with an expanded sampling of fossil calibration (Supplementary Information Table S6, see appendix). We repeated the analyses with two approaches to defining the maximum age calibration. The first used the moon-forming impact (4.52 Ga), under the provision that no forms of life are likely to have survived this event. The second relaxed this assumption, instead using the estimated age of the universe (13.7 Ga) as a maximum. Analyses were performed as above.

3.4 Results and Discussion

Finding ancient vertically-evolving genes

To estimate the AB branch length and the phylogeny of prokaryotes using a dataset that resolves some of the issues identified above, we performed a meta-analysis of several previous studies to identify a consensus set of vertically-evolving marker genes (see Appendix: Table 10). We identified unique markers from these analyses by reference to the COG ontology[53,224], extracted homologous sequences from a representative sample of 350 archaeal and 350 bacterial genomes (Figure 21), and performed iterative phylogenetics and manual curation to obtain a set of 54 markers that recovered archaeal and bacterial monophyly (see Methods).

Prior to manual curation, non-ribosomal markers had a greater number of HGTs and cases of mixed paralogy. In particular, for the original set of 95 unique COG families (see ‘Phylogenetic analyses’ in Methods), we rejected 41 families based on the inferred ML trees, either due to a large degree of HGT, paralogous gene families or LBA. For the remaining 54 markers, the ML trees contained evidence of occasional recent HGT events. Strict monophyly was violated in 69% of the non-ribosomal and 29% of the ribosomal families.

We manually removed the individual sequences which violated domain monophyly before re-alignment, trimming, and subsequent tree inference (see Methods). These results imply that manual curation of marker genes is important for deep phylogenetic analyses, particularly when using non-ribosomal markers.

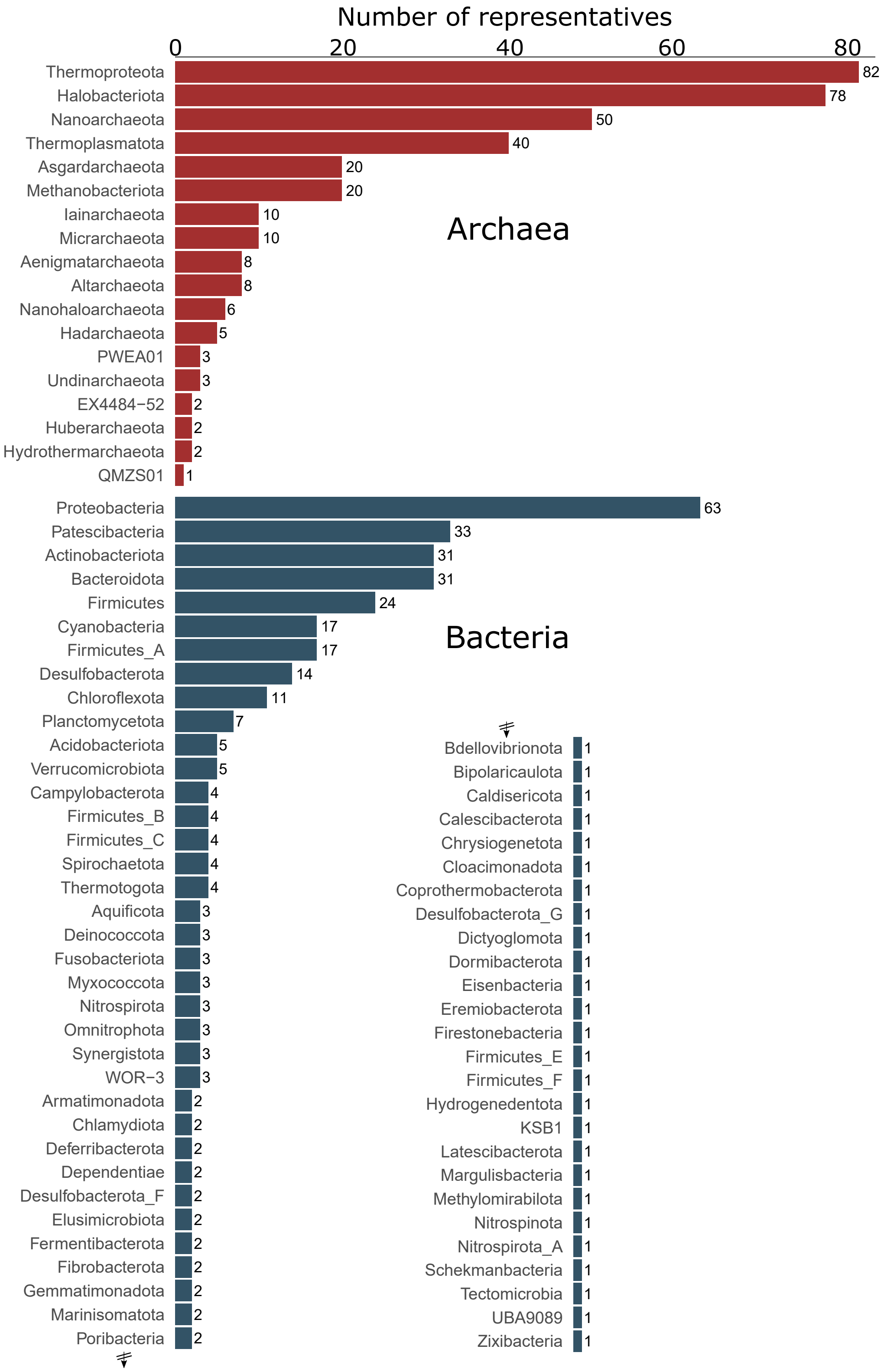

Figure 21 – Count of phyla from the marker set presented here. GTDB-defined phyla for 700 archaeal and bacterial genomes in our marker set analysis adapted from Moody et al.[176].

	Pruned	**Unpruned**
AB branch length	1.945	1.734

Table 2 – The effect of including known HGTs and paralogues in a concatenation on AB branch length.Unpruned refers to a ML tree inferred under LG+C60+F+G from a concatenation of the 54 markers without manual curation and removal of HGTs, the pruned set is the same markers but with the manual curation step, i.e. HGTS and/or paralogous sequences are pruned before sequence alignment and concatenation.

Comparison of within-domain split scores for these 54 markers indicated that markers that better resolved established relationships within each domain also supported a longer AB branch length (Figure 22). Further, the AB branch length inferred from a concatenation of the 54 marker genes increased moderately following pruning of recent HGTs consistent with the hypothesis that non-modelled inter-domain HGTs reduce the overall estimate of AB branch length when included in concatenations (see Table 2).

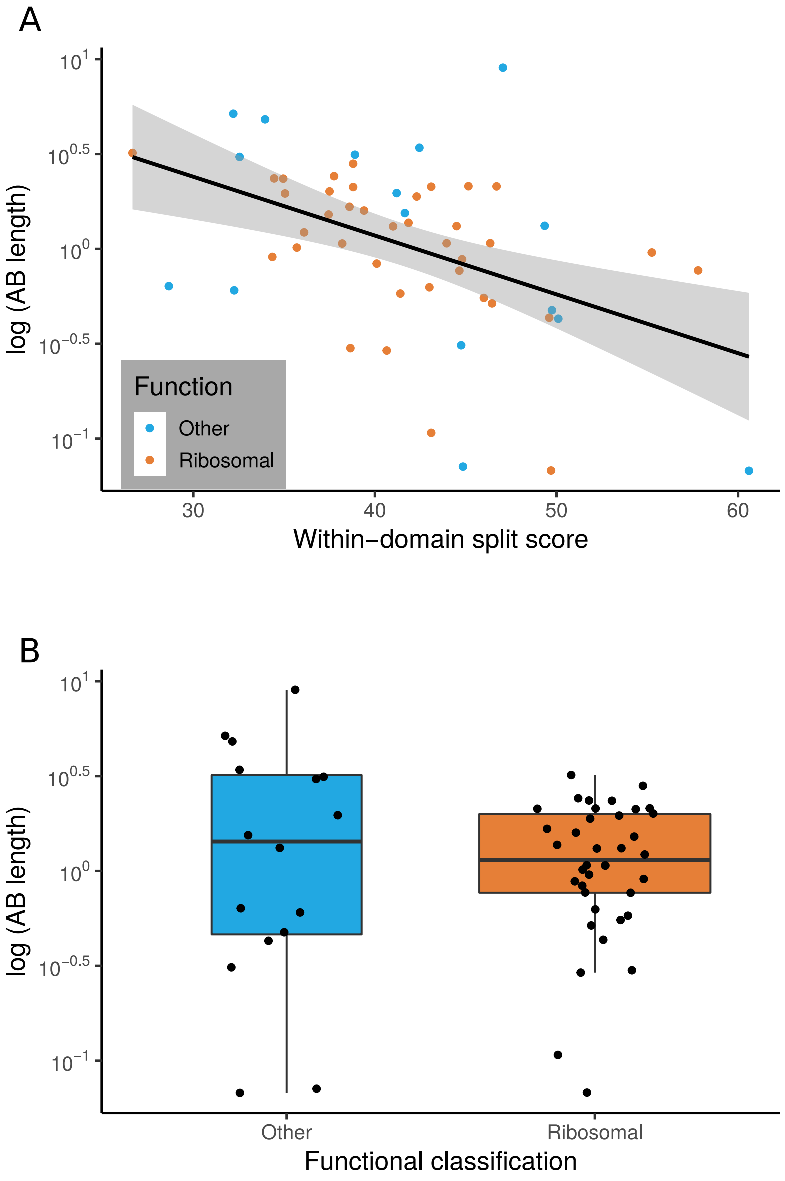

Figure 22 – The relationship between marker gene verticality, AB branch length, and functional category. (A) Vertically-evolving phylogenetic markers have longer AB branches. The plot shows the relationship between a proxy for marker gene verticality, within-domain split score (a lower split score denotes better recovery of established within-domain relationships, see Methods), and AB branch length (in expected number of substitutions/site) for the 54 marker genes. (Continued on the following page.)

Figure 22 – Continued. Marker genes with higher split scores (that split established monophyletic groups into multiple subclades) have shorter AB branch lengths (p = 0.0311, R = 0.294). Split scores of ribosomal and non-ribosomal markers were statistically indistinguishable (p = 0.828, Figure 2, Figure Supplement 1). (B) Among vertically-evolving marker genes, ribosomal genes do not have a longer AB branch length. The plot shows functional classification of markers against AB branch length using 54 vertically-evolving markers. We did not obtain a significant difference between AB branch lengths for ribosomal and non-ribosomal genes (P = 0.6191, Wilcoxon rank-sum test). The grey zones represent the standard error of the regression line.

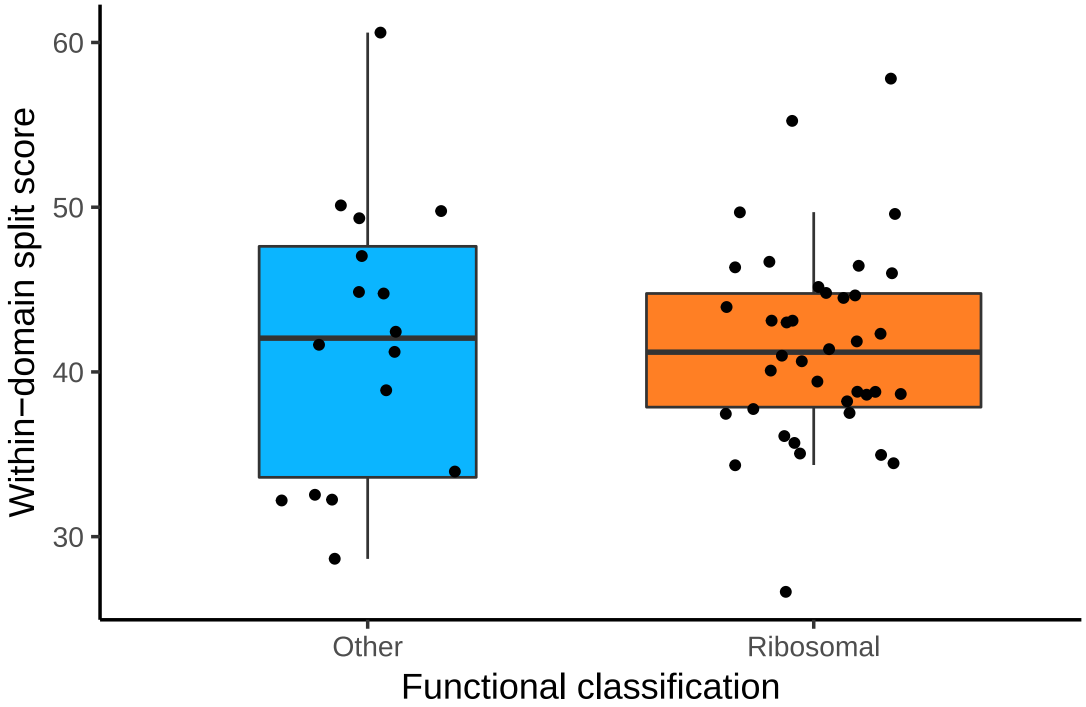

Figure 23 – Among vertically-evolving marker genes, the split scores of ribosomal and non-ribosomal proteins are statistically indistinguishable. The plot shows functional classification of markers (ribosomal markers or other) against the split score (a higher split score denotes greater disagreement with established within-domain relationships) using 54 markers from the new analysis. After removing genes that do not appear to have been vertically inherited since the divergence of Archaea and Bacteria, split scores of ribosomal and nonribosomal markers were statistically indistinguishable (P = 0.8275, Wilcoxon rank sum test).

	AB branch length		Total tree length		AB branch length as a proportion of total tree length	
	Ribosomal	**Non-ribosomal**	**Ribosomal**	**Non-ribosomal**	**Ribosomal**	**Non-ribosomal**
27 marker set	1.9541	3.7723	250.7255	239.8203	0.0078	0.0157
54 marker set	1.8647	2.5414	271.3327	288.8470	0.0069	0.0088

Table 3 – AB branch lengths and AB branch lengths as a proportion of total tree length inferred from ribosomal and non-ribosomal concatenates are similar. The data do not support a faster evolutionary rate for ribosomal proteins on the AB branch compared to other kinds of ancient proteins.

Trees inferred from ribosomal marker genes do not have a longer AB branch length

Traditional universal marker sets include many ribosomal proteins[161,177,160,27,54,225]. If ribosomal proteins experienced accelerated evolution during the divergence of Archaea and Bacteria, this might lead to the inference of an artefactually long AB branch length[168,30]. To investigate this, we plotted the inter-domain branch lengths for 38 and 16 ribosomal and non-ribosomal genes, respectively, comprising the 54 marker genes set. We found no evidence that there was a longer AB branch associated with ribosomal markers than for other vertically-evolving "core" genes (Figure 22B; mean AB branch length for ribosomal proteins 1.35 substitutions/site, mean for non-ribosomal 2.25 substitutions/site).

In order to investigate further, we concatenated ribosomal proteins and non-ribosomal proteins from the top 27 marker set and the 54 marker set (see Table 3). For both sets of markers (the 27 markers consisting of 21 ribosomal & 6 non-ribosomal proteins; the 54 marker set consisting of 38 ribosomal & 16 non-ribosomal proteins), we see a different pattern. In both cases we see an increase in the AB branch length for the concatenation inferred from the 16 non-ribosomal markers, as opposed to the 6 non-ribosomal markers. The trees inferred from the ribosomal concatenations had a consistently shorter AB length which was not markedly different in either the 38 or 16 ribosomal sets. Taken together, these results suggest that on average, ribosomal markers do not behave very differently from non-ribosomal markers or if so, appear to be have shorter AB branch lengths.

Four of the non-ribosomal markers 22 were shown to have higher AB lengths than any of the ribosomal markers. Although our results disagree with those of Zhu et al.[30], we do find some overlap in the markers they found to be outliers, a single-gene: rpoC (RNA-polymerase, subunit-β). Although the long AB branch reported here is nowhere near as long as reported by[30], one explanation could be an error in orthologue identification. Our manual curation process may go someway in alleviating this issue, but we still find a long AB stem. Another explanation could be due to the structural differences between the archaeal and bacterial RNA polymerase, of which there are many, the large number of homologous RNA polymerase proteins in Archaea could also reduce the need for such stringent conservation of sites[226].

Substitutional saturation and poor model fit contribute to underestimation of AB branch length

For the 27 most vertically evolving genes as ranked by within-domain split score, we performed an additional round of single gene tree inference and manual review to identify and remove remaining sequences which had evidence of HGT or represented distant paralogues. The resulting single gene trees are provided in the Data Supplement (10.6084/m9.figshare.13395470).

To evaluate the relationship between site evolutionary rate and AB branch length, we created two concatenations: the fastest sites (comprising the sites with highest probability of being in the fastest Gamma rate category; 868 sites) and the slowest sites (sites with highest probability of being in the slowest Gamma rate category, 1604 sites) and compared relative branch lengths inferred from the entire concatenate, using IQ-TREE 2 to infer site-specific rates (Figure 24).

Notably, the proportion of inferred substitutions that occur along the AB branch differs between the slow-evolving and fast-evolving sites. As would be expected, the total tree length measured in substitutions per site is shorter from the slow-evolving sites, but the relative AB branch length is longer (1.2 substitutions/site, or ~2% of all inferred substitutions, compared to 2.6 substitutions/site, or ~0.04% of all inferred substitutions for the fastest-evolving sites; see Figure 25 for absolute tree size comparisons).

Since we would not expect the distribution of substitutions over the tree to differ between slow-evolving and fast-evolving sites, this result suggests that some ancient changes along the AB branch at fast-evolving sites have been overwritten by more recent events in evolution — that is, that substitutional saturation leads to an underestimate of the AB branch length.

Another factor that has been shown to lead to an underestimation of genetic distance on deep branches is a failure to adequately model the site-specific features of sequence evolution[77,214,186,54,30]. Amino acid preferences vary across the sites of a sequence alignment, due to variation in the underlying functional constraints[77,80,227].

The consequence is that, at many alignment sites, only a subset of the twenty possible amino acids are tolerated by selection. Standard substitution models such as LG+G4+F are site-homogeneous, and approximate the composition of all sites using

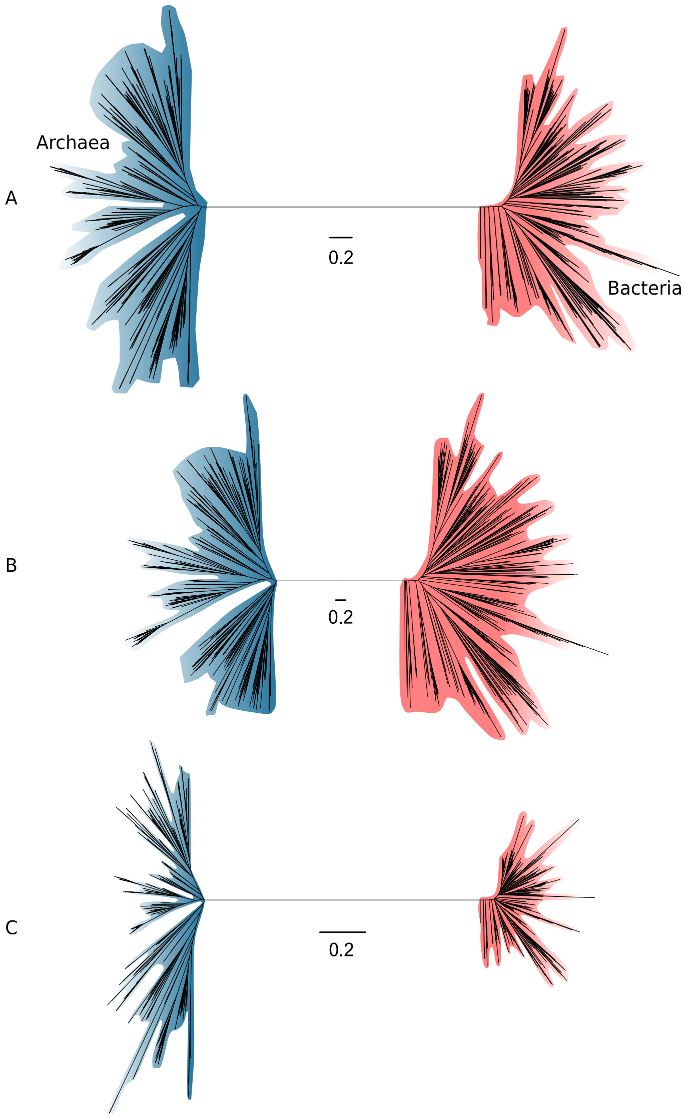

Figure 24 – Slow- and fast-evolving sites support different shapes for the universal tree. (A) Tree of Archaea (blue) and Bacteria (red) inferred from a concatenation of 27 core genes using the best-fitting model (LG+C60+G4+F); (B) Tree inferred from the fastest-evolving sites; (C) Tree inferred from the slowest-evolving sites. To facilitate comparison of relative diversity, scale bars are provided separately for each panel; for a version of this figure with a common scale bar for all three panels, see Figure 3 Figure Supplement 1. Slow-evolving sites support a relatively long inter-domain branch and less diversity within the domains (that is, shorter between-taxa branch lengths within domains). This suggests that substitution saturation (overwriting of earlier changes) may reduce the relative length of the AB branch at fast-evolving sites and genes.

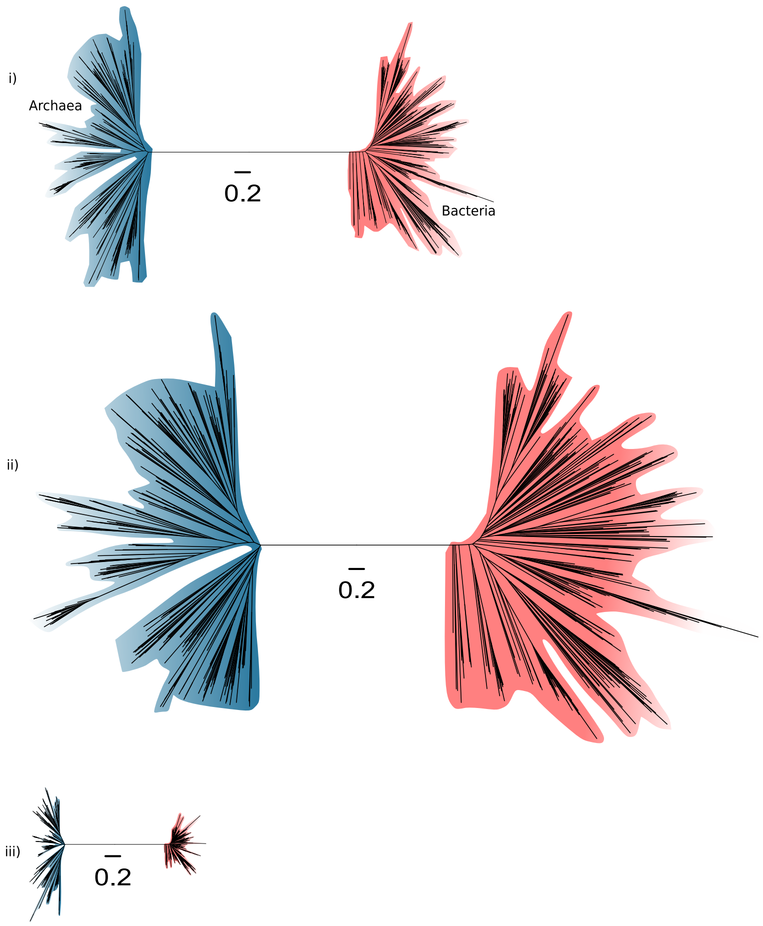

Figure 25 – Slow- and fast-evolving sites support different shapes for the universal tree. (i) Tree of Archaea (blue) and Bacteria (red) inferred from a concatenation of 27 core genes using the best-fitting model (LG+C60+G4+F); (ii) Tree inferred from the fastest-evolving sites; (iii) Tree inferred from the slowest-evolving sites. Identical scale bars are provided for comparison.

the average composition across the entire alignment. Such models underestimate the rate of evolution at highly constrained sites because they do not account for the high number of multiple substitutions that occur at such sites. The effect is that site-homogeneous models underestimate branch lengths when fit to site-heterogeneous data. Site-heterogeneous models have been developed that account for site-specific amino acid preferences, and these generally show improved fit to real protein sequence data (reviewed in Williams et al.[78]).

To evaluate the impact of substitution models on estimates of AB branch length, we assessed the fit of a range of models to the full concatenation using the Bayesian information criterion (BIC) in IQ-TREE 2. The AB branch length inferred under the best-fit model, the site-heterogeneous LG+C60+G4+F model, was 2.52 substitutions/site, ∼1.7-fold greater than the branch length inferred from the site-homogeneous LG+G4+F model (1.45 substitutions/site). Thus, substitution model fit has a major effect on the estimated length of the AB branch, with better-fitting models supporting a longer branch length (Table 4).

The same trends are evident when better-fitting site-heterogeneous models are used to analyse the expanded marker set: considering only the top 5% of genes by ΔLL score, the AB branch length is 1.2 under LG+G4+F, but increases to 2.4 under the best-fitting LG+C60+G4+F model (Figure 26). These results are consistent with Zhu et al.[30], who also noted that AB branch length increases as the model fit improves for the expanded marker dataset.

Overall, these results indicate that difficulties with modelling sequence evolution, either due to substitutional saturation or failure to model variation in site compositions, lead to an under-estimation of the AB branch length, both in published analyses and for the analyses of the new dataset presented here. As substitution models improve, we would therefore predict the estimates of the AB branch length to increase further.

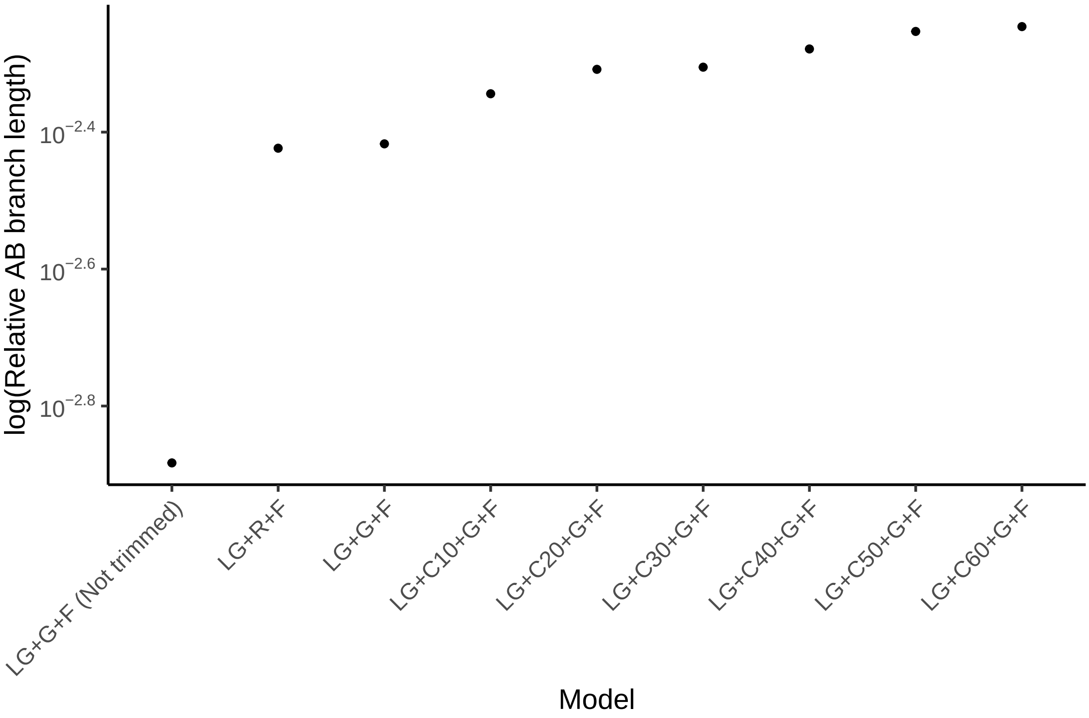

Figure 26 – The effect of modelling site compositional heterogeneity on AB branch length. Increasing the number of protein mixture profiles, as well as trimming poorly-aligned positions, is associated with a change in AB branch length on the expanded marker set. All analyses used LG exchangeabilities, four rate categories (Gamma-distributed or freely estimated), and included a general composition vector containing the empirical amino acid frequencies (+F). Modelling of site heterogeneity with the C10-C60 models increases the inferred AB branch length ∼2-fold. Trimming poorly-aligned sites slightly increases the AB branch estimation whereas relaxing the gamma rate categories slightly decreases estimation of AB branch length. LG (LG substitution matrix), G (four gamma rate categories), F (empirical site frequencies estimated from the data), C10-60 (number of protein mixture profiles used) R (four free rate categories which relax the assumption of a gamma distribution for rates, BMGE (trimming using Block Mapping and Gathering with Entropy)[212].

Substitution model	**BIC (ΔBIC)**	**AB branch length**
LG+G4+F	5935950.053	1.4491
LG+C20+G4+F	(152046.1)	2.1394
LG+C40+G4+F	(179126.7)	2.4697
LG+C60+G4+F	(189063.8)	2.5178

Table 4 – The inferred AB branch length from a concatenation of the top 27 markers (chosen by within-domain split score, see Appendix: Table 10) using a simple model versus models which account for site compositional heterogeneity. Models that account for across-site compositional heterogeneity fit the data better (as assessed by lower BIC scores) and infer a longer AB branch length.

Our maximum likelihood phylogeny of the primary domains inferred from the most vertical marker genes

The phylogeny of the primary domains of life inferred from the 27 most vertically-evolving genes using the best-fitting LG+C60+G4+F model (Figure 29) is consistent with recent single-domain trees inferred for Archaea and Bacteria independently[24,53,52], although the deep relationships within Bacteria are poorly resolved, with the exception of the monophyly of Gracilicutes (Figure 29).

Our results are also in good agreement with a recent estimate of the universal tree based on a different marker gene selection approach[228]. In that study, marker genes were selected based on Tree Certainty, a metric that quantifies phylogenetic signal based on the extent to which markers distinguish between different resolutions of conflicting relationships[229].

Our analysis placed the Candidate Phyla Radiation (CPR)[25] as a sister lineage to Chloroflexi (Chloroflexota) rather than as a deep-branching bacterial superphylum. While this contrasts with initial trees suggesting that CPR may represent an early diverging sister lineage of all other Bacteria[25,27,29], our finding is consistent with recent analyses that have instead recovered CPR within the Terrabacteria[24,228,23]. Together, these analyses suggest that the deep-branching position of CPR in some trees may be a result of long branch attraction, a possibility that has been raised previously[27,230].

The deep branches of the archaeal subtree are generally well-resolved and recover DPANN (51% bootstrap support), and Asgards (100% bootstrap support), and TACK Archaea (75% bootstrap support) as monophyletic clades in agreement with a range of previous studies[164,53,164,181,52]. We also find support for the placement of Methanonatronarchaeia[231] distant to Halobacteria as one of the earliest branches of the Methanotecta (Figure 27) in agreement with recent analyses, suggesting that their initial placement with Halobacteria[231] may be an artefact of compositional attraction[232,53,51,35].

We obtained moderate (92%) bootstrap support for the branching of some euryarchaeota with the TACK+Asgard clade: the Hadesarchaea+Persephonarchaea were resolved as the sister group to TACK+Asgards with moderate (92%) support, with this entire lineage branching sister to a strongly supported (100%) clade

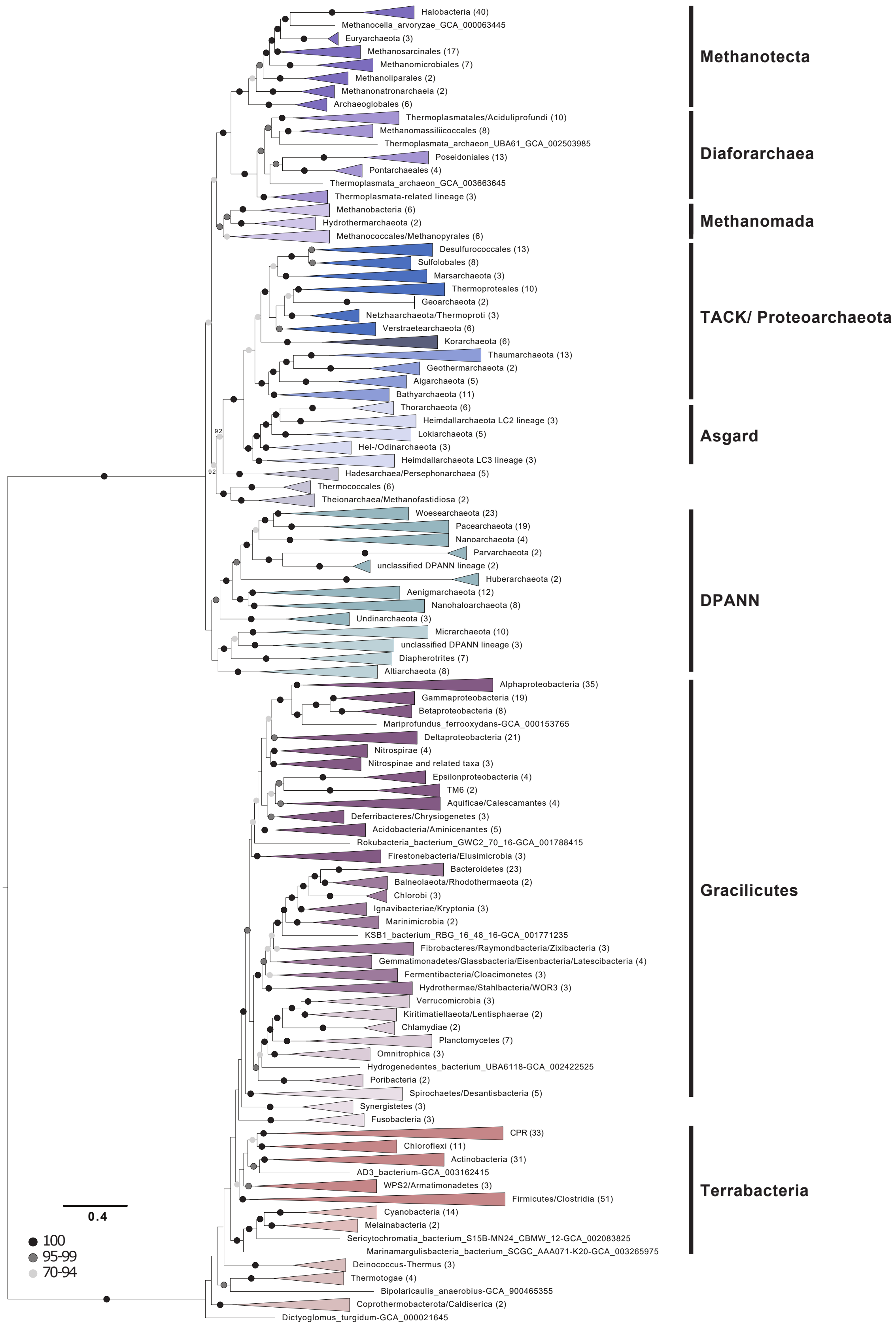
Halobacteria (40)
Methanocella_arvoryzae_GCA_000063445
Euryarchaeota (3)
Methanosarcinales (17)
Methanomicrobiales (7)
Methanoliparales (2)
Methanonatronarchaeia (2)
Archaeoglobales (6)
Thermoplasmatales/Aciduliprofundi (10)
Methanomassiliicoccales (8)
Thermoplasmata_archaeon_UBA61_GCA_002503985
Poseidoniales (13)
Pontarchaeales (4)
Thermoplasmata_archaeon_GCA_003663645
Thermoplasmata-related lineage (3)
Methanobacteria (6)
Hydrothermarchaeota (2)
Methanococcales/Methanopyrales (6)
Desulfurococcales (13)
Sulfolobales (8)
Marsarchaeota (3)
Thermoproteales (10)
Geoarchaeota (2)
Netzhaarchaeota/Thermoproti (3)
Verstraetearchaeota (6)
Korarchaeota (6)
Thaumarchaeota (13)
Geothermarchaeota (2)
Aigarchaeota (5)
Bathyarchaeota (11)
Thorarchaeota (6)
Heimdallarchaeota LC2 lineage (3)
Lokiarchaeota (5)
Hel-/Odinarchaeota (3)
Heimdallarchaeota LC3 lineage (3)
Hadesarchaea/Persephonarchaea (5)
Thermococcales (6)
Theionarchaea/Methanofastidiosa (2)
Woesearchaeota (23)
Pacearchaeota (19)
Nanoarchaeota (4)
Parvarchaeota (2)
unclassified DPANN lineage (2)
Huberarchaeota (2)
Aenigmarchaeota (12)
Nanohaloarchaeota (8)
Undinarchaeota (3)
Micrarchaeota (10)
unclassified DPANN lineage (3)
Diapherotrites (7)
Altiarchaeota (8)
Alphaproteobacteria (35)
Gammaproteobacteria (19)
Betaproteobacteria (8)
Mariprofundus_ferrooxydans-GCA_000153765
Deltaproteobacteria (21)
Nitrospirae (4)
Nitrospinae and related taxa (3)
Epsilonproteobacteria (4)
TM6 (2)
Aquificae/Calescamantes (4)
Deferribacteres/Chrysiogenetes (3)
Acidobacteria/Aminicenantes (5)
Rokubacteria_bacterium_GWC2_70_16-GCA_001788415
Firestonebacteria/Elusimicrobia (3)
Bacteroidetes (23)
Balneolaeota/Rhodothermaeota (2)
Chlorobi (3)
Ignavibacteriae/Kryptonia (3)
Marinimicrobia (2)
KSB1_bacterium_RBG_16_48_16-GCA_001771235
Fibrobacteres/Raymondbacteria/Zixibacteria (3)
Gemmatimonadetes/Glassbacteria/Eisenbacteria/Latescibacteria (4)
Fermentibacteria/Cloacimonetes (3)
Hydrothermae/Stahlbacteria/WOR3 (3)
Verrucomicrobia (3)
Kiritimatiellaeota/Lentisphaerae (2)
Chlamydiae (2)
Planctomycetes (7)
Omnitrophica (3)
Hydrogenedentes_bacterium_UBA6118-GCA_002422525
Poribacteria (2)
Spirochaetes/Desantisbacteria (5)
Synergistetes (3)
Fusobacteria (3)
CPR (33)
Chloroflexi (11)
Actinobacteria (31)
AD3_bacterium-GCA_003162415
WPS2/Armatimonadetes (3)
Firmicutes/Clostridia (51)
Cyanobacteria (14)
Melainabacteria (2)
Sericytochromatia_bacterium_S15B-MN24_CBMW_12-GCA_002083825
Marinamargulisbacteria_bacterium_SCGC_AAA071-K20-GCA_003265975
Deinococcus-Thermus (3)
Thermotogae (4)
Bipolaricaulis_anaerobius-GCA_900465355
Coprothermobacterota/Caldiserica (2)
Dictyoglomus_turgidum-GCA_000021645
92
92
Methanotecta
Diaforarchaea
Methanomada
TACK/ Proteoarchaeota
Asgard
DPANN
Gracilicutes
Terrabacteria
0.4
● 100
● 95-99
● 70-94

Figure 27 – A phylogeny of Archaea and Bacteria inferred from a concatenation of 27 marker genes. Continued on next page.

Figure 27 – Continued. Consistent with some recent studies[53,164,181,52], we recovered the DPANN, TACK and Asgard Archaea as monophyletic groups. Although the deep branches within Bacteria are poorly resolved, we recovered a sister group relationship between CPR and Chloroflexota, consistent with a recent report[24]. The tree was inferred using the best-fitting LG+C60+G4+F model in IQ-TREE 2[72]. Branch lengths are proportional to the expected number of substitutions per site. Support values are ultrafast (UFBoot2) bootstraps[201]. Numbers in parenthesis refer to the number of taxa within each collapsed clade. Please note that collapsed taxa in the Archaea and Bacteria roughly correspond to order- and phylum-level lineages, respectively. Adapted from[176].

comprising Theionarchaea, Methanofastidiosa and Thermococcales. However, the position of these lineages was sensitive to the marker gene set used. As part of a robustness test, we also inferred an additional tree from a 25-gene subset, excluding two genes that have complex evolutionary histories in Archaea[222] (Figure 28).

In this analysis, these Archaea instead branched with Methanomada with high support (98%), highlighting the difficulty of placing these lineages in the archaeal tree. Euryarchaeotal paraphyly has been previously reported[233,181,52], though the extent of euryarchaeotal paraphyly and the lineages involved has varied among analyses.

A basal placement of DPANN within Archaea is sometimes viewed with suspicion[234] because DPANN genomes are reduced and appear to be fast-evolving, properties that may cause LBA artefacts[235] when analyses include Bacteria. However, in contrast to CPR, with which DPANN share certain ecological and genomic similarities (e.g. host dependency, small genomes, limited metabolic potential), the early divergence of DPANN from the archaeal branch has received support from a number of recent studies[236,237,53,238,52,39] though the inclusion of certain lineages within this radiation remains controversial[234,51].

While more in-depth analyses will be needed to further illuminate the evolutionary history of DPANN and establish which archaeal clades constitute this lineage, our work is in agreement with current literature and a recently established phylogeny-informed archaeal taxonomy[238].

A broader observation from our analysis is that the phylogenetic diversity of the archaeal and bacterial domains, measured as substitutions per site in this consensus set of vertically-evolving marker genes, appears to be similar (Figure 24A; the mean root to tip distance for archaea: 2.38, for bacteria: 2.41, the range of root to tip distances for archaea: 1.79-3.01, for bacteria: 1.70-3.17). Considering only the slowest evolving category of sites, branch lengths within Archaea are actually longer than within Bacteria (Figure 24C). This result differs from some published trees[27,30] in which the phylogenetic diversity of Bacteria has appeared to be significantly greater than that of Archaea.

By contrast to those earlier studies, we analysed a set of 350 genomes from each domain, an approach which may tend to reduce the differences between them. While we had to significantly downsample the sequenced diversity of Bacteria, our sampling nonetheless included representatives from all known major lineages of both domains, and so might be expected to recover a difference in diversity, if present. Our analyses and a number of previous studies[27,26,168,30] indicate that the choice of marker genes has a profound impact on the apparent phylogenetic diversity of certain prokaryotic groups; for instance, in the proportion of bacterial diversity composed of CPR[27,28].

Our results demonstrate that slow and fast-evolving sites from the same set of marker genes support different tree shapes and branch lengths; it therefore seems possible that between-dataset differences are due, at least in part, to evolutionary rate variation within and between marker genes.

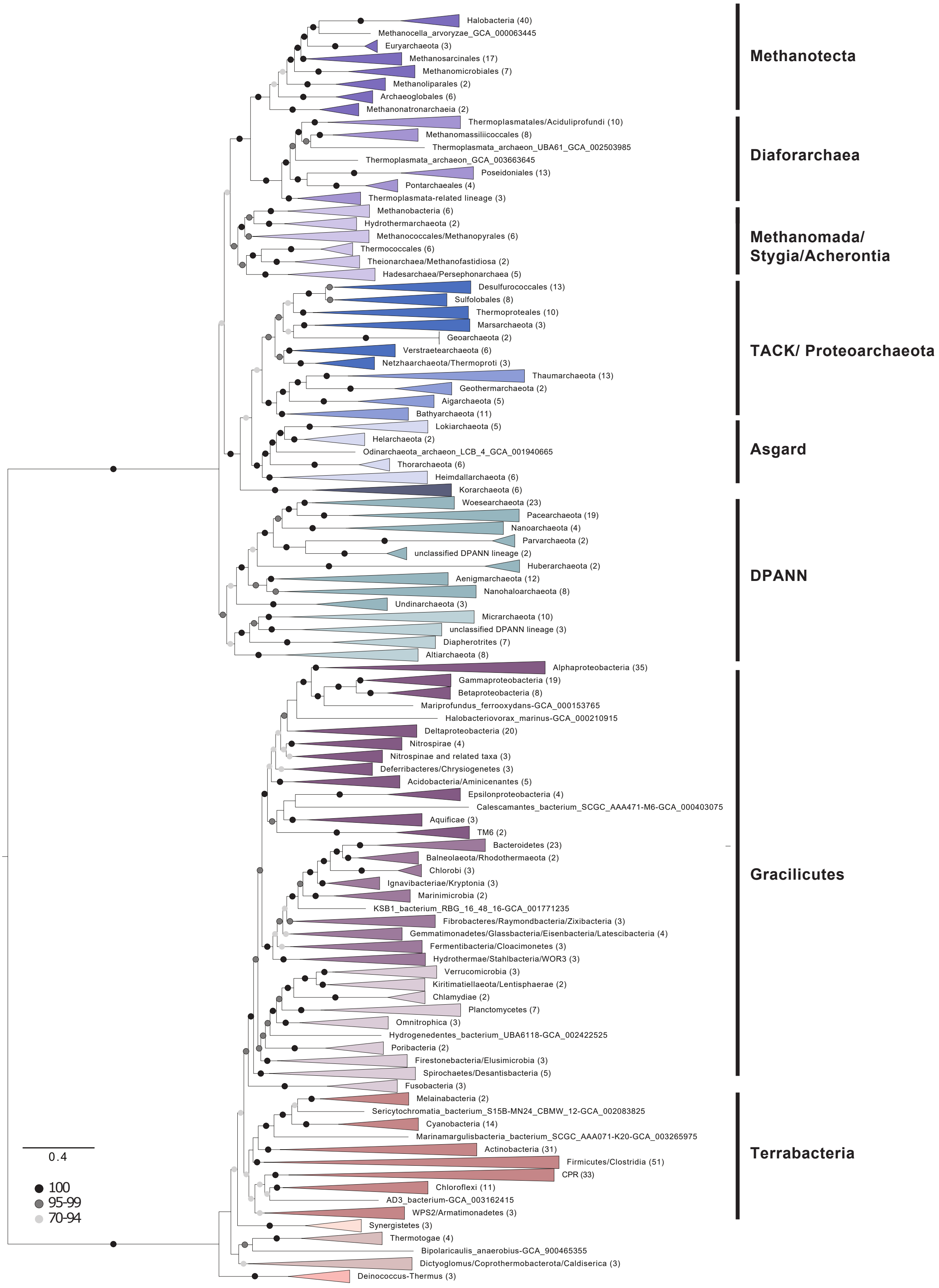
Halobacteria (40)
Methanocella_arvoryzae_GCA_000063445
Euryarchaeota (3)
Methanosarcinales (17)
Methanomicrobiales (7)
Methanoliparales (2)
Archaeoglobales (6)
Methanonatronarchaeia (2)
Thermoplasmatales/Aciduliprofundi (10)
Methanomassiliicoccales (8)
Thermoplasmata_archaeon_UBA61_GCA_002503985
Thermoplasmata_archaeon_GCA_003663645
Poseidoniales (13)
Pontarchaeales (4)
Thermoplasmata-related lineage (3)
Methanobacteria (6)
Hydrothermarchaeota (2)
Methanococcales/Methanopyrales (6)
Thermococcales (6)
Theionarchaea/Methanofastidiosa (2)
Hadesarchaea/Persephonarchaea (5)
Desulfurococcales (13)
Sulfolobales (8)
Thermoproteales (10)
Marsarchaeota (3)
Geoarchaeota (2)
Verstraetearchaeota (6)
Netzhaarchaeota/Thermoproti (3)
Thaumarchaeota (13)
Geothermarchaeota (2)
Aigarchaeota (5)
Bathyarchaeota (11)
Lokiarchaeota (5)
Helarchaeota (2)
Odinarchaeota_archaeon_LCB_4_GCA_001940665
Thorarchaeota (6)
Heimdallarchaeota (6)
Korarchaeota (6)
Woesearchaeota (23)
Pacearchaeota (19)
Nanoarchaeota (4)
Parvarchaeota (2)
unclassified DPANN lineage (2)
Huberarchaeota (2)
Aenigmarchaeota (12)
Nanohaloarchaeota (8)
Undinarchaeota (3)
Micrarchaeota (10)
unclassified DPANN lineage (3)
Diapherotrites (7)
Altiarchaeota (8)
Alphaproteobacteria (35)
Gammaproteobacteria (19)
Betaproteobacteria (8)
Mariprofundus_ferrooxydans-GCA_000153765
Halobacteriovorax_marinus-GCA_000210915
Deltaproteobacteria (20)
Nitrospirae (4)
Nitrospinae and related taxa (3)
Deferribacteres/Chrysiogenetes (3)
Acidobacteria/Aminicenantes (5)
Epsilonproteobacteria (4)
Calescamantes_bacterium_SCGC_AAA471-M6-GCA_000403075
Aquificae (3)
TM6 (2)
Bacteroidetes (23)
Balneolaeota/Rhodothermaeota (2)
Chlorobi (3)
Ignavibacteriae/Kryptonia (3)
Marinimicrobia (2)
KSB1_bacterium_RBG_16_48_16-GCA_001771235
Fibrobacteres/Raymondbacteria/Zixibacteria (3)
Gemmatimonadetes/Glassbacteria/Eisenbacteria/Latescibacteria (4)
Fermentibacteria/Cloacimonetes (3)
Hydrothermae/Stahlbacteria/WOR3 (3)
Verrucomicrobia (3)
Kiritimatiellaeota/Lentisphaerae (2)
Chlamydiae (2)
Planctomycetes (7)
Omnitrophica (3)
Hydrogenedentes_bacterium_UBA6118-GCA_002422525
Poribacteria (2)
Firestonebacteria/Elusimicrobia (3)
Spirochaetes/Desantisbacteria (5)
Fusobacteria (3)
Melainabacteria (2)
Sericytochromatia_bacterium_S15B-MN24_CBMW_12-GCA_002083825
Cyanobacteria (14)
Marinamargulisbacteria_bacterium_SCGC_AAA071-K20-GCA_003265975
Actinobacteria (31)
Firmicutes/Clostridia (51)
CPR (33)
Chloroflexi (11)
AD3_bacterium-GCA_003162415
WPS2/Armatimonadetes (3)
Synergistetes (3)
Thermotogae (4)
Bipolaricaulis_anaerobius-GCA_900465355
Dictyoglomus/Coprothermobacterota/Caldiserica (3)
Deinococcus-Thermus (3)
Methanotecta
Diaforarchaea
Methanomada/
Stygia/Acherontia
TACK/ Proteoarchaeota
Asgard
DPANN
Gracilicutes
Terrabacteria
0.4
● 100
● 95-99
● 70-94

Figure 28 – A phylogeny of Archaea and Bacteria inferred from a concatenation of 25 marker genes. Continued on next page.

Figure 28 – Continued. In addition to our focal analysis, we inferred a tree from a 25-gene subset of the 27 highest-ranked marker genes. This analysis excluded two genes (COG0480: Translation elongation factor EF-G and COG5257: Translation initiation factor 2, gamma subunit) that, while scoring well by split score, had either low representation in Bacteria or had previously been suggested to have a complex evolutionary history in Archaea[222]. The tree topologies and AB branch lengths (25-marker supermatrix: 2.3738 substitutions/site, 27-marker supermatrix: 2.5178 substitutions/site) were closely similar between the 27- and 25-gene analyses, with the exception of several lineages that proved difficult to place. In the 25-gene analysis, the Korarchaeota branched outside the TACK+Asgard clade with moderate (92%) bootstrap support, and the Hadesarchaea, Persephonarchaea, Theionarchaea, Methanofastidiosa and Thermococcales formed a clade sister to Methanomada (98% bootstrap support). The tree was inferred using the best-fitting LG+C60+G4+F model in IQ-TREE 2[72]. Branch lengths are proportional to the expected number of substitutions per site. Support values are ultrafast (UFBoot2) bootstraps[201]. Numbers in parenthesis refer to the number of taxa within each collapsed clade. Please note that collapsed taxa in the Archaea and Bacteria roughly correspond to order- and phylum-level lineages, respectively. Adapted from Moody et al.[176].

Difficulties in estimating the age of the last universal common ancestor

While a consensus may be emerging on the topology of the universal tree, estimates of the ages of the deepest branches, and their length in geological time remain highly uncertain. The fossil record of early life is incomplete and difficult to interpret[239], and in this context molecular clock methods provide a means of combining the abundant genetic data available for modern organisms with the limited fossil record to improve our understanding of early evolution[216].

The 381 gene dataset was suggested to be useful[30] for inferring deep divergence times, because age estimates of the last universal common ancestor (LUCA) from this dataset using a strict molecular clock were in agreement with the geological record: a root (LUCA) age of 3.6-4.2 Ga was inferred from the entire 381 gene dataset, consistent with the earliest fossil evidence for life[216,215]. By contrast, analysis of ribosomal markers alone[30] supported a root age of ∼7 Ga, which might be considered implausible because it is older than the age of the Earth and Solar System (with the moon-forming impact occurring ∼4.52 Ga[240,241]).

The published molecular clock analyses[30] made use of concatenation-based branch lengths in which topological disagreement among sites is not modelled, and are likely to be affected by the impact of nonvertical marker genes and substitutional saturation on branch length estimation discussed above. Consistent with this hypothesis, divergence time inference using the same method on the 5% most-vertical subset (Table 1) of the expanded marker set (as determined by ΔLL; this set of 20 genes includes only one ribosomal protein, see Table 1), resulted in age estimates for LUCA that exceed the age of the Earth, $>\sim$ 5.5 Ga (Figure 5), approaching the age inferred from the ribosomal genes (7.46-8.03 Ga).

These results (Figure 5) suggest that the apparent agreement between the fossil record and divergence times estimated from the expanded gene set may be due, at least in part, to the shortening of the AB branch due to phylogenetic incongruence among marker genes.

In the original analyses, the age of LUCA was estimated using a strict clock with a single calibration constraining the split between Cyanobacteria and Melainabacteria derived from estimates of the Great Oxidation Event and a secondary estimate of the age of cyanobacteria derived from an independent analysis[244]. The combination

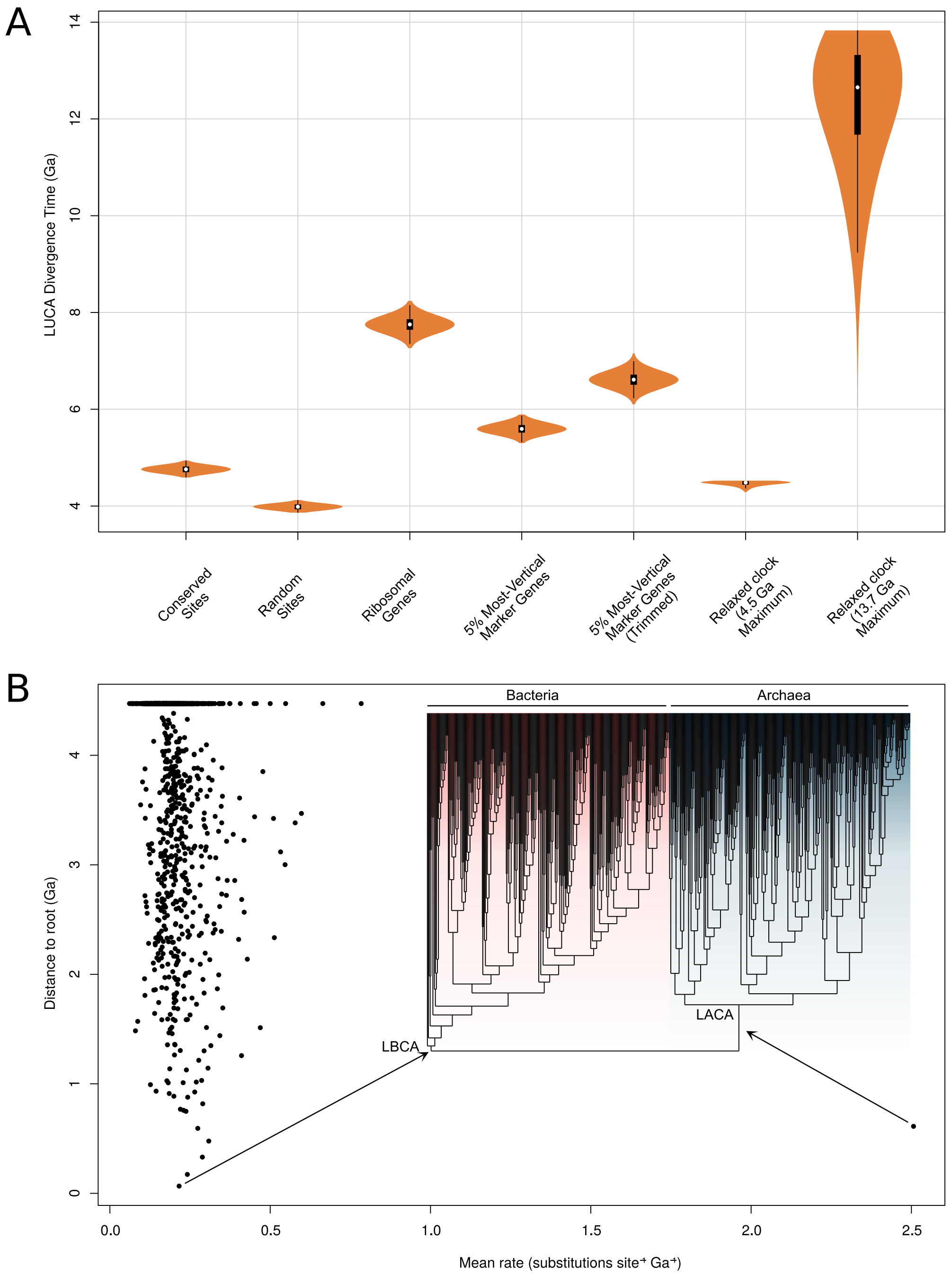

Figure 29 – Molecular clock estimates of LUCA and LACA age are uncertain due to a lack of deep calibrations and maximum ages for microbial clades. Continued on next page.

Figure 29 – Continued. (A) Posterior node age estimates from Bayesian molecular clock analyses of 1) Conserved sites as estimated previously[30]; 2) Random sites[30] 3) Ribosomal genes[30] 4) The top 5% of marker gene families according to their ΔLL score (including only 1 ribosomal protein) and 5) The same top 5% of marker genes trimmed using BMGE[212] to remove poorly-aligned sites. In each case, a strict molecular clock was applied, with the age of the Cyanobacteria-Melainabacteria split constrained between 2.5 and 2.6 Ga. In 6) and 7) an expanded set of fossil calibrations were implemented with a relaxed (lognormal) molecular clock. In 6) a soft maximum age of 4.520 Ga was applied, representing the age of the moon-forming impact[242]. In 7) a soft maximum age corresponding to the estimated age of the universe[243] was applied. (B) Inferred rates of molecular evolution along the phylogeny in a relaxed clock analysis where the maximum age was set to 4.520 Ga. The rate of evolution along the archaea stem lineage was a clear outlier (mean = 2.51, 95% HPD = 1.6-3.5 subs. site-1 Ga-1).

of a strict clock and only two calibrations is not sufficient to capture the variation in evolutionary rate over deep timescales[245].

To investigate whether additional calibrations might help to improve age estimates for deep nodes in the universal tree, we performed analyses on our new 27 marker gene dataset using two different relaxed clock models (with branchwise independent and autocorrelated rates) and 7 additional calibrations (Table 5). Unfortunately, all of these were minimum age calibrations with the exception of the root (for which the moon-forming impact 4.52Ga[242] provides a reasonable maximum), due to the difficulty of establishing uncontroversial maximum ages for microbial clades. Maximum age constraints are essential to inform faster rates of evolution because, in combination with more abundant minimum age constraints, they imply that a given number of substitutions must have accumulated in, at most, a certain interval of time. In the absence of other maximum age constraints, the only lower bound on the rate of molecular evolution is provided by the maximum age constraint on the root (LUCA).

These new analyses indicated that even with additional minimum age calibrations, the age of LUCA inferred from the 27-gene dataset was unrealistically old, falling close to the maximum age constraint in all analyses even when the maximum was set to the age of the known universe (13.7Ga[243]; Figure 27). Inspection of the inferred rates of molecular evolution across the tree (Figure 27B) provides some insight into these results: the mean rate is low (mean = 0.21, 95% credibility interval = 0.19-0.22 subs. site-1 Ga-1), so that long branches (such as the AB stem), in the absence of other information, are interpreted as evidence of a long period of geological time. These low rates likely result both from the limited number of calibrations and, in particular, the lack of maximum age constraints.

An interesting outlier among inferred rates is the LUCA to LACA branch, which has a rate tenfold greater than the average (mean = 2.51, 95% HPD = 1.6-3.5 substitutions per site^{-1} Ga $^{-1}$). The reason is that calibrations within Bacteria imply that LBCA cannot be younger than 3.227 Ga (Manzimnyama Banded Ironstone Formation provides evidence of cyanobacterial oxygenation[246], Supplementary Information Table S6, see Appendix); as a result, with a 4.52Ga maximum the LUCA to LBCA branch cannot be longer than 1.28Ga. By contrast, the early branches of the archaeal tree are poorly constrained by fossil evidence. Analysis without the 3.227Ga constraint resulted in overlapping age estimates for LBCA (4.47-3.53Ga) and LACA (4.37-3.44Ga). Finally, analysis of the archaeal and bacterial subtrees independently (that is, without the AB branch, rooted on LACA and LBCA, respectively) resulted in LBCA and LACA ages that abut the maximum root age (LBCA: 4.52-4.38Ga; LACA: 4.52-4.14Ga). This analysis demonstrates that, under these analysis conditions, the inferred age of the root (whether corresponding to LUCA, LACA, or LBCA) is strongly influenced by the prior assumptions about the maximum age of the root.

In sum, the agreement between fossils and age estimates from the expanded gene set appears to result from the impact of phylogenetic incongruence on branch length estimates. Under more flexible modelling assumptions the limitations of current clock methods for estimating the age of LUCA become manifest: the sequence data only contain limited information about the age of the root, with posterior estimates driven by the prior assumptions about the maximum age of the root. This analysis implies several possible ways to improve age estimates of deep branches in future analyses. More calibrations, particularly maximum age constraints and calibrations within Archaea, are essential to refine the current estimates. Given the difficulties in

Node	Fossil	Minimum	Soft Maximum
LUCA	Strelley Pool Formation, Western Australia	3347 Ma	4520 Ma
Total group Cyanobacteria	Manzimnyama Banded Ironstone Formation, South Africa	3225 Ma	4520 Ma
Crown group Cyanobacteria	*Bangiomorpha pubescens*	1033 Ma	4520 Ma
Heterocystous Cyanobacteria	*Anhuthruix magna*	720 Ma	4520 Ma
Alphaproteobacteria	*Bangiomorpha pubescens*	1033 Ma	4520 Ma
Anoxychlamydiales	*Bangiomorpha pubescens*	1033 Ma	4520 Ma
Lokiarchaeota	Changzhougou Formation, Northern China	1619?	4520 Ma

Table 5 – A list of fossil calibrations employed in relaxed molecular clock analyses. All calibrations were modelled as uniform distributions between a hard minimum and a soft maximum. The probability that the maximum could be exceeded was modelled as a 2.5% probability tail. With the exception of Heterocystous Cyanobacteria and Anoxychlamydiales[176], the calibrations were taken from Betts et al.[216].

establishing maximum ages for archaeal and bacterial clades, constraints from other sources such as donor-recipient age constraints inferred from HGTs[247,248,249,250], or clock models that capture biological opinion about rate shifts in early evolution, may be particularly valuable.

3.5 Conclusion

We established a set of 27 highly vertically evolving marker gene families and found no evidence that ribosomal genes overestimate stem length; since they appear to be transferred less frequently than other genes, our analysis affirms that ribosomal proteins are useful markers for deep phylogeny and that, in general, they are not intrinsically worse or better than other functional classifications of proteins.

We show that the inclusion of HGTs and paralogous sequences artificially reduces the inferred stem length, and as such manual curation is a fundamental stage in the selection of universal marker gene families. We also find that substititional saturation may also artificially reduce the archaeal-bacterial stem, and model selection accounting for compositional and rate heterogeneity are equally important.

Our divergence time estimates highlight the limit of the molecular clock when a limited number of fossil calibrations are available, but also show that using sets of markers which artificially reduce the inferred stem length is not a solution to the old molecular clock estimates for early life.

In general, high-verticality markers, regardless of functional category, supported a longer AB branch length. Furthermore, our phylogeny was consistent with recent work on early prokaryotic evolution, resolving the major clades within Archaea and nesting the CPR within Terrabacteria. Notably, our analyses suggested that both the true Archaea-Bacteria branch length and the phylogenetic diversity of Archaea, may be underestimated by even the best current models, a finding that is consistent with a root of the tree of life between the two prokaryotic domains.

4 OTHER APPLICATIONS OF PHYLOGENETICS

Author's contribution

Data from this chapter form part of three publications[251,252,253]:

Antón, Z., Weijman, J.F., Williams, C., **Moody, E.R.R**, Mantell, J., Yip, Y.Y., Cross, J.A., Williams, T.A., Steiner, R.A., Crump, M.P. and Woolfson, D.N., 2021. Molecular mechanism for kinesin-1 direct membrane recognition. *Science Advances*, 7(31), p.eabg6636.

Ferguson, DCJ., Mokim, JH., Meinders, M., **Moody, E.R.R**., Williams., TA, Cooke., S., Trakarnsanga, K., Daniels, DE., Ferrer-Vicens, I., Shoemark, D,. Tipgomut, C., Macinnes, KA., Wilson, MC., Singleton, BK., Frayne, J., 2021. Characterization and evolutionary origin of novel C2H2 zinc finger protein (ZNF648) required for both erythroid and megakaryocyte differentiation in humans. *Haematologica*, 106(11), p.2859.

Simonetti, B., Guo, Q., Gimenez-Andres, M., Chen, K.E., **Moody, E.R.R**, Evans, A.J., Danson, C.M., Williams, T.A., Collins, B.M. and Cullen, P.J., 2021. Mechanistic basis for SNX27-Retromer coupling to ESCPE-1 in promoting endosomal cargo recycling. *bioRxiv.*

The phylogenetic and evolutionary analyses were undertaken by Edmund R. R. Moody in each of the above papers, as such the methods sections correspond to the methods sections in the above papers (these methods sections were written entirely by Edmund R. R. Moody). Where necessary, a discussion and contextualisation of the research of the other authors is included. This chapter was written by Edmund R. R. Moody in its entirety.

4.1 Abstract

Traditionally phylogenetics has been used to infer the vertical relationships between species. However, other exciting applications of the methods are now being applied. Here, we outline some of these other uses of phylogenetic methods and present three case studies, examining the evolution at the sub-protein level, the protein level, and at the protein complex level. Specifically, the evolution of the kinesin-light-chain

component of the motor protein kinesin-1, the evolution of a novel zinc-finger protein (ZNF648) involved in blood-cell differentiation, and the evolution of the mechanisms behind the protein complex responsible for aspects of protein recycling sorting, retromer and its binding mechanism with sorting-nexin protein (SNX27) and other associated proteins.

4.2 Introduction

Unlike the previous chapters, which seek to use a combination of phylogenetics and phylogenomics in an attempt to further our understanding of the tree of life and early evolutionary history, this chapter uses phylogenetic techniques in a practical sense to examine molecular evolution in terms of protein structure and organisation. Phylogenetic techniques have been used previously to help in the classification of genes and proteins[224], as well as to aid in solving structural protein problems[254]. They can be used for the classification of organisms[255], for evolutionarily defined taxonomies through differentiating between species[256] and subspecies[257] for conservation reasons[258]. In fact phylogenetic diversity[259] is used as a major metric for determining diversity in conservation research[260].

Another exciting recent application of developing phylogenetic methods is in linguistic evolution[261] and wider cultural historical applications, ranging from medicinal plant usage to afterlife beliefs[262,263,264,265,266].

More biologically grounded applications involve estimating species divergence times through the 'Molecular Clock', the idea that in general, evolutionary time is proportional to the number of amino acid substitutions[63]. This has been refined over the years as to incorporate additional biological knowledge, as we understand that a constant rate of evolution over all branches in the tree is not biologically realistic[267], and the use of fossil data for independent time calibrations to convert the 'relative' timescales of substitutions per site to actual time estimates[268] is required for dated time-trees.

Another useful application lies within phylogeographical estimations[269], using a combination of genotypes (whether genetic, genomic or single nucleotide polymorphisms), geographical data, and phylogenetics/phylogenomics to understand how modern distributions of genotypes have arisen, and the historical and geographical processes behind them[270]. Phylogenies can also be used in the court room, one of the most famous examples of this was the 'Florida dentist' case in the 1990s, where a parsimony-based phylogeny was used to prove that, beyond all reasonable doubt, the patients of the dentist had become infected with human immunodeficiency virus (HIV) from said dentist[271]. Similar methods, albeit using probabilistic models, are being employed for other similar cases to date[272,273].

Various websites incorporate SNP-based phylogenies to determine your recent familial ancestry using direct-to-consumer genetic testing[274]. Such data has also been used recently in helping identify genomic regions associated with protection against severe COVID-19 infection[275]. Phylogeny has also played a large part in understanding the COVID-19 pandemic, from its likely origins as a zoonotic event emerging from a market in Wuhan[276] to the evolution and spread of the virus, as well as the impact of the vaccine[277,278]. Phylogenies were also used in combination with biochemical techniques to understand the evolution of the SARS-CoV-1 strain of *Betacoronavirus* and its closest relatives, and inform on the evolution of the structure on both a protein and sub-protein level[279].

The focus of this chapter is to demonstrate phylogenetic applications of a range of evolutionary biochemical questions, specifically at the sub-protein level, the protein level, and at the protein complex level. We examine the evolution of the kinesin-light-chain component in kinesin-1, the evolution of a novel zinc-finger protein (ZNF648), and the evolution of the mechanisms behind sorting nexin protein (SNX27) binding to retromer and its associated proteins.

4.3 Evolution of membrane recognition in Kinesin-1

4.3.1 Introduction

Kinesins are a superfamily of motor proteins responsible for intra-cellular transport of cellular cargo, organelles (such as mitochondria and lysosomes), and also play an important role in mitosis[280]. Kinesins use microtubules as a network to move around the cytoskeleton similar to the way trains move on railway tracks. Amongst the kinesin protein superfamily, kinesin-1 was the first to be identified in the 1980s[281] from the cytoplasm of *Architeuthis dux* (giant squid) nerve cells.

Kinesin-1's structure is made up of two light chains (KLC) and two heavy chains (KHC) (Figure 30). The adenosine-triphosphate (ATP) dependent motor is found within the KHCs, which form two globular 'feet' that attach to the microtubule 'tracks'. The KLCs bind to the cellular cargo, which allow it to be transported. The kinesin-1 complex moves in a step-wise movement, where one 'foot' is propelled forward via a conformational change through hydrolyzing ATP. This foot then binds to the microtubule, followed by the other foot undergoing a similar movement.

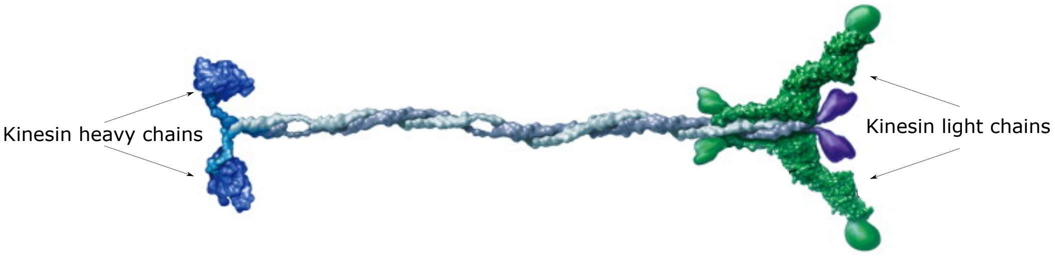

Figure 30 – Structure of kinesin-1, kinesin heavy chains (blue) comprise the motor component of the protein which attach to the microtubules, kinesin light chains (green) comprise the locations which bind to the cargo. Figure adapted from Vale[282].

The KHCs also possess a long tail which coils around the second KHC. At the distal end of the coiled KHC tails are the KLC proteins[283]. KLC1 in *Homo sapiens* is composed of a heptad repeat which binds to the KHC followed by 6 TPR tandem repeats, with an intrinsically disordered C-terminal domain (Figure 31).

Through a collaborative project, the molecular mechanism responsible for mediation between membrane-bound cargo and the KLCs of kinesin-1 was identified[251], and through secondary structure prediction it was suggested that an amphipathic helix (AH) is present near the C-terminus within an otherwise disordered region of the full length KLC1 protein[251]. An amphipathic helix is a protein sequence which possesses both a hydrophobic and hydrophilic face on each side of a helix. In this case, this motif allows for the direct binding of kinesin-1 to phospholipid membranes[251]. Through biochemical synthesis of a peptide including the KLC amino acid sequence, my collaborators[251] found that in the presence of a membrane this α-helix is formed. This was demonstrated through the use of circular dichroism spectroscopy and nuclear magnetic resonance. The AH region was confirmed to bind to lipid membranes by using lysosomes with cosedimentation assays, the inclusion of a helix disrupting proline mutation in the AH region was enough to prevent membrane binding[251].

In *H. sapiens* there are multiple paralogous KLC genes, and across these paralogues alternative splicing appears to be the control mechanism for the inclusion of the amphipathic helix in the KLC protein. The longer splicing variant possesses the

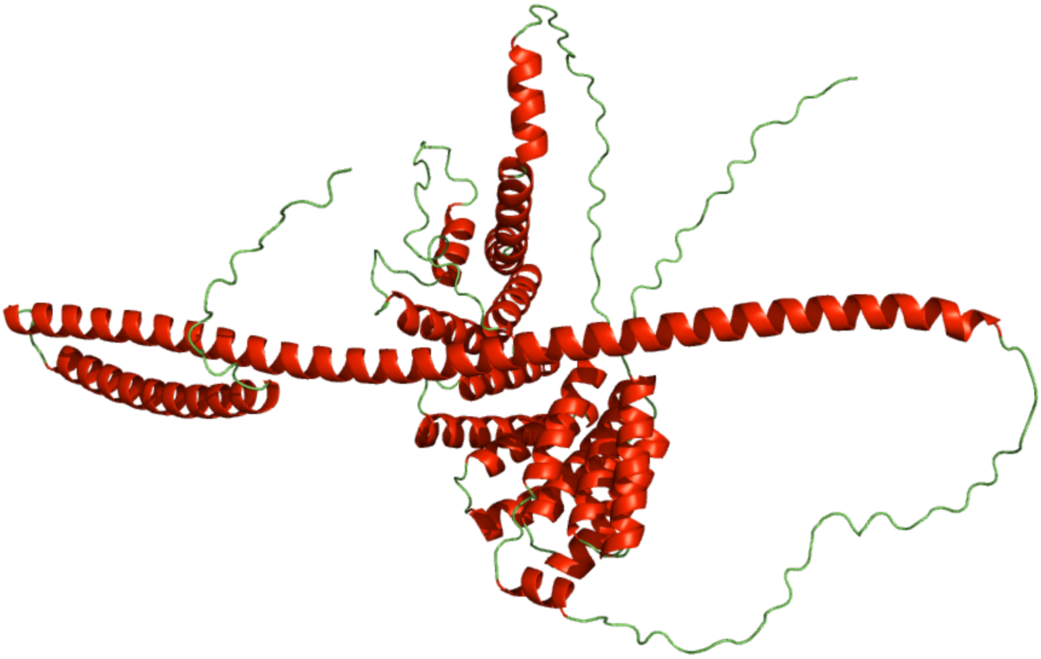

Figure 31 – Alphafold structural prediction for KLC1[284,285], produced with PyMOL[286]. C-terminal on the right, the AH region is believed to be present in this 'disordered' region. Multiple TPR repeats present nearer the in the middle of protein. α-helices in red, disordered region in green.

amphiphathic helix, whereas the shorter spicing variant does not. In this case-study, we examine the evolutionary history of the KLC proteins, and the presence and absence of the the AH motif in KLC isoforms across the lineages where they are found.

4.3.2 Methods

Initially, orthologous KLC1, KLC2, KLC3 and KLC4 sequences were downloaded from ENSEMBL, PMID: 31691826, followed by searching using BLAST to identify potential KLC homologues from an increased range of taxa. MAFFT L-INS-i[220] was used for alignment under default parameters. Maximum likelihood tree inference was done using IQ-TREE 2.0.6[72]. Initial trees were done using LG+F+G, and the final representative tree was inferred using the built-in model finder within IQ-TREE[72] with the Bayesian information criterion (BIC) for model selection. Several rounds of tree inference and were used to identify orthologous and paralogous sequences, in addition to within-species duplication events. The best fitting model for the representative tree using AIC, BIC and AICc was 'JTT + R6'.

4.3.3 Results and Discussion

Initial exploration was done using a hybrid approach, with the multiple sequence alignment of KLC1 metazoan sequences from the ENSEMBL database[287], and using the KLC1 sequence from *H. sapiens* as a query for a BLAST search against non-metazoan opisthokont sequences. After taking the top 100 sequences and including a bacterial tetratricopeptide repeat (TPR) protein as an outgroup, these sequences were aligned and a tree inferred under a fast maximum likelihood model (LG+F+G) (Figure 32).

Although the bacterial outgroup does resolve more closely (when unrooted) to the clade of non-metazoan opisthokont sequences, it does not preclude the possibility of the presence of KLC orthologues in the opisthokont lineage, and as (aside from the outgroup) the only sequences included in this initial search were opisthokonts, then further exploration was necessary to determine when the ancestral KLC gene emerged.

Further BLAST searching revealed annotations of multiple KLC proteins in various metazoan groups. *H. sapiens* possess four paralogues of KLC: KLC1, KLC2,

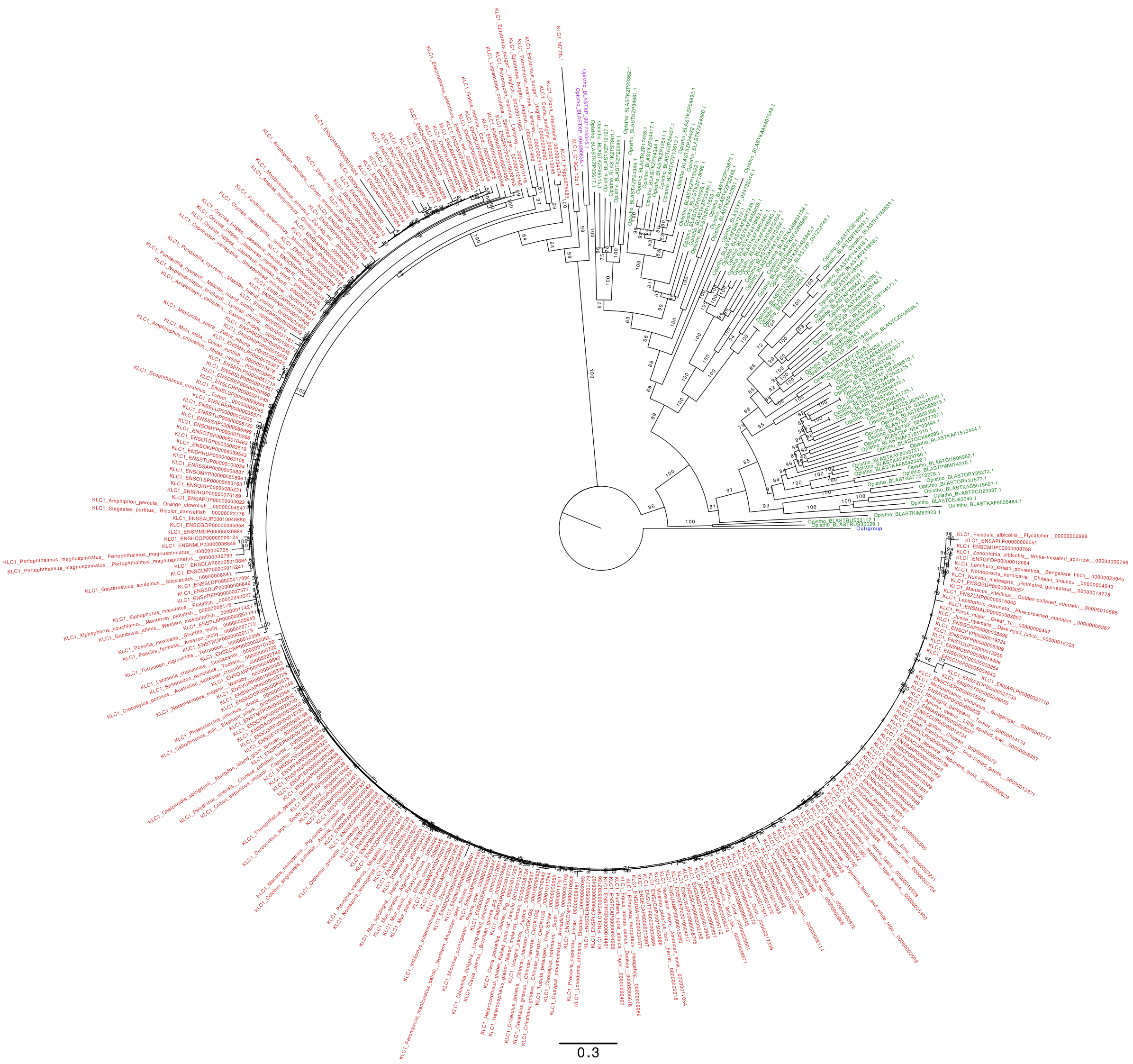

Figure 32 – Preliminary maximum likelihood tree of KLC1. (LG+F+G, 1000 ultrafast bootstrap replicates) tree topology inferred from a combination of top BLAST hits against non-metazoan opisthokonts (green) and the metazoan sequences for KLC1 from ENSEMBL (red), the only non-metazoan sequences to group with Metazoa were those of choanoflagellates (purple). A bacterial TPR protein is included as an outgroup (blue).

KLC3 and KLC4. These four sequences were used as query sequences for the BLAST against representative sequences from across Eukaryota and Archaea, and only unique sequences from the search were taken. Significant hits were only found within opisthokonts, Viridiplantae and members of the Rhizaria, Alveolata, Stramenopila and Asgard archaea which were included and aligned. We inferred a maximum likelihood tree under LG+F+G (Figure 33).

Although there appears to be a small number of fungi grouping with the metazoan KLC sequences in this tree, a reverse BLAST search showed that these sequences were actually a contamination of the genome assemblies: the hits were annotated as being from the rain forest green plant, the Malletwood or Silver Leaf, *Rhodamnia argentea*, but when using that sequence as a query, the top hit is in *Fragariocoptes setiger*, a parasitic mite which causes galls in plants[288]. In this tree non-choanozoan eukaryotes mainly grouped together. The archaeal and bacterial TPR proteins formed a clade with a small number of eukaryotic sequences containing TPR motifs. As fungi and plant sequences were grouping together to the exclusion of the choanozoan sequences. From this we could determine that the orthologous KLC sequences were present only in choanozoans, and the proteins found in the other eukaryotes must be distant paralogues. The KLC query sequences with BLAST failed to find similar sequences in Filasterea and Ichthyosporea. The choanoflagellates, *Salpingoeca rosetta* and *Monosiga brevicollis*, possessed a single-copy orthologue of KLC1. KLC was found in all metazoan phyla, with the four KLC paralogues being found in gnathostomes and placed as distinct monophyletic clades in regard to each other. *Caenorhabditis elegans* and *Petromyzon marinus* (the sea lamprey) also possessed multiple copies of KLC genes, which warranted further investigation (Figure 34).

The ML tree depicted a *C. elegans* (or, potentially, nematode-specific) duplication. This is evidenced by the branch lengths between multiple *C. elegans* tips (annotated as klc-1 and klc-2), although they are monophyletic with the other included Ecdysozoa. *Drosophila melanogaster* is placed as the sister group as is to be expected, which suggests that this duplication was limited to at least nematodes if not specific to *C. elegans*. The duplication of the *P. marinus* sequences was less clear, with low bootstrap support separating the agnathan sequences (including a sequence from the hagfish *Eptatretus burgeri*) from other vertebrates. Whilst it does appear

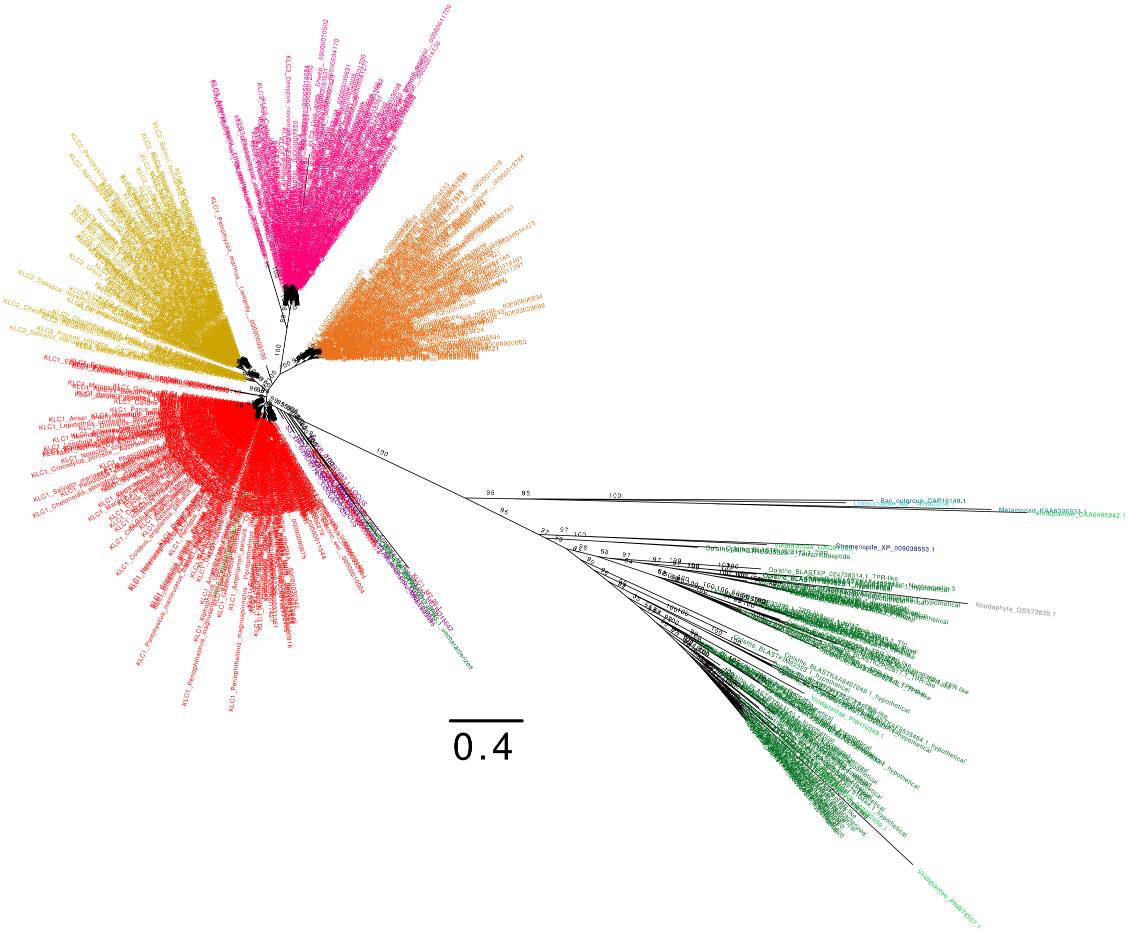

Figure 33 – An unrooted maximum likelihood tree of paralogous KLC genes. (LG+F+G, with 1000 ultrafast bootstrap replicates) phylogeny of the KLC1 (red), KLC2 (yellow), KLC3 (pink), KLC4 (orange). A clade containing sequences from Viridiplantae (light green) and non-metazoan opisthokonts (dark green), sequences from Rhodaphyta (grey), stramenopiles (navy blue) is recovered, with the bacterial and archaeal sequences (light blue) grouping (with a metamonad (grey blue) and an another paralogous plant sequence)

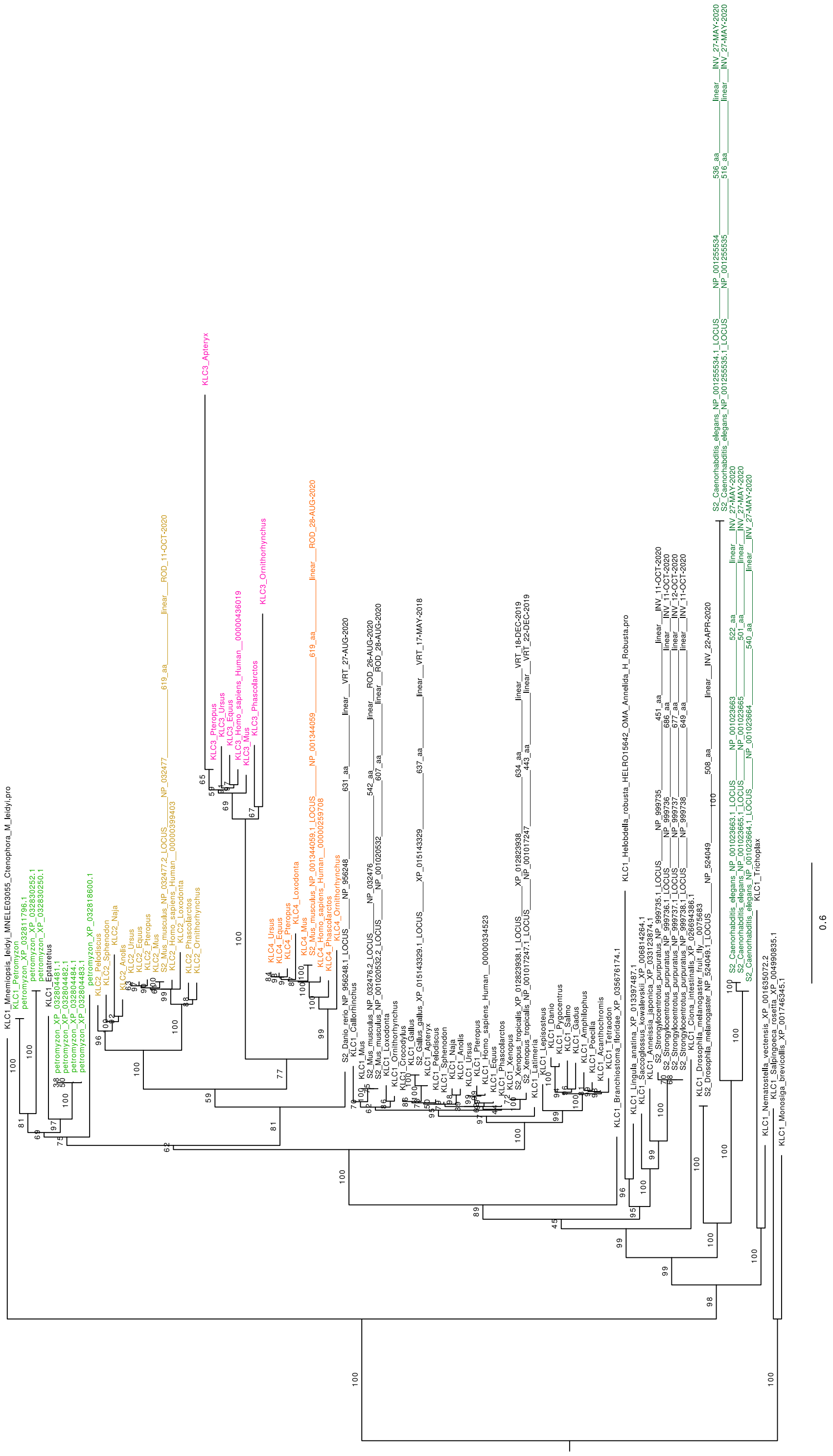
KLC1_Mnemiopsis_leidyi_MNELE03055_Ctenophora_M_leidyi.pro
KLC1_Petromyzon
petromyzon_XP_032811796.1
petromyzon_XP_032830252.1
petromyzon_XP_032830250.1
KLC1_Eptatretus
petromyzon_XP_032804481.1
petromyzon_XP_032804482.1
petromyzon_XP_032804484.1
petromyzon_XP_032804483.1
petromyzon_XP_032818600.1
KLC2_Pelodiscus
KLC2_Sphenodon
KLC2_Naja
KLC2_Anolis
KLC2_Ursus
KLC2_Equus
KLC2_Pteropus
KLC2_Mus
S2_Mus_musculus_NP_032477.2_LOCUS_______NP_032477________________619_aa____________linear___ROD_11-OCT-2020
KLC2_Homo_sapiens_Human__00000399403
KLC2_Loxodonta
KLC2_Phascolarctos
KLC2_Ornithorhynchus
KLC3_Apteryx
KLC3_Pteropus
KLC3_Ursus
KLC3_Equus
KLC3_Homo_sapiens_Human__00000436019
KLC3_Mus
KLC3_Phascolarctos
KLC3_Ornithorhynchus
KLC4_Ursus
KLC4_Equus
KLC4_Pteropus
KLC4_Loxodonta
KLC4_Mus
S2_Mus_musculus_NP_001344059.1_LOCUS_______NP_001344059_____________619_aa____________linear___ROD_28-AUG-2020
KLC4_Homo_sapiens_Human__00000259708
KLC4_Phascolarctos
KLC4_Ornithorhynchus
S2_Danio_rerio_NP_956248.1_LOCUS_______NP_956248________________631_aa____________linear___VRT_27-AUG-2020
KLC1_Callorhinchus
KLC1_Mus
S2_Mus_musculus_NP_032476.2_LOCUS_______NP_032476________________542_aa____________linear___ROD_26-AUG-2020
S2_Mus_musculus_NP_001020532.2_LOCUS_______NP_001020532_____________607_aa____________linear___ROD_28-AUG-2020
KLC1_Loxodonta
KLC1_Ornithorhynchus
KLC1_Crocodylus
KLC1_Gallus
S2_Gallus_gallus_XP_015143329.1_LOCUS_______XP_015143329_____________637_aa____________linear___VRT_17-MAY-2018
KLC1_Apteryx
KLC1_Pelodiscus
KLC1_Sphenodon
KLC1_Naja
KLC1_Anolis
KLC1_Ursus
KLC1_Pteropus
KLC1_Homo_sapiens_Human__00000334523
KLC1_Equus
KLC1_Phascolarctos
KLC1_Xenopus
S2_Xenopus_tropicalis_XP_012823938.1_LOCUS_______XP_012823938_____________634_aa____________linear___VRT_18-DEC-2019
S2_Xenopus_tropicalis_NP_001017247.1_LOCUS_______NP_001017247_____________443_aa____________linear___VRT_22-DEC-2019
KLC1_Latimeria
KLC1_Lepisosteus
KLC1_Danio
KLC1_Pygocentrus
KLC1_Salmo
KLC1_Gadus
KLC1_Amphilophus
KLC1_Poecilia
KLC1_Acanthochromis
KLC1_Tetraodon
KLC1_Branchiostoma_floridae_XP_035676174.1
KLC1_Helobdella_robusta_HELRO15642_OMA_Annelida_H_Robusta.pro
KLC1_Lingula_anatina_XP_013397487.1
KLC1_Saccoglossus_kowalevskii_XP_006814264.1
KLC1_Anneissia_japonica_XP_033123874.1
S2_Strongylocentrotus_purpuratus_NP_999735.1_LOCUS_______NP_999735________________451_aa____________linear___INV_11-OCT-2020
S2_Strongylocentrotus_purpuratus_NP_999736.1_LOCUS_______NP_999736________________686_aa____________linear___INV_11-OCT-2020
S2_Strongylocentrotus_purpuratus_NP_999737.1_LOCUS_______NP_999737________________677_aa____________linear___INV_12-OCT-2020
S2_Strongylocentrotus_purpuratus_NP_999738.1_LOCUS_______NP_999738________________649_aa____________linear___INV_11-OCT-2020
KLC1_Ciona_intestinalis_XP_026694386.1
KLC1_Drosophila_melanogaster_fruit_fly__0075683
S2_Drosophila_melanogaster_NP_524049.1_LOCUS_______NP_524049________________508_aa____________linear___INV_22-APR-2020
S2_Caenorhabditis_elegans_NP_001255534.1_LOCUS_______NP_001255534_____________536_aa____________linear___INV_27-MAY-2020
S2_Caenorhabditis_elegans_NP_001255535.1_LOCUS_______NP_001255535_____________516_aa____________linear___INV_27-MAY-2020
S2_Caenorhabditis_elegans_NP_001023663.1_LOCUS_______NP_001023663_____________522_aa____________linear___INV_27-MAY-2020
S2_Caenorhabditis_elegans_NP_001023665.1_LOCUS_______NP_001023665_____________501_aa____________linear___INV_27-MAY-2020
S2_Caenorhabditis_elegans_NP_001023664.1_LOCUS_______NP_001023664_____________540_aa____________linear___INV_27-MAY-2020
KLC1_Trichoplax
KLC1_Nematostella_vectensis_XP_001635072.2
KLC1_Salpingoeca_rosetta_XP_004990835.1
KLC1_Monosiga_brevicollis_XP_001746345.1
0.6

Figure 34 – Rooted maximum likelihood tree of paralogous KLC genes. (LG+F+G, 1000 ultrafast bootstrap replicates) KLC2 sequences (yellow), KLC3 (pink), KLC4 (orange), *P. marinus* paralogues (light green), *Caenorhabditis elegans* (dark green), other KLC sequences are KLC1 (black).

that there are multiple paralogous KLC sequences within *Petromyzon marinus*, it is not clear whether this was the same duplication event which lead to the KLC duplication in the common ancestor of gnathostomes (Figure 33), or if this was a separate duplication event entirely. Well supported monophyletic clades of KLCs 1-4 are present across extant gnathostomes: mammals, reptiles, amphibians, actinopterygians, and Chondrichthyes. This suggests that the four KLC paralogues were present in the last common ancestor of gnathostomes. A more refined selection of taxa from across Choanozoa was chosen for secondary structure analysis to determine the extent of the AH motif. The inferred phylogeny (using model fitting) from this representative alignment (Figure 35) is consistent with the earlier phylogenies showing the emergence of the multiple KLC paralogues is present by at least the last common ancestor of gnathostomes. Secondary structure analysis shows that the 20 residues long AH motif (located in the c-terminal domain) is present across choanozoans and unless there were multiple independent losses of the AH motif across Ecdysozoa (within at least Arthropoda and Nematoda) then the last common ancestor of ecdysozoans had undergone a secondary loss of the AH motif. Interestingly, the ancestral agnathan likely had the ability to use alternative-splicing as a means of synthesizing a KLC isoform in a long and short-form, with the short form lacking the AH motif. Other planulozoans display a range of isoforms, but only the ecdysozoan sequences appear to have lost sites homologous to the AH motif in even their longest of isoforms. In gnathostomes, the KLC1, KLC2, KLC3 and KLC4 all possess an AH motif, with the hydrophobic residues generally more conserved[251].

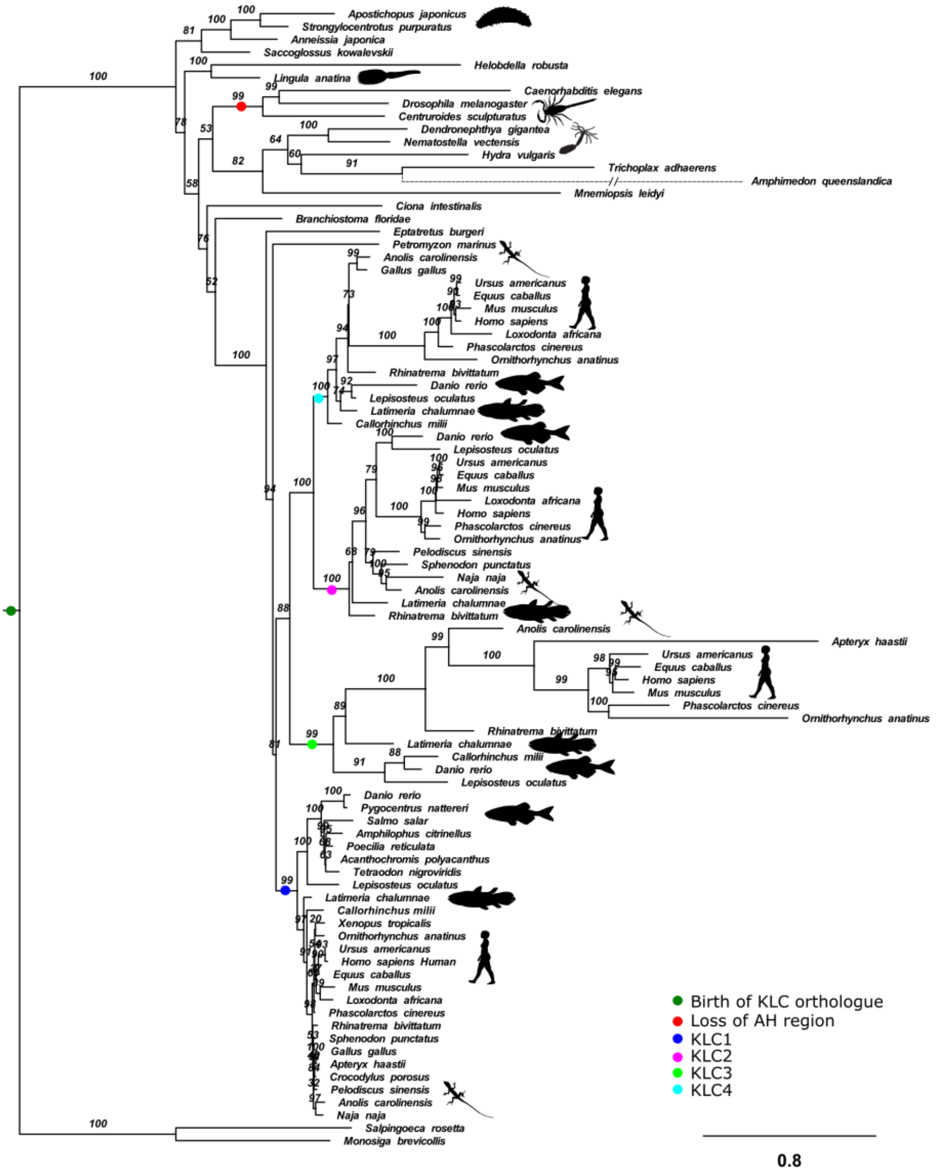

Figure 35 – Representative tree for KLC1. Maximum likelihood phylogeny under the best fitting model (JTT + R6, with 1000 ultrafast bootstrap replicates). The emergence of the four KLC paralogues can be seen in the vertebrate stem. The loss of the AH region in even the longest isoforms of KLC is shown by a red dot on the ecydsozoan stem.

4.3.4 Conclusion

The phylogenetic analyses in combination with the broader work of the paper[251] establishes the amphipathic helix‘s role in mediating membrane binding, and the ancestral presence of the helix in the last common ancestor of Choanozoa. The four KLC paralogues, established here as being present in the last vertebrate common ancestor also play an important role in regulating membrane binding to kinesin-1. The different sequence composition of the KLC paralogues may allow vertebrates more control and sensitivity for certain proteins[251]. The long and short isoforms of KLC1 across Metazoa may also play a role in additional sensitivity, and an important consideration for the future, is to ask why the ecdysozoan lineage has lost the ability to produce a long-isoform containing the amphipathic helix. Kinesin-1's presence in choanoflagellates also warrants further investigation.

4.4 Evolution of the ZNF648 protein family

4.4.1 Introduction

The zinc finger is a structural motif found in proteins which hold zinc ions in place as part of a co-ordination complex (similar to the haem group coordinating with iron ions in haemoglobin). Zinc finger domains will generally have multiple zinc finger motifs, and zinc finger proteins are generally involved in cell development and differentiation as well as the suppression of tumors[289]. There are multiple classes of zinc fingers including the gag knuckle, Treble clef and Zinc ribbon, which vary in their composition and their folding.

The first zinc finger to be discovered was from the popular model organism *Xenopus laevis*, the African clawed frog, where it was found that a protein transcription factor IIIA (TFIIIA) was required for the transcription of 5S RNA[290]. Within TFIIIA a biochemical study showed the presence of significant zinc content and a repeating pattern within the sequence, characterized by two cysteine and two histidine residues[291]. These multiple repeating motifs were theorized to fold into 'finger-like' shapes, which hold the DNA in place[291,290]. Subsequent analysis later confirmed the structure of the motif which is made up of anti-parallel β-sheets and an α-helix which coordinate with the zinc ion in the middle. This provides the α-helix a surface area in which to bind within the major groove of DNA, the amino acid

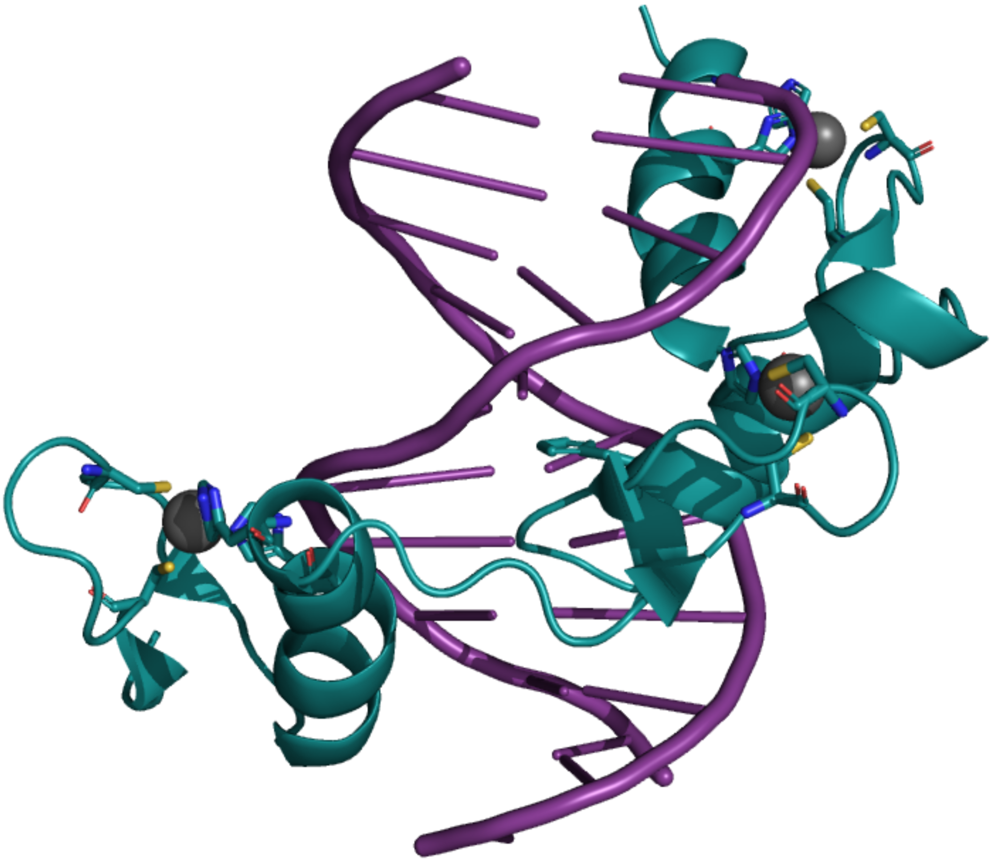

Figure 36 – Structure of a protein (teal) containing three C_2H_2 motifs, binding to the major groove of DNA (purple), cysteine and histidine residues forming the C_2H_2 coloured according to their elements, coordinating with zinc ions (grey). Protein structure determined by Elrod-Erickson et al.[292] (PDB: 1A1L), figure generated with PyMol[286]

residues on the exposed loops of the α-helix giving specificity to the nucleotide to which they bind (Figure 36).

These motifs are referred to as C_2H_2 motifs and there are hundreds of genes containing these motifs across metazoan genomes. C_2H_2 zinc finger proteins represent a large group of regulatory proteins which have undergone mammalian specific expansion events, and also primate-specific expansion events[293]. In humans they represent the largest group of transcription factors[293], however the specific gene regulatory networks or the DNA sequences they bind to are unknown. It is even unclear whether they bind to DNA specifically, as it has been shown that they can bind to proteins and RNA as well[289,294]. The majority of the nearly 700 C_2H_2[295] zinc finger proteins found in humans are yet to have their functions discovered but those that have, are responsible for an extensive range of roles[293]. For example: FOG, KLF1, GATA 1 and 2 are all C_2H_2 zinc finger proteins involved in the development and differentiation of erythrocytes in humans[252].

In a collaborative project[252] my colleagues discovered a novel zinc-finger protein (ZNF648), which when over-expressed results in faster rates of erythroid differentiation and when knocked down impeded erythroid and megakaryocyte differentiation[252]. The structure of ZNF648 is composed of a repeating series of multiple C_2H_2 motifs in the C-terminal domain, and an N-terminal domain with no known motifs or structure (Figure 37). In addition to binding directly to DNA, my collaborators suspect that the additional C_2H_2 repeats may have roles linked to protein-protein or protein-RNA interactions[252]. My role in this work was to investigate the evolution of the protein and its constituent domains with the aim of determining the protein's function.

4.4.2 Methods

Initial ZNF648 sequences were retrieved from ENSEMBL (ENSG00000179930[287]). Multiple BLAST searches were used to identify other ZNF648 sequences from RefSeq[296]. These included several unannotated sequences. More divergent ZNF648 sequences were searched for using sensitive Hidden Markov Model (HMM) profiles against a local database of representative metazoan proteomes using HMMER3[199]. Motif identification was done using ExPASY prosite[297]. All-versus-all BLASTp searches[298] were used to calculate percentage identities of both terminal regions. Heatmaps used for visualization created in R[204] with ggplot[206] and Viridis https://github.com/sjmgarnier/viridis.

Alignments were inferred with MAFFT L-INS-i[221]. All phylogenies were inferred using maximum likelihood models with IQ-TREE 1.6.10[299]. Initial phylogenies were inferred using LG+F+G. The representative phylogeny of ZNF648 was inferred using LG+C60+G+F[77,71], selected under BIC using model testing[299]. We used 1000 ultrafast bootstrap replicates[201] to determine support values.

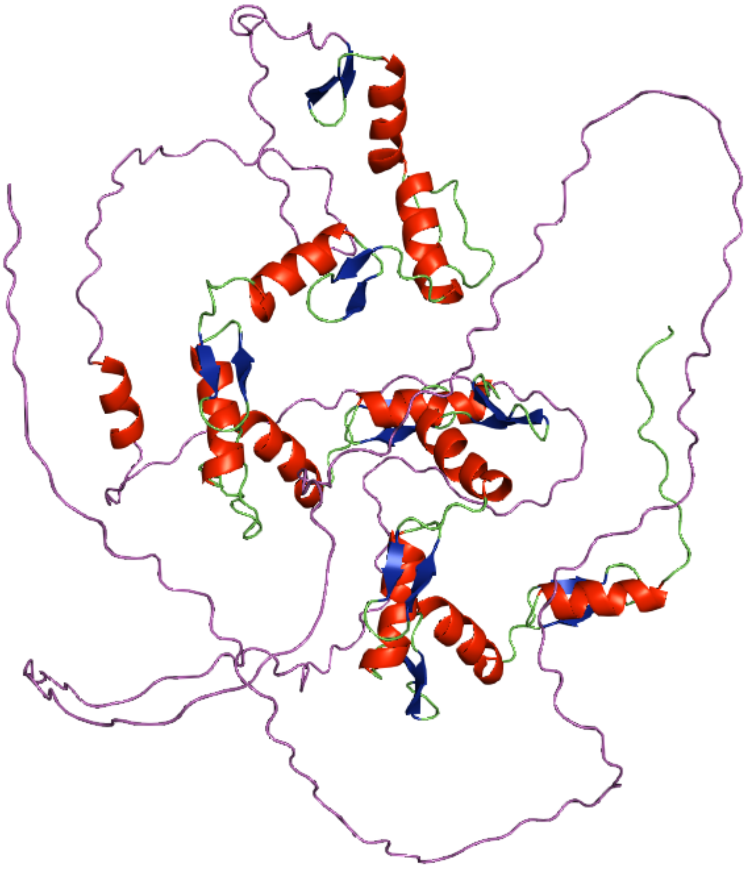

Figure 37 – Alphafold structural prediction for *H. sapiens* ZNF648[284,285], produced with with PyMOL[286]. Disordered N-terminal domain on the left (purple), C-terminal on the right (green), multiple C_2H_2 motifs (α-helices in red, β-sheets in blue) present throughout the C-terminal domain.

4.4.3 Results and Discussion

Phylogenetic range of ZNf648

In order to get an initial grasp on the the range of ZNF648, preliminary testing using profile Hidden Markov Models (HMMER[199]) was done on a representative selection of metazoan genomes. The quick maximum likelihood tree inferred from an alignment of these HMMER hits suggested that the gene evolved in primatomorpha, possibly from a duplication event in the common ancestor of boreoeuthereans (with a subsequent loss in multiple boreoeutherian groups), as the tree shows the primatomorph sequences grouping more closely with carnivoran sequences than in Glires. These results turned out to be a result of carnivoran and primatomorph ZNF648 sequences being found using HMMER, but with paralogous sequences being the top hits, this may have been due to the protein structure being composed of several repeating C_2H_2 motifs, as there are hundreds of proteins across Metazoa with a similar structural organisation.

Switching to a more traditional approach proved to be more fruitful. Using ENSEMBL[287], a much broader range of (annotated) ZNF648 sequences were found, which were aligned ZNF648 sequences with other similar zinc finger proteins and a maximum likelihood tree was inferred (Figure 38). A highly supported monophyletic clade of single-copy ZNF648 sequences in Osteichthyes suggests that the ZNF648 orthologue was present in at least the common ancestor of bony fish, however the automated ENSEMBL pipeline failed to recover the ZNF648 orthologue in multiple teleost species as well as amphibians, lepidosaurians, aves, non-placental mammals and lagomorphs, in addition to multiple individual taxa across the entire tree — but a high number of independent loss events seemed unlikely.

Additional rounds of extensive BLAST searches and tree inference recovered ZNF648's presence in birds and amphibians, however the secondary independent losses in non-placental mammals, lepidosaurians and lagomorphs appear to be legitimate. In order to verify this, we also performed both protein BLASTp against the refseq protein database and translated protein query search tBLASTn against RNA refseq database[296], including the genomes of lampreys and chondrichthyans but no sequences were found which grouped with ZNF648 more closely than other zinc-finger paralogues. Surprisngly, *Pelodiscus sinensis* and *Apteryx owenii* were both annotated as ZNF808 in ENSEMBL, however they fall within the ZNF648 clade when inferring a tree from an alignment including these sequences (as well as other ZNF648 and ZNF808 sequences) and their placement follows the species tree, which suggests, along with the issues found earlier with HMMER and ENSEMBL — that automated pipelines struggle somewhat with this protein. A representative list of ZNF648 sequences were chosen and aligned for a maximum likelihood tree inference (with model selection, LG+C60+F+G) highlighting the independent secondary losses in multiple groups (Figure 39).

Molecular evolution

After establishing the taxonomic range of ZNF648, we wanted to examine the evolution and structure of the C and N terminal domains. Interestingly, when BLASTing the C-terminal domain, many similar hits were found for multiple different zinc finger proteins, whereas the N-terminal domain as a query only gets very closely related sequences as a hit. Although the whole ZNF648 sequence is highly conserved across mammalian taxa, it appears the C-terminus domain is more strongly conserved (94%+, Figure 40) than the N-terminus domain (50%+, Figure 41)[252], which is less conserved within mammalian species and even less conserved across more distantly related taxa. We find stronger N-terminal conservation in more closer related taxa, for example: Haplorhini, Cichliformes, Cryptodira etc. (Figure 41).

This suggests that there are different selection pressures at play on the C-terminal and N-terminal domain, with the N-terminal domain under less functional constraint. This could be the result of functional constraints on specific lineages. For example, the majority of mammals have highly similar N-terminal regions, with the exception of *Mus musculus* — however, we know that murids have evolved much more quickly on average than other mammal groups such as hominids[300]. One possible contributory factor for this could be due to the lack of nuclei in mature mammalian erythrocytes. As the C-terminal domain is so highly similar across all taxa with ZNF648, it is likely that it is performing the same function, which, if similar to other C_2H_2 proteins, is binding to the major grove of DNA using the exposed charges from amino acid side-chains on the α-helix.

Using PHYRE2[301] (software for determining protein fold recognition through finding previously known protein structures), the C-terminal domain was predicted to be a series of C_2H_2 motifs, whereas the N-terminal domain had no similar

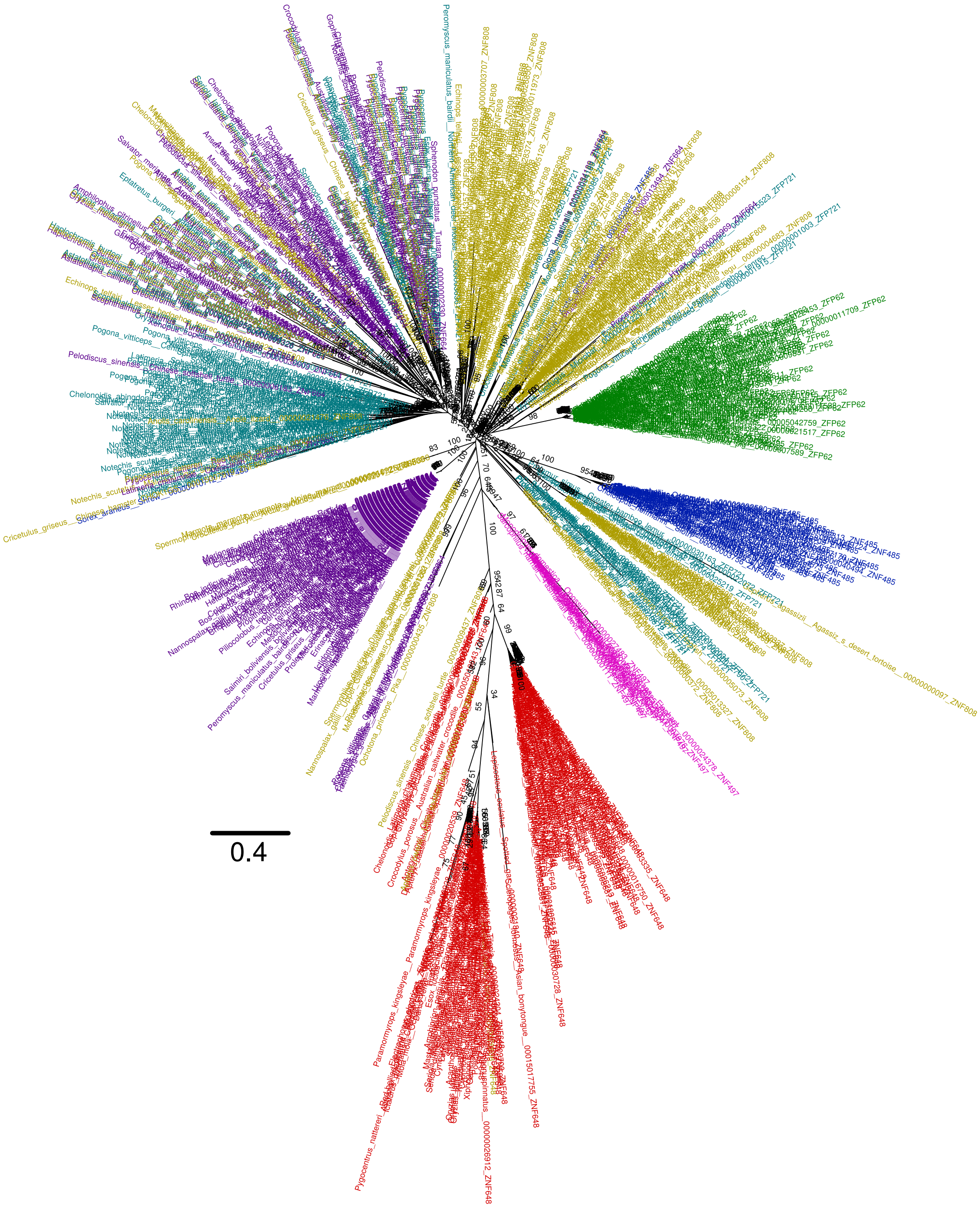

Figure 38 – Unrooted maximum likelihood tree of paralogous zinc-finger proteins. Maximum likelihood tree inferred (LG+F+G, 1000 ultrafast bootstrap replicates) using zinc-finger proteins from ENSEMBL, ZNF648 (red), ZNF808 (olive), ZFP62 (green), ZNF664 (purple), ZNF (pink), ZFP721 (teal). Note the two ZNF808 sequences within the 100% bootstrap supported monophyletic clade of ZNF648. The polyphyly of many of the clades is indicative of the issues with automatic annotation regarding zinc-finger proteins.

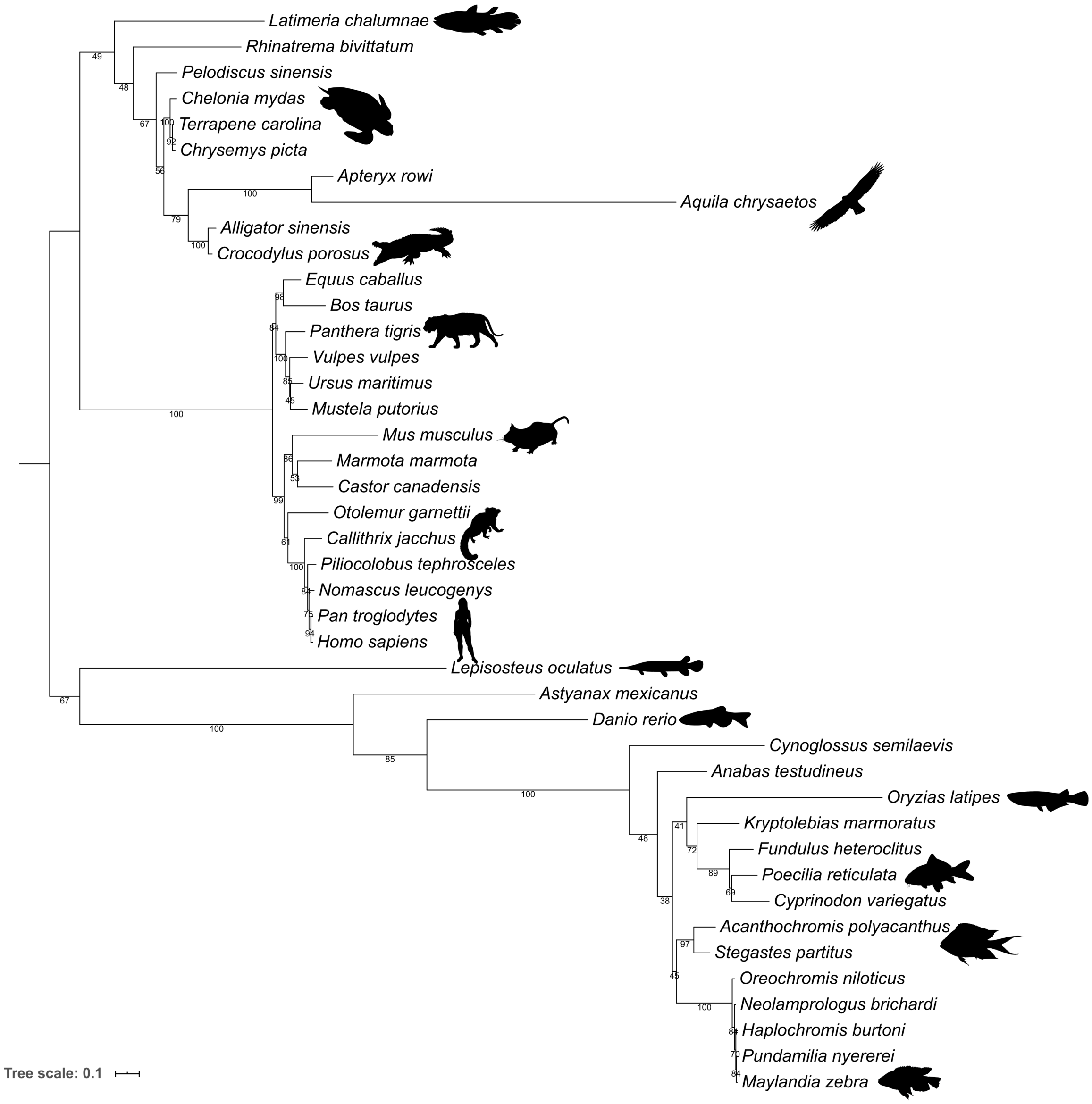

Figure 39 – Maximum likelihood tree of ZNF648. A representative maximum likelihood model with using the best fitting model (LG+F+G+C60) under model finder in IQ-TREE, 1000 ultrafast bootstrap replicates support values shown.

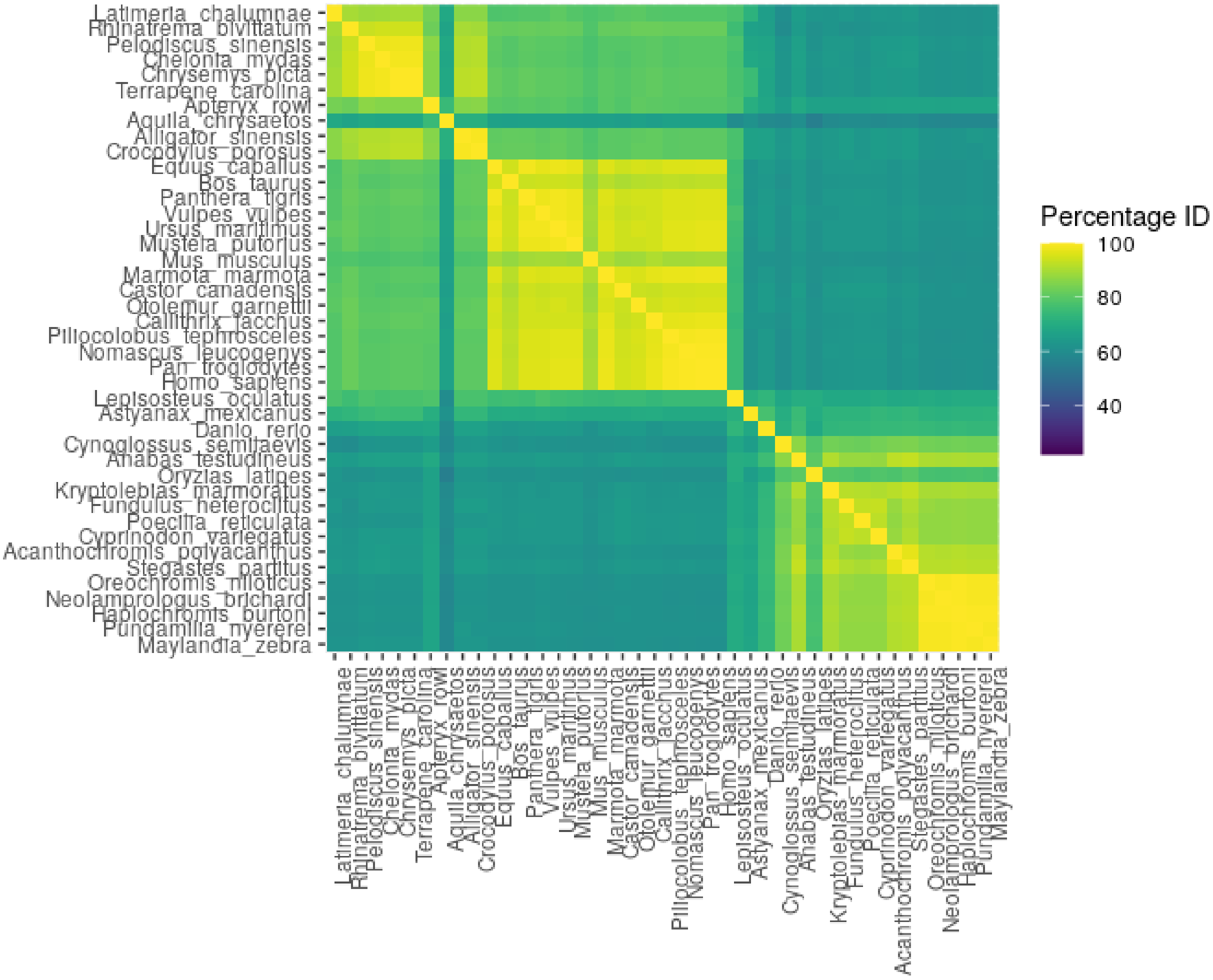

Figure 40 – Heatmap of sequence similarity between the C-terminal domains of representative taxa selection. Corresponding *H. sapiens* amino acid residues 279-568. Generally most of the C-terminal domain sequences are well conserved across the taxa selection, however we see much higher levels of conservation across more closely related species, i.e. within mammals or reptiles.

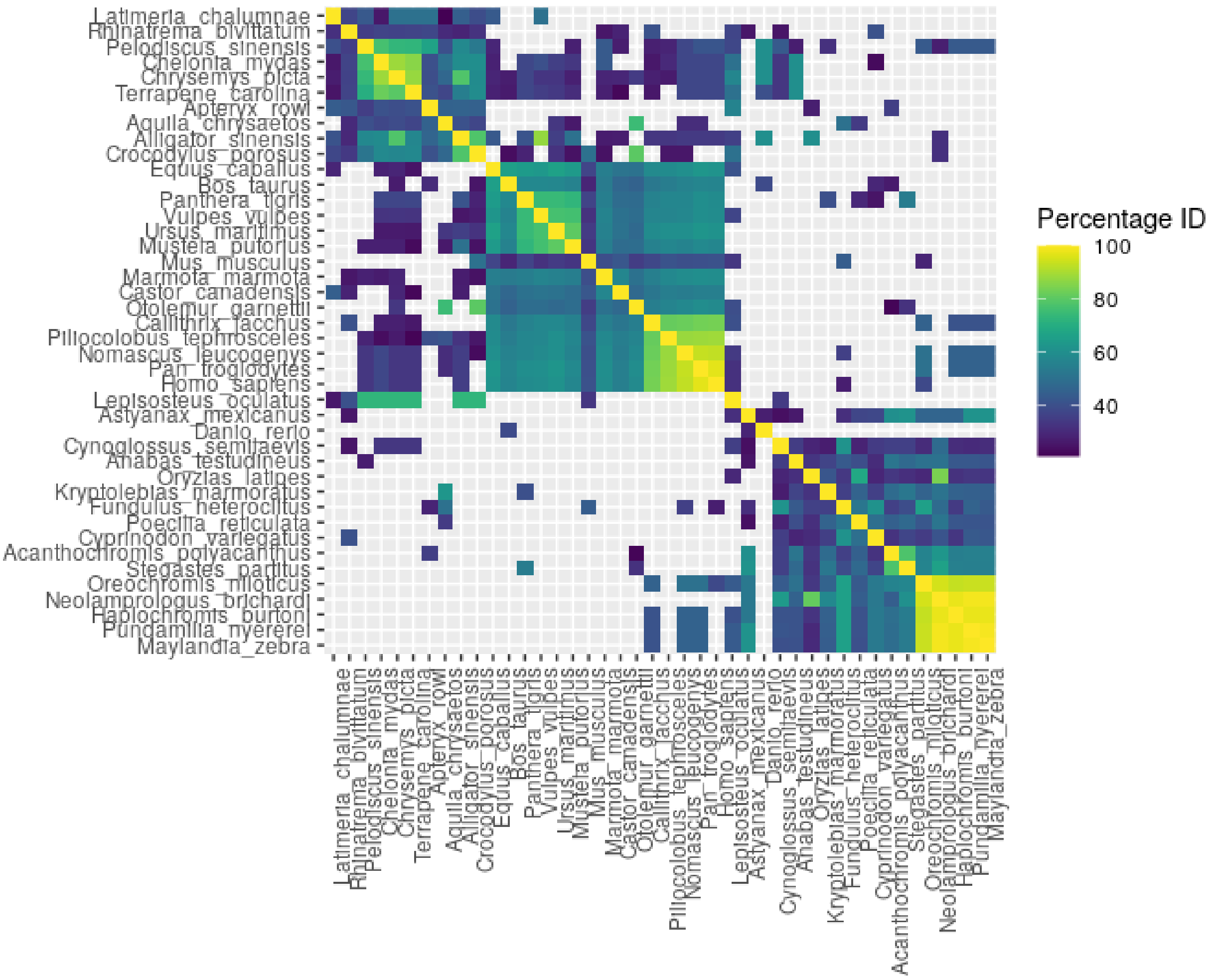

Figure 41 – Heatmap of sequence similarity between the N-terminal domains of representative taxa selection. Corresponding *H. sapiens* amino acid residues 1-278. In comparison with the C-terminal domain, the N-terminal domain is far less conserved across the taxa selection, with many sequences having little to no similarity between them and more distantly related groups, i.e. mammals with reptiles.

structures in any protein database, either suggesting a novel functional domain or an intrinsically disordered domain. For other zinc finger proteins, the non-C_2H_2 containing domain may be for protein-protein interaction, or perform some other as of yet unknown function or may adopt some structure in the presence of another macromolecule[302,303]. Incidentally, a low similarity between proteins is not necessarily indicative of a different function or structure[128], but might also be explained by a large species-level divergence or co-evolution with other proteins. For example, if the N-terminal domain of ZNF648 were binding to a protein which was evolving quickly due to functional necessity, the changes in the N-terminal region of ZNF648 could be compensatory as a result of this, with the structure and/or function being retained.

To assess how the number of C_2H_2 motifs evolved over time, we used EXPASY's prosite[297]. Different lineages do have different numbers of C_2H_2 motifs (Table 6). Mammals have a conserved C-terminal domain containing 10 C_2H_2 zinc-finger motifs, the only amphibian we could find with the sequence had 9 C_2H_2s, whereas the coelocanth also has 11. Reptiles display the largest fluctuation with *Crocodylus porosus* having 11, but with *Alligator sinensis* and the turtle sequences having a seemingly completely novel additional run of C_2H_2 zinc fingers within their N-terminal region, with *Terrapene carolina* having six C_2H_2 fingers in this additional run, and the other turtles and alligator all having seven. Birds seemed to have gone the other way and appear to have lost multiple C_2H_2 motifs, with *Aquila chrysaetos* having seven and *Apteryx rowi* having nine. The variation in the number of C_2H_2 motifs across Osteichthyes (7-18) suggests that the functional constraints of the protein are changing through time.

One possible explanation for the stark differences in conservation across the N-terminal and C-terminal domains is that they have separate evolutionary histories and arose from a gene fusion event. To test this we split the unaligned sequences based on the structural prediction of where the C_2H_2 motifs begin in the sequence, we then aligned the N-terminal and C-terminal domains separately, and inferred the maximum likelihood trees (Figures 42 and 43). These trees both follow the same species tree, with some taxa jumping around, but that is to be expected given the reduced signal, suggesting that they evolved together but at different rates.

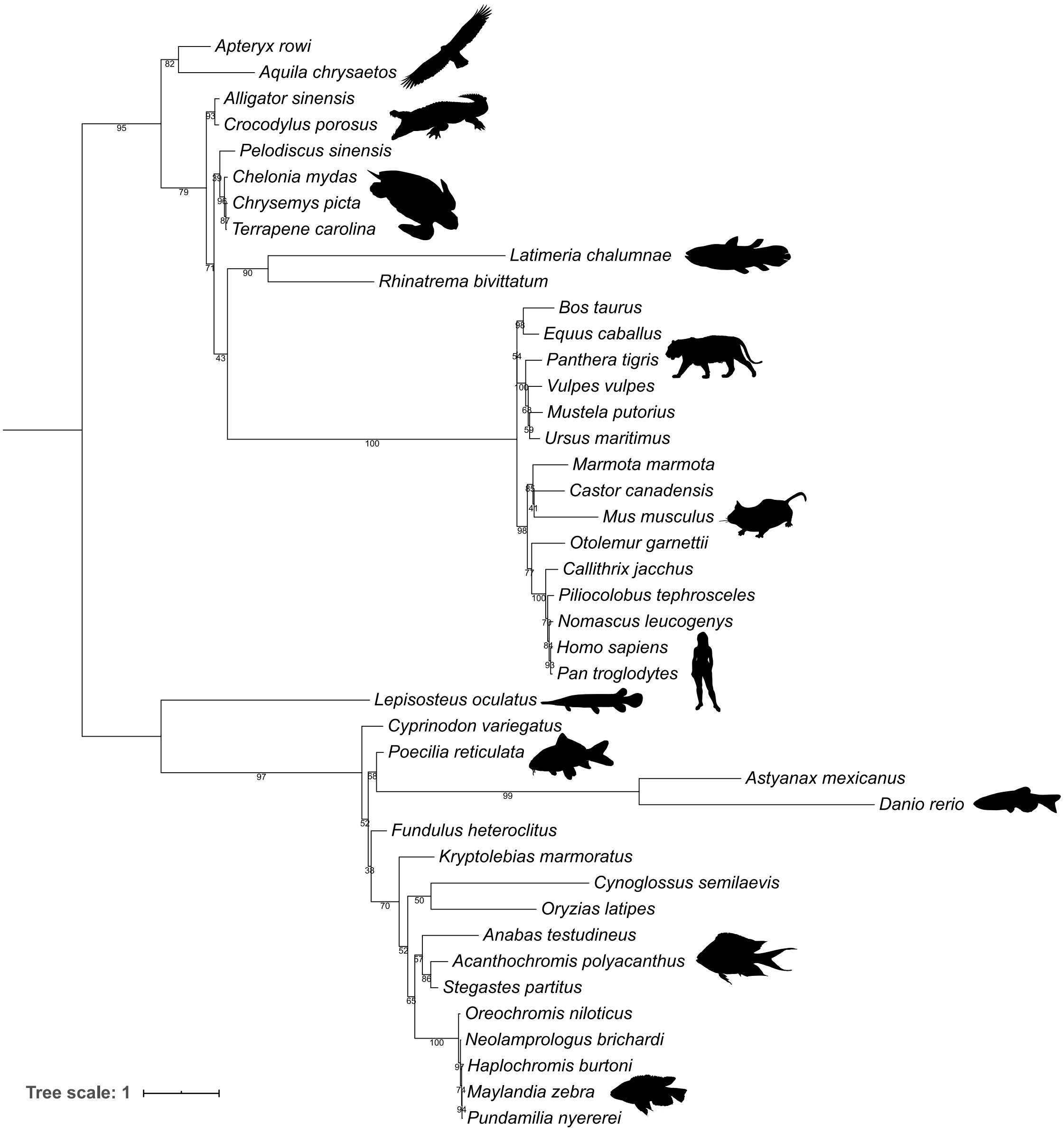

Figure 42 – Maximum likelihood tree under best fitting model (LG+C60+F+G) from a representative taxa selection. On the N-terminal domain *H. sapiens* amino acid residues 1-278, we find many short branches (within clades) due to the high amount of conservation between closely related species, but long branches between clades, i.e. mammals, reptiles and teleosts.

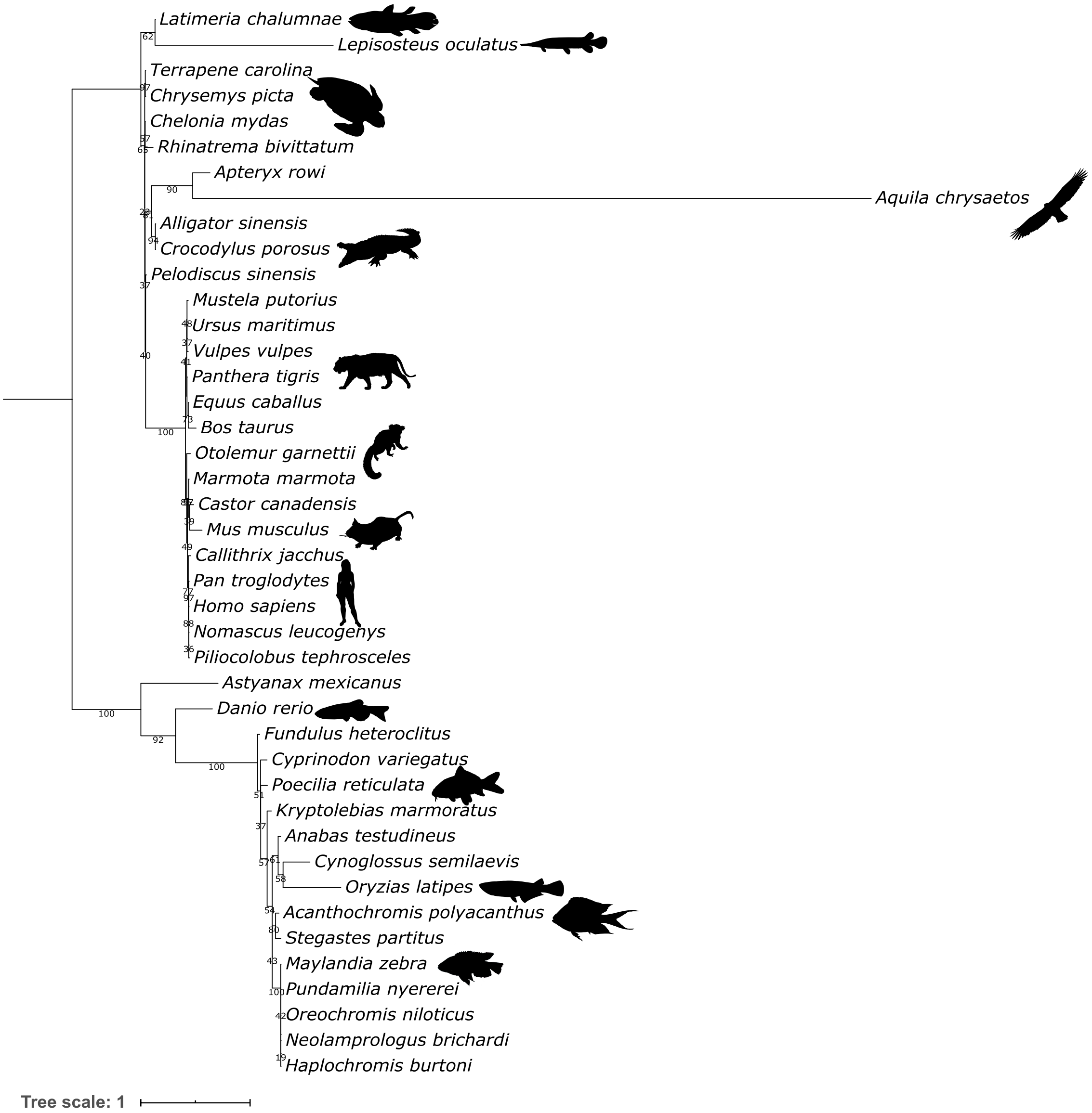

Figure 43 – Maximum likelihood tree under best fitting model (LG+C60+F+G) from a representative taxa selection. On the C-terminal domain *H. sapiens* amino acid residues 279-568. We find most branches are short due to the high level of conservation across the C_2H_2 motifs, however a strikingly long branch is present in the *Aquila chrysaetos*, most likely due to the reduced number of C_2H_2 motifs in that orthologue (seven).

Group	**Number of C_2H_2 Zinc Finger Motifs**
Mammalia	10
Eureptillia	7-18
Lissamphibia	9
Latimeria	11
Lepiosteus	7
Euteleostei	10
Ostariophysi	7-11

Table 6 – Number of C_2H_2 Zinc finger motifs retained across groups of taxa. There is as much variation within the reptiles as the rest of the orthologous ZNF648. We see multiple losses from the (predicted) ancestral 11 C_2H_2 fingers.

4.4.4 Conclusion

Together the phylogenetic analysis combined with the laboratory work[252] show that the presence of ZNF648 across the osteichthyan lineage is indicative of the need for erythroid and megakaryocyte differentiation. This is but a small piece of the puzzle of the evolution of blood but is consistent considering that the markedly different gnathostome haemoglobin[304] evolved somewhere after the split from lampreys and the common ancestor of gnathostomes[305]. This provides a potential explanation for the absence of ZNF648 in *P. marinus* which has multiple paralogous haem-containing-globins[306], however this does not explain the absence in chondrichthyan taxa (Figure 39), one possible explanation could be down to the duplication of the α and β subunits of haemoglobin in the ancestral gnathostome[307], could chondrichthyans have evolved an alternative solution for ZNF648's role?

The multiple independent losses of ZNF648, but the ubiquitous nature of haemoglobin in gnathostomes suggests that many groups such as amphibians (the only member of which a ZNF648 orthologue could be found was in the caecilian *Rhinatrema bivattatum*), marsupials, monotremes and Lepidosauria have developed other solutions which replace the need for ZNF648. The stark contrast in the conservation of the N- and C-terminal domains could suggest different protein-protein interactions in the disordered region of the N-terminal domain across groups or perhaps even a different role completely. In birds, the small number of sequences found had a reduced number of C_2H_2 finger motifs. Perhaps we are seeing a degradation of the protein across Aves, and over the next few million years these species will have lost the ZNF648 completely. Other reptiles, however, such as the turtles and alligator have perhaps taken a different route, with an increase in the number of C_2H_2 motifs, one possibility is that the role of ZNF648 is functionally redundant with another paralogous zinc-finger protein (Figure 38), and whether or not a given gnathostome has kept the ZNF648 orthologue or perhaps some other paralogue is sheer chance.

4.5 Evolution of the SNX27-Retromer-ESCPE-1 coupling mechanisms

4.5.1 Introduction

Most eukaryotic cells contain an organelle composed of multiple flattened stacks of membranes called the Golgi apparatus. The Golgi apparatus is responsible for packaging, distributing and collecting useful protein products from within the cell. The cis Golgi network is responsible for collecting protein cargo produced in the endoplasmic reticulum, where they are then processed before exiting the Golgi apparatus within a vesicle from the trans Golgi network (TGN) before being sent to their respective destinations (the cell's plasma membrane through exocytosis or perhaps a lysosome).

In addition to exocytosis, the TGN also plays a role in the endocytic pathway. Endosomes are vesicle-like organelles responsible for holding material that has entered through the plasma-membrane. This material can then either be disposed of in the lysosome, or allow important materials (such as receptor proteins) to be recycled. The retromer is one of the protein complexes responsible for the recycling of endosomal material, either back to the TGN for re-processing or back to the cell's surface plasma membrane. The retromer complex is found across all eukaryotes, but both its function and structure differ somewhat in plants[308] and in yeast (where it was first discovered)[309] in comparison to the metazoan retromer.

In *Saccharomyces cerevisiae* the retromer is a stable heteropentamer composed of multiple vacuole protein sorting (Vps) proteins: Vps35, Vps29, Vps26, Vps5 and Vps17. In humans and other mammals, the paralogous sorting nexins, SNX1 and SNX2 are orthologous to VPS5 — similarly SNX5 and SNX6 are orthologues of

VPS17. The retromer complex is more ephemeral, and consists of Vps35, Vps26 and Vps29[309], which is able to couple with SNX1/2 and SNX5/6, a sorting nexin complex known as ESCPE-1 (Endosomal SNX-BAR sorting complex for promoting exit 1) or SNX27, another sorting nexin responsible for recycling cargo back to the plasma-membrane[253]. As part of the recycling process, the SNX27-retromer binds with ESCPE-1 in order to allow the cargo to be rescued and moved to ESCPE-1, from SNX27-retromer[253].

In a collaborative project[253] my colleagues combined biochemical, structural and cellular approaches to discover the mechanisms behind SNX27-retromer ESCPE-1 coupling. They confirmed the FERM[310] domain's importance in retromer mediated lysosomal retrieval, the binding of SNX27 to SNX1 and SNX2 of the ESCPE-1 complex through CRISPR gene-editing, and the expression of mutants lacking the FERM domain[253]. Through site-directed mutagenesis they also found the specific motifs in SNX1 and SNX2 which associate with the SNX27 FERM domain which are acidic-asparagine-leucine-phenylalanine (aDLF)[253] (Figure 44). They also confirmed that the association of ESCPE-1 with SNX27 is necessary for retrieving the cargo and then recycling it[253] through a series of rescue experiments on knockout HeLa cells, and mutants lacking the ability to bind to ESCPE-1[253]. My role in the project was to determine about the evolutionary history of the interactions between SNX27, retromer, and ESCPE-1 at both the protein and motif level, through the phylogenetic analysis of SNX27, SNX1, SNX2, Vps5 and Vps26.

4.5.2 Methods

Homologous sequences were retrieved using BLAST[298] against all non-redundant GenBank coding sequences. Ctenophore data was retrieved using BLAST on sequences from the Mnemiopsis Genome Project Portal[311,312,313]. Sequences were aligned using MAFFT L-INS-i[221]. Iterative rounds of tree inference were used to identify duplication events (IQ-TREE 2.1.4[72]). Exploratory trees were inferred under the LG+G+F model using 1000 ultrafast bootstrap replicates. Orthologues were found in a range of taxa, and more rounds of tree inferences allowed us to identify paralogous gene sequences. More refined maximum likelihood trees were inferred under the model with the lowest BIC score: we included site-rate heterogeneous models (LG+C10...C60)[80], derived amino acid frequencies from the data (+F), and

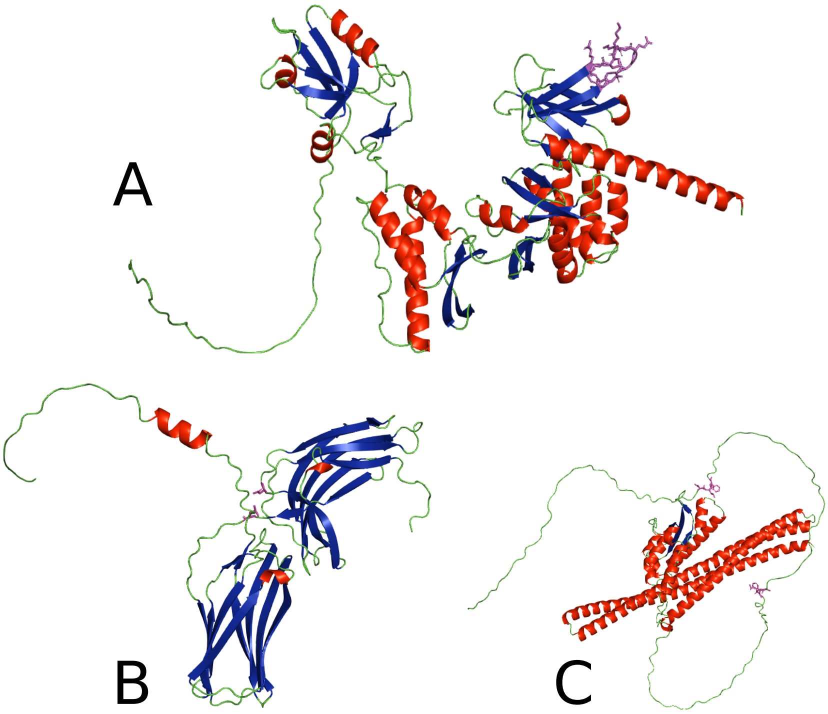

Figure 44 – Alphafold structural prediction of SNX27, VPS26B and SNX1[284,285]. SNX27 (A), the PDZ β-hairpin loop is in pink, with residues as sticks, VPS26B (B) with the D44 and L154 residues shown as sticks in pink, and SNX1 (C) with the acidic-DLF motifs shown as pink sticks. α-helices in red, beta-sheets in blue. Produced with with PyMOL[286]

allowed for across-site rate variation using either a Gamma distribution with four discrete rate categories (+G)[314] or using the free rate model (+R)[315,316]. Support was estimated using 10000 ultrafast bootstraps[201]. LG+C30+F+G was the model selected for the final representative SNX1 and SNX2 trees, and LG+C20+F+G were selected for the SNX27 and VPS26 trees.

4.5.3 Results and Discussion

SNX27

To find the taxonomic range of SNX27, we retrieved the top hit from a BLASTp search against a range of taxa, as SNX27 has not been described in the literature as being present in *Saccharomyces cerevisiae.* We included representatives from Metazoa, Filasteria, Ichthyosporea, choanoflagellates, and Fungi with the top hit from Amoebozoa as an outgroup. The inferred topology is poorly resolved, with low bootstraps, but places the Amoeba sequence as sister to one of the fungal sequences. Surprisingly, the sequence from the ichthyosporean is also included in the same clade along with the other fungal sequence. This did not definitively limit SNX27 to any of the groups present, as although it is odd to have the ichthyosporean sequence within the fungal clade, with such low support and only a small number of non-metazoan representatives. The sequences roughly follow the species tree (with a small number of exceptions), therefore it is possible that all the sequences were orthologues of SNX27. To refine the search, we included the top two hits from the metazoan sequences (annotated as SNX17) as well as the top hits from Viridaeplantae. The inferred topology from this expanded search resolved a monophyletic clade of SNX17 sequences from across Parahoxozoa (Cnidaria, Bilateria, and Placozoa) as well as the choanoflagellate sequences.

We found a clade containing fungal sequences, amoebozoan sequences, as well as the second top hit from *Ciona intestinalis*, *Caenorhabditis elegans* and *Branchiostoma floridae*, along with the ichthysporan sequence from *Sphaeroforma arctica.* Another clade included the majority of sequences annotated as SNX27, but also included the plant sequence for *Rhodamina argentea* which was surprising. As a result, I performed a reciprocal BLAST search on this sequence, the closest hits were ants, suggesting that is a contamination error, as the other hits from plants were not recovered (e-value threshold of 0.05). After removing the contaminated sequence, and the sequences

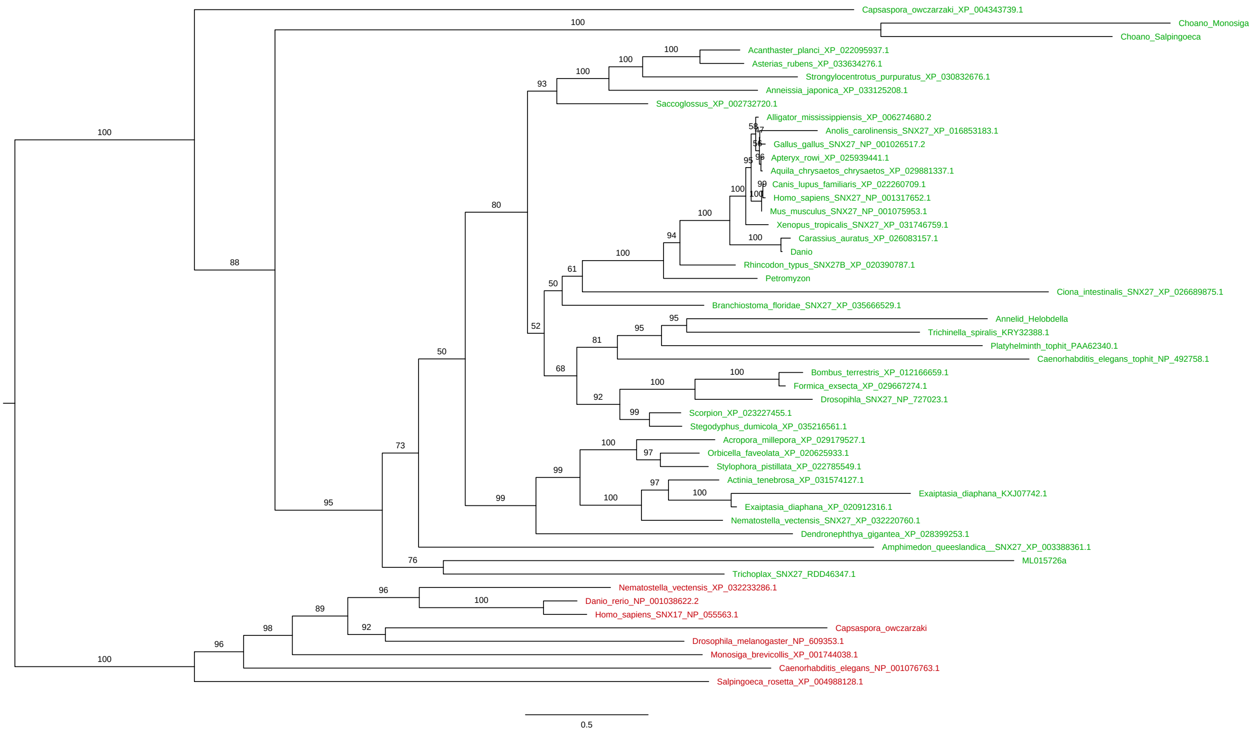
Capsaspora_owczarzaki_XP_004343739.1
100
Choano_Monosiga
Choano_Salpingoeca
100
Acanthaster_planci_XP_022095937.1
100
Asterias_rubens_XP_033634276.1
100
Strongylocentrotus_purpuratus_XP_030832676.1
93
Anneissia_japonica_XP_033125208.1
Saccoglossus_XP_002732720.1
Alligator_mississippiensis_XP_006274680.2
100
58
47
Anolis_carolinensis_SNX27_XP_016853183.1
56
Gallus_gallus_SNX27_NP_001026517.2
96
95
Apteryx_rowi_XP_025939441.1
Aquila_chrysaetos_chrysaetos_XP_029881337.1
100
99
Canis_lupus_familiaris_XP_022260709.1
100
Homo_sapiens_SNX27_NP_001317652.1
80
Mus_musculus_SNX27_NP_001075953.1
100
Xenopus_tropicalis_SNX27_XP_031746759.1
94
100
Carassius_auratus_XP_026083157.1
100
Danio
88
Rhincodon_typus_SNX27B_XP_020390787.1
61
Petromyzon
50
Ciona_intestinalis_SNX27_XP_026689875.1
Branchiostoma_floridae_SNX27_XP_035666529.1
95
Annelid_Helobdella
52
95
Trichinella_spiralis_KRY32388.1
81
Platyhelminth_tophit_PAA62340.1
50
Caenorhabditis_elegans_tophit_NP_492758.1
68
100
Bombus_terrestris_XP_012166659.1
100
Formica_exsecta_XP_029667274.1
92
Drosopihla_SNX27_NP_727023.1
99
Scorpion_XP_023227455.1
Stegodyphus_dumicola_XP_035216561.1
73
100
Acropora_millepora_XP_029179527.1
97
Orbicella_faveolata_XP_020625933.1
Stylophora_pistillata_XP_022785549.1
99
97
Actinia_tenebrosa_XP_031574127.1
95
99
100
100
Exaiptasia_diaphana_KXJ07742.1
Exaiptasia_diaphana_XP_020912316.1
Nematostella_vectensis_SNX27_XP_032220760.1
Dendronephthya_gigantea_XP_028399253.1
Amphimedon_queeslandica__SNX27_XP_003388361.1
76
ML015726a
Trichoplax_SNX27_RDD46347.1
96
Nematostella_vectensis_XP_032233286.1
100
Danio_rerio_NP_001038622.2
89
Homo_sapiens_SNX17_NP_055563.1
98
92
Capsaspora_owczarzaki
100
96
Drosophila_melanogaster_NP_609353.1
Monosiga_brevicollis_XP_001744038.1
Caenorhabditis_elegans_NP_001076763.1
Salpingoeca_rosetta_XP_004988128.1
0.5

Figure 45 – Maximum likelihood tree of SNX27 and SNX17. Maximum likelihood (LG+C20+F+G, best fitting model according to BIC, 1000 ultrafast bootstrap replicates) topology of SNX27 (green) with SNX17 (red) as an outgroup.

leading to very long branches, a monophyletic clade of non-filozoan sequences emerges with Fungi, Ichthyosporea, and amoebazoans. From what we know about the species tree, if the fungal and ichthyosporean sequences were SNX27 orthologues, they should group within the other opisthokont sequences rather than with the amoebozoans. After establishing the presence of SNX27 in filozoans, additional iterative rounds of BLASTing, alignment, and tree inference were performed in order to have a balanced taxonomic representation across Filozoa for the motif evolution analyses. The final tree includes SNX17 as an outgroup (Figure 45).

VPS26

In order to find the range of orthologous VPS26 sequences, *H. sapiens* VP26 was used as the query sequence, and the top hits from a range of obazoan taxa were used, as well as the top hit from a range of plants as an outgroup. The inferred topology from this analysis places plants as the outgroup, it resolves Fungi as a monophyletic clade with Ichthyosporea as a sister group, and with choanozoans broadly following the species tree. There is a duplication event in the common ancestor of at least *H. sapiens* and *D. rerio*. Since further rounds of testing found three paralogues of VPS26 in *H. sapiens* and *P. marinus*, and two in plants and fungi; it seems that the inferred topology broadly follows the species tree.

This analysis confirms the duplication of VPS26 happened multiple times within eukaryotes. The first duplication leading to paralogues VPS26C and VPS26B must have occurred within or before the last common ancestor of plants and humans, which many would agree would be the last eukaryotic common ancestor[317]. However, there is debate on the topic, others have suggested perhaps the root of eukaryotes lies somewhere else, and that the last common ancestor of plants and animals would be the ancestral neokaryote[318].

Another duplication has taken place in the common ancestor of vertebrates as evidenced by the presence of VPS26A and VPS26B in *P. marinus*, *D. rerio* and *H. sapiens* and their placement as expected similar to the conventional species tree. A separate duplication event has also taken place within plants, and although they are annotated as VPS26A and VPS26B, the duplication is not orthologous to the duplication of VPS26A and VPS26B in vertebrates (Figure 46).

SNX1 and SNX2

As we know SNX1 and SNX2 are present in *H. sapiens*, and VPS5 is an orthologue found in yeast. Retromer is present across eukaryotes, but as its structure differs across the domain[308], we included outgroup representatives from Archaea, plants, and amoebozoa, as well as a range of sequences from Fungi, Metazoa and top hits from the limited pool of choanoflagelates, filasterians and ichthyosporeans present on Genbank[75]. The resulting topology is poorly supported, and we suspect suffers from a mixture of lack of signal and long branch attraction. Aside from a well supported (99% bootstrap support) metazoan clade, other expected groups, such as Fungi and Filasteria, were not found to be monophyletic and had poor support. As we were mainly interested in the evolution of the mechanism between ESCPE-1 and SNX27-retromer, and the presence of a SNX1 orthologue is already known in opisthokonts, we narrowed the search to metazoans and choanoflagellates (the closest paralogue to SNX1 and SNX2 within *Capsapora* was SNX7). The duplication of SNX1 and SNX2 as with VPS26 occurred within the common ancestor of vertebrates

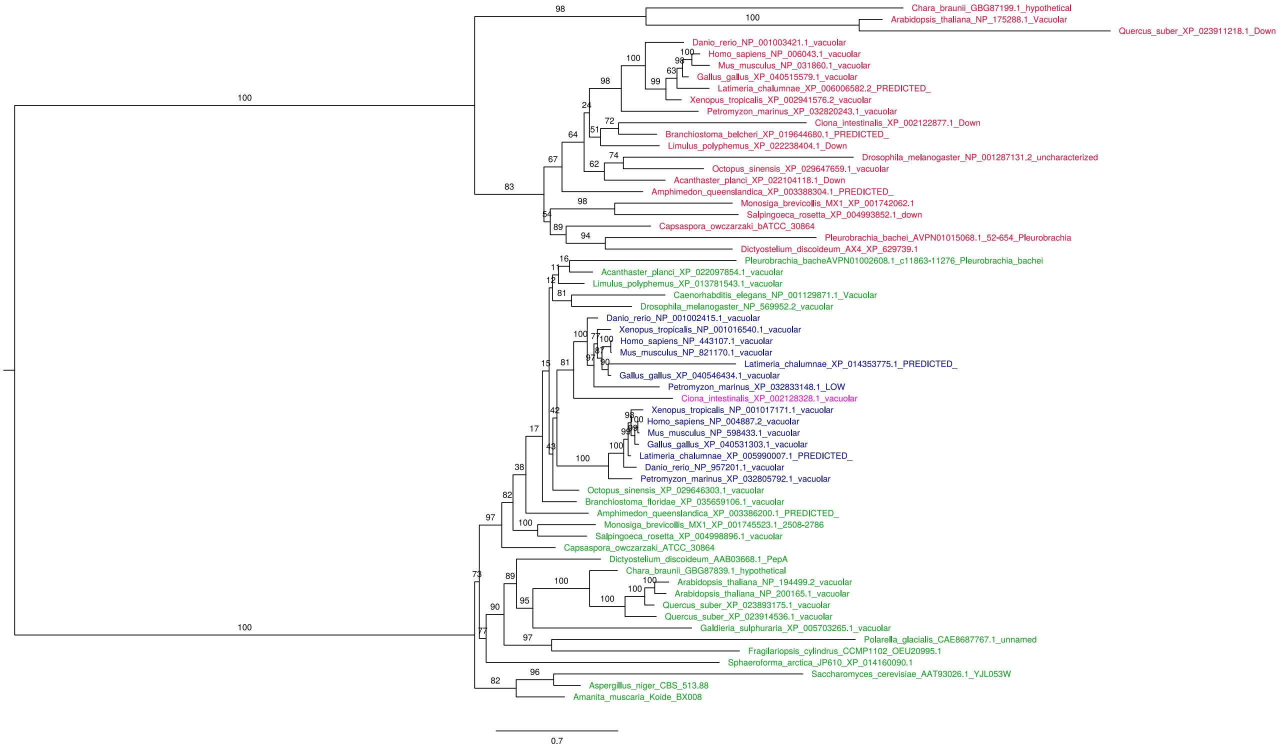
Chara_braunii_GBG87199.1_hypothetical
Arabidopsis_thaliana_NP_175288.1_Vacuolar
Quercus_suber_XP_023911218.1_Down
Danio_rerio_NP_001003421.1_vacuolar
Homo_sapiens_NP_006043.1_vacuolar
Mus_musculus_NP_031860.1_vacuolar
Gallus_gallus_XP_040515579.1_vacuolar
Latimeria_chalumnae_XP_006006582.2_PREDICTED_
Xenopus_tropicalis_XP_002941576.2_vacuolar
Petromyzon_marinus_XP_032820243.1_vacuolar
Ciona_intestinalis_XP_002122877.1_Down
Branchiostoma_belcheri_XP_019644680.1_PREDICTED_
Limulus_polyphemus_XP_022238404.1_Down
Drosophila_melanogaster_NP_001287131.2_uncharacterized
Octopus_sinensis_XP_029647659.1_vacuolar
Acanthaster_planci_XP_022104118.1_Down
Amphimedon_queenslandica_XP_003388304.1_PREDICTED_
Monosiga_brevicollis_MX1_XP_001742062.1
Salpingoeca_rosetta_XP_004993852.1_down
Capsaspora_owczarzaki_bATCC_30864
Pleurobrachia_bachei_AVPN01015068.1_52-654_Pleurobrachia
Dictyostelium_discoideum_AX4_XP_629739.1
Pleurobrachia_bacheAVPN01002608.1_c11863-11276_Pleurobrachia_bachei
Acanthaster_planci_XP_022097854.1_vacuolar
Limulus_polyphemus_XP_013781543.1_vacuolar
Caenorhabditis_elegans_NP_001129871.1_Vacuolar
Drosophila_melanogaster_NP_569952.2_vacuolar
Danio_rerio_NP_001002415.1_vacuolar
Xenopus_tropicalis_NP_001016540.1_vacuolar
Homo_sapiens_NP_443107.1_vacuolar
Mus_musculus_NP_821170.1_vacuolar
Latimeria_chalumnae_XP_014353775.1_PREDICTED_
Gallus_gallus_XP_040546434.1_vacuolar
Petromyzon_marinus_XP_032833148.1_LOW
Ciona_intestinalis_XP_002128328.1_vacuolar
Xenopus_tropicalis_NP_001017171.1_vacuolar
Homo_sapiens_NP_004887.2_vacuolar
Mus_musculus_NP_598433.1_vacuolar
Gallus_gallus_XP_040531303.1_vacuolar
Latimeria_chalumnae_XP_005990007.1_PREDICTED_
Danio_rerio_NP_957201.1_vacuolar
Petromyzon_marinus_XP_032805792.1_vacuolar
Octopus_sinensis_XP_029646303.1_vacuolar
Branchiostoma_floridae_XP_035659106.1_vacuolar
Amphimedon_queenslandica_XP_003386200.1_PREDICTED_
Monosiga_brevicollis_MX1_XP_001745523.1_2508-2786
Salpingoeca_rosetta_XP_004998896.1_vacuolar
Capsaspora_owczarzaki_ATCC_30864
Dictyostelium_discoideum_AAB03668.1_PepA
Chara_braunii_GBG87839.1_hypothetical
Arabidopsis_thaliana_NP_194499.2_vacuolar
Arabidopsis_thaliana_NP_200165.1_vacuolar
Quercus_suber_XP_023893175.1_vacuolar
Quercus_suber_XP_023914536.1_vacuolar
Galdieria_sulphuraria_XP_005703265.1_vacuolar
Polarella_glacialis_CAE8687767.1_unnamed
Fragilariopsis_cylindrus_CCMP1102_OEU20995.1
Sphaeroforma_arctica_JP610_XP_014160090.1
Saccharomyces_cerevisiae_AAT93026.1_YJL053W
Aspergillus_niger_CBS_513.88
Amanita_muscaria_Koide_BX008
0.7

Figure 46 – Maximum likelihood tree of paralogous VPS26 sequences. Maximum likelihood tree (LG+C20+F+G, chosen according to BIC, 1000 bootstrap replicates), tree rooted between VPS26C (red), and VPS26B (green), we see the duplication occured likely in the common ancestor of vertebrates (both paralogues in blue), with *Ciona intestinalis* (pink) being placed within the clade of duplicates, making it difficult to distinguish between a duplication and loss in tunicates.

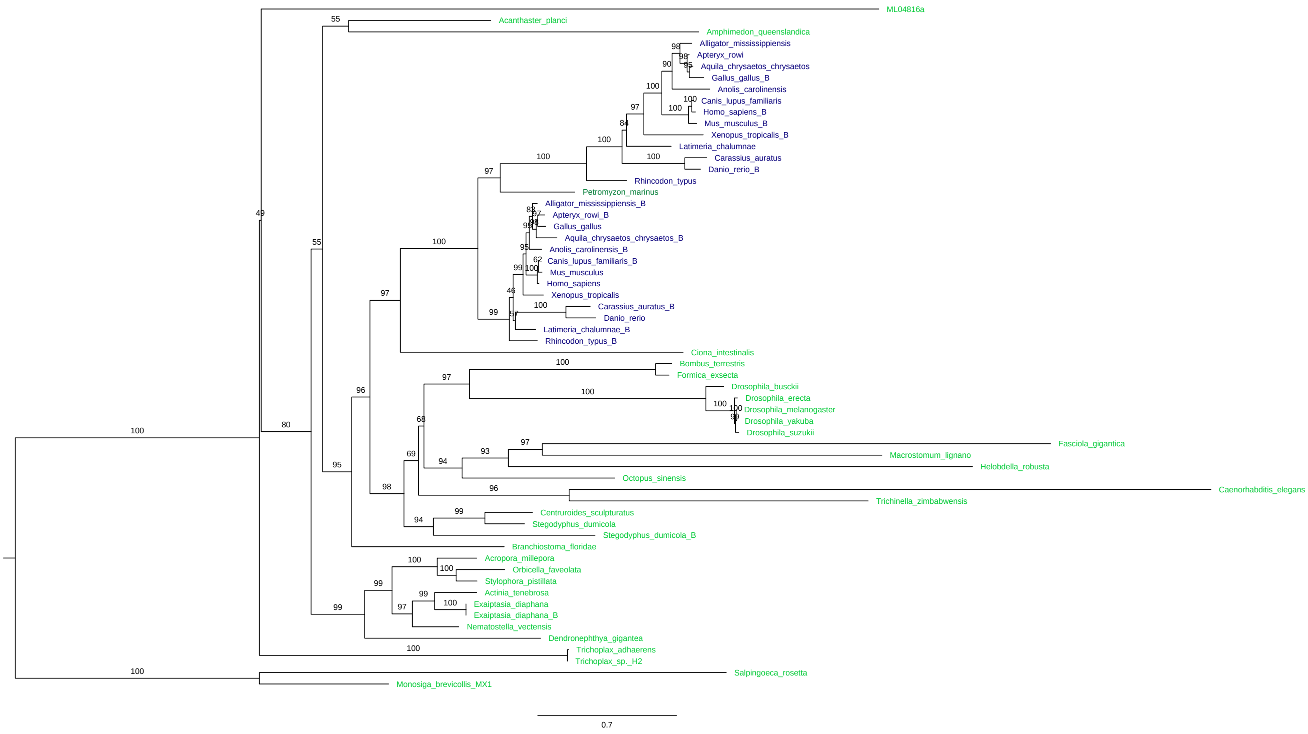

Figure 47 – Rooted maximum likelihood topology (LG+C20+F+G, 1000 bootstrap replicates) of SNX1 (green) and SNX2 (blue). Tree shows the duplication occurred by at least the common ancestor of gnathostomes, *P. marinus* is placed between the two clades, suggesting either duplication in vertebrate common ancestor, and subsequent loss in the *P. marinus* or the duplication happened afterwards in the common ancestor of *Rhincodon typus* and *H. sapiens*.

(or at least gnathostomes), as only one orthologue was present in *P. marinus*.

SNX5 and SNX6

SNX5 and SNX6 were not the main focus of the work, but some preliminary tests show multiple duplications, with vertebrate SNX5, SNX6, and SNX32 all being paralogues of each other. However due to subsequent loss of a paralogue in arthropods and low support, the resulting topology resolves an independent duplication in *C. intestinalis*, *Hydra vulgaris*, *Nematostella vectensis*, *B. floridae* and *Acanthaster planci*. This history seems unlikely unless multiple independent duplication and subsequent loss events occurred. As the tree support is so low, it is difficult to exclude the possibility that SNX5 and SNX6 are paralogues from a duplication event which dates back to at least the last common parahoxozoan (only one orthologue was present in

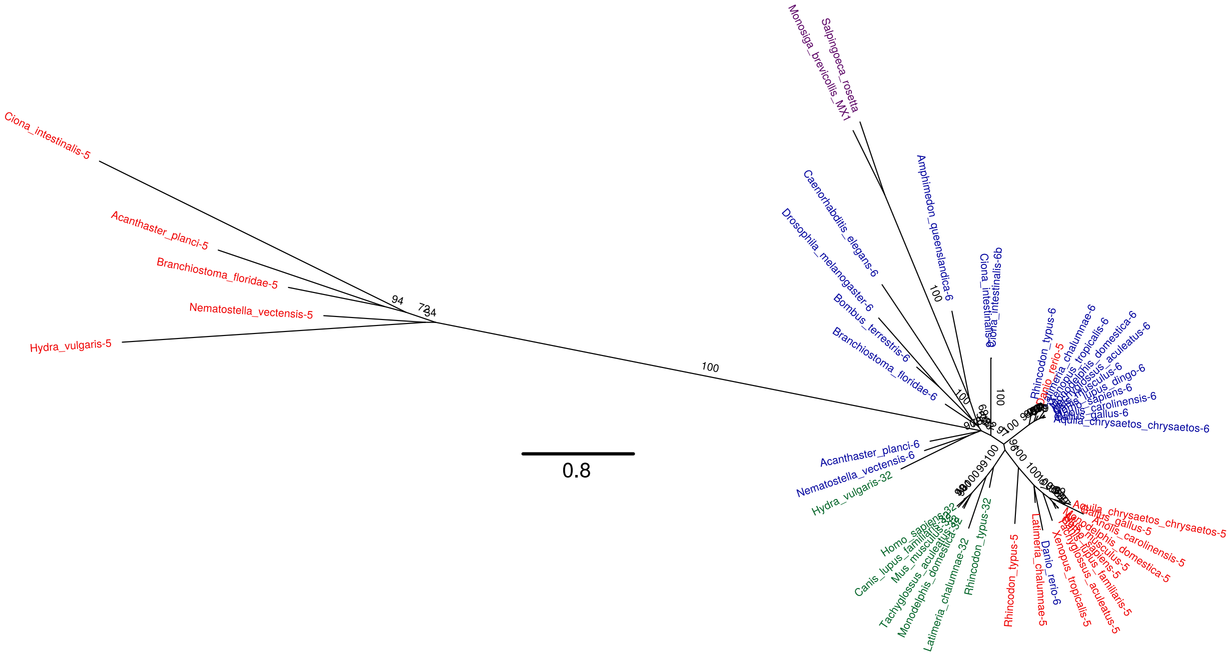

Figure 48 – Unrooted maximum likelihood (LG, 1000 bootstrap replicates) topology of SNX5 (red), SNX6 (blue) and SNX32 (green).

Amphemedon queenslandica and in the choanoflagellate sequences). An additional duplication of SNX5 is also likely to have occurred in vertebrates, resulting in SNX32, however it must be noted that the annotation of some sequences is confusing and incorrect (a SNX32 sequence is present in *H. vulgaris* and *D. rerio*'s SNX5 and SNX6 are inverted).

Evolution of binding mechanism

SNX27-Retromer

Using the multiple-sequence alignments of the representative phylogenies above with alphafold structural prediction we have found that although SNX27 appears to be present across Filozoa, the exposed β-hairpin[253] found in *H. sapiens* residues: 67-79, has an orthologous motif present only in Chozoanoa, and not present in the *Capsaspora owczarzaki* sequence included in the alignment. This β-hairpin mediates the binding of VPS26 to SNX27[319,253]. Across Choanoza, L67 and L74 which are a vital component of the β-hairpin motif, are present in SNX27 orthologues. The corresponding groove in VPS26 includes, residues D44 and L154 (VPS26A, *H. sapiens*), and D42, L152 (VPS26B, *H. sapiens*)[319,253]. The D44 residue is present across Filozoa, but the L154 is constrained to Choanozoa, implying either the common ancestor of Filozoa either binds with less affinity, or the corresponding residues in SNX27 were also suitably different to establish a similar mechanism. The binding mechanism between SNX27 and retromer, as we understand it, evolved in the last common ancestor of choanozoans, as seen in Figure 49.

SNX27-ESCPE-1

Through biochemical testing and structural prediction, my colleagues determined that ESCPE-1 and SNX27 associate through the FERM domain of SNX27 and specific DLF binding motifs within SNX1/2[253]. SNX1 is present in all eukaryotes, but the duplication of SNX1 (and origin of SNX2) likely happened in the common ancestor of Vertebrata (see above). The binding residues of the FERM domain in SNX27 necessary for the interaction with ESCPE-1 are R437, K495, K496, R498, and K501. Although all of these residues are present by the last common ancestor of cnidarians and bilaterians, the exact evolution of the FERM domain remains unclear. R347, R498, and K501 are present in the last metazoan ancestor, whereas K495 and K496 are present in the common ancestor of Bilateria and Cnidaria (see Figure 49). In the ESCPE-1 dimer, the SNX1/2 component in *H. sapiens* contains an acidic DLF motif within the N-terminus. The first aDLF motif can be traced back to the common ancestor of *Porifera* and *H. sapiens*, suggesting that the SNX27-ESCPE-1 interaction had evolved by the time of last common ancestor of Metazoa. However, surprisingly, the homologous motif is missing in Cnidaria, Placozoa and Ctenophora, which indicate secondary losses of the motif in these groups. There is a second aDLF motif in *H. sapiens* which motif is present across vertebrates and is likely to have existed in the the last common ancestor of Vertebrata, which is also where the paralagous SNX2 most likely duplicated. Theses analyses suggest the SNX27:ESCPE-1 interaction had emerged after the SNX27-retromer link during early metazoan evolution, rather than the common ancestor of choanoflagellates and Metazoa.

4.5.4 Conclusion

Combining my phylogenetic work and the work undertaken by my colleagues[253]. We establish that the ancestral SNX27 in the common ancestor of filozoans likely did not use the same mechanism to bind to the retromer complex. This had evolved

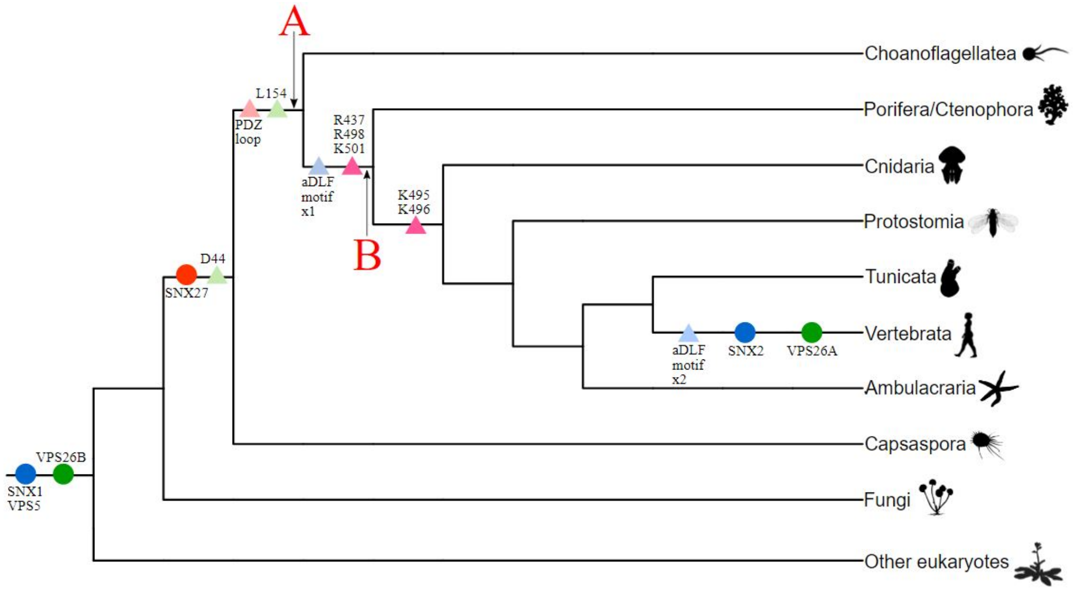

Figure 49 – Summary cladogram incorporating the results of multiple phylogenies. SNX1/VPS5 and VPS26B are present in the common ancestor of all eukaryotes. SNX27 emerged in the common ancestor of Filozoa, along with the *H. sapiens* D44 residue in the VPS26B protein component of retromer which is thought to bind to SNX27. The common ancestor of Choanozoans (A) would have possessed the PDZ loop within SNX27 which binds to VPS26B, and the complimentary *H. sapiens* L154 residue had also evolved by this point. We show that the last common ancestor of Metazoa had the acidic-DLF motif with SNX1 as well as a proportion of the residues of the FERM domain in SNX27 (B), more of these FERM domain residues had been established by the common ancestor of Parahoxozoa. Along the vertebrate stem we find the duplication of SNX1 (SNX2), and VPS26B (VPS26A), along with an additional acidic-DLF motif within the SNX1/2 paralogues.

by the last common ancestor of Choanozoa, consisting of the PDZ β-hairpin loop in SNX27 and the L154 and D44 residues in VPS26. The additional interaction responsible for the switch-over of protein cargo from SNX27, to ESCPE-1, meanwhile has emerged by the last common ancestor of cnidarians and vertebrates, but we do see the acidic-DLF motif found in the last common ancestor of metazoans.

4.6 Conclusion

These three case studies show how the application of phylogenetic methods can be used to help shed light on the evolution of proteins at the sub-protein level: through this study we have learnt that within the kinesin-1 light chains, the mechanisms for attaching to the cargo had evolved by the last common ancestor of choanozoans, the presence of of KLC1 in choanoflagellates is interesting, because we know these unicellular organisms to be far more complex than first thought[320]. Understanding cellular transport in unicellular organisms could help us understand how these complicated processes (first thought to be unique to multicellular organisms[321]) evolved. The multiple duplication events leading to the emergence of the four distinct KLC paralogues in vertebrates are similar to the duplication of the components of the retromer and its associated ESCPE-1 complex. The subsequent loss of the amphipathic-helix in ecdysozoans could suggest an alternative mechanism for binding to cargo in kinesin-1 light chain within these organisms.

Likewise the phylogeny of ZNF648 shows that although the two domains of a cell can be totally different regarding the conservation of their sites, they still follow the same evolutionary history — for example, the lack of hits when using BLAST[298] and HMMER[199] on the N-terminal domain for example is a strong word of caution for relying too much on sequence similarity and goes some way in highlighting the importance of using phylogeny to uncover the history of the genes.

Finally, we see how the SNX27 protein and the mechanisms which allow the protein to attach to retromer and the ESCPE-1 complex evolved gradually through the metazoan lineage, sometime after the constituent parts had all come into existence.

5 CONCLUDING REMARKS

Author's contribution

This chapter was written by Edmund. R.R Moody in its entirety.

In chapter one, we look at how the tree of life has grown from early philosophical beginnings to a highly complex and thriving field at the forefront of scientific research. Although early questions were centred around the presence of three[43] or two[111] domains life, and the taxonomic implications[124,117] of said discussion, scientific consensus is now broadly in alignment that there are indeed two primary domains of life[136,138,27,52,30,54] - albeit slightly different to the 'parkaryotes' and 'karyotes' tree envisaged by Lake[112].

Scientific discussion has become more nuanced. We now wrestle with whether or not it is right to use whole-genome data[139,147,166] due to the impact of HGT and conflicting gene histories, or whether we should be using vast numbers of genes[148] or a small set of markers unlikely to be impacted by duplication, transfer and loss[129].

One of the questions we set out to answer was how do different marker sets affect the inferred phylogeny of the tree of life? In chapter two, we showed that using genes with conflicting evolutionary histories artificially reduce the length of the deepest branch in the tree of life. We also have shown how important the manual curation of contamination, the removal of paralogous and transferred gene sequences, and the danger of inferring a species tree from a concatenation of these conflicting genes under the same model. With more time, I think it would be useful to examine how the removal of individual contaminant sequences affects the inferred branch length and whether there is a way to automate the 'cleaning' of sequences, as the future of the field will seek to incorporate the massive increase of information available from whole genomes. The verticality metrics we explored in chapter two may offer one potential means of evaluating the utility of new marker genes in the future. For example, after inferring a selection of orthologous genes in a given taxon selection, one could provide a sensible constraint, here we used the monophyly of Archaea and Bacteria, but the monophyly of any given group or known topological placement (i.e. mammals as sister to reptiles in a vertebrate phylogeny) word work as a set of constraints. After establishing said constraints, run a free tree search and a constrained tree search and calculate the difference in log-likelihood, only those within a decided threshold should be kept. Another round of automated cleaning could involve calculating the split score (as we did) for known taxonomic tanks (the exact rank, i.e. phylum, would be determined by the breadth of the dataset being used). Again, only those with a split score lower than a decided threshold should be retained. These steps could be useful in reducing the amount of manual curation needed.

In chapter three, we build on this and show that regardless of functional classification, a core set of markers which represents the underlying species tree infers a long archaeal-bacterial branch length. However, we also show the beneficial effects of using appropriate phylogenetic models which account for heterogeneous rate and compositions across sites. As computer power increases, so does our ability to incorporate biological knowledge: the ultimate goal would be to use a model of evolution which accounts for different compositions[163] across branches, and eventually all tips. As it stands these methods are not without limitations, but we are beginning to see the emergence of models which do account for branch-wise compositional heterogeneity[34]. Using and adapting these models for the tree of life are the next steps.

The field has moved beyond[52] the broad-scale questions of the differences between prokaryotes and eukaryotes[124], and whether or not there are two[111] or three domains of life[43] and what we should call them. Although the apparent paraphyly[54] of Archaea continues to be questioned[322]. The majority of new research in the field pertains to the precise location in the tree of life of potentially massive groups of recently discovered prokaryotic taxa, which scientists have only very recently began to culture[169,235,230,56]. Through a phylogenomic analysis using only vertically evolving marker genes, we have inferred a prokaryotic tree (Figure 50), where we help resolve the placement of some of these groups, i.e. CPR within Terrabacteria, or DPANN as basal within Archaea.

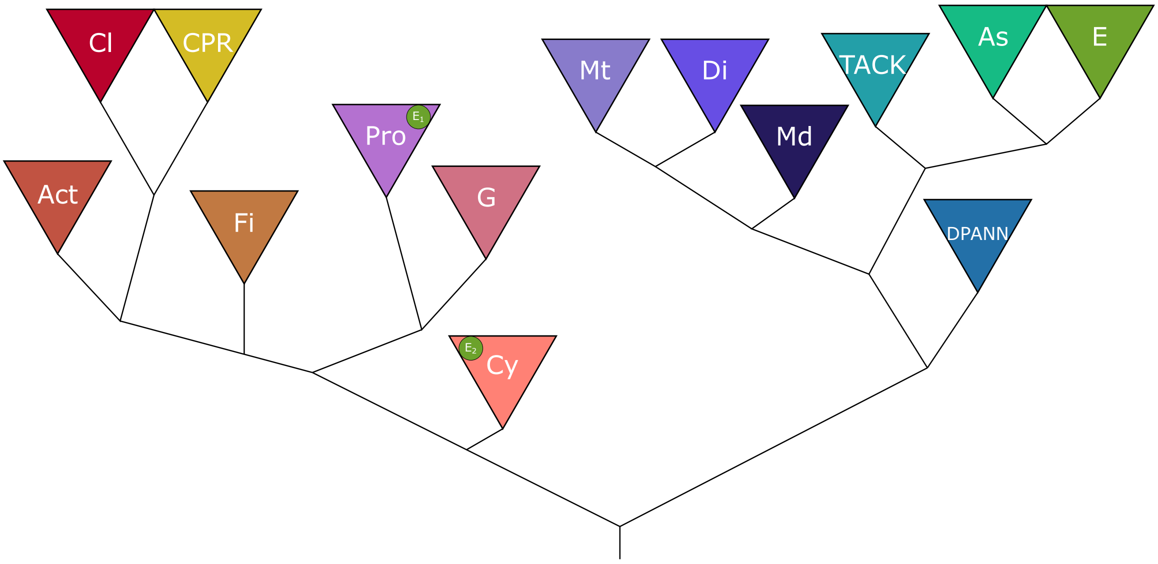

Figure 50 – A cartoon figure summarising the tree inferred from our vertically evolving markers concatenation. This figure sums up the tree inferred from the vertically evolving marker genes in chapter three. We find Terrabacteria to be paraphyletic with Cyanobacteria placed outside Actinobacteria, Chloroflexi, CPR and Firmicutes. Gracilicutes are monophyletic, and Proteobacteria are represented here to show the location of the ancestral mitochondria. Within Archaea we find DPANN as the most basal group, and we find Euryarchaeota to be paraphyletic, with Hadesarchaea and Persephonarchaea grouping closer to TACK and Asgard (not shown here). Act = Actinobacteria, Cl = Chloroflexi, CPR = Candidate Phyla Radiation, Fi = Firmicutes, Pro = Proteobacteria, G = other gracilicutes, Cy = Cyanobacteria, Mt = Methanotecta, Di = Diaforarchaea, Md = Methanomada, DPANN = Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota & Nanohaloarchaeota, TACK = Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota, As = Asgard archaea, E = Eukaryota, E1 = ancestral mitochondrion, E2 = ancestral chloroplast

However, it would be untrue to say that we have solved the prokaryotic tree of life. Our tree of life has not resolved the bacterial domain in a way in which we would expect, since previous focused analyses have shown the robustness of a monophyletic Terrabacteria containing Cyanobacteria[323,24]. Nevertheless, our tree is consistent with previous analyses[55] on the archaeal domain regarding the placement of the asgard archaea as sister to eukaryotes, a monophyletic TACK and a basal DPANN. Interestingly, our finding of a monophyletic Euryarchaeota is at odds with recent work[55,53]. These disagreements are likely the result of using such distantly related taxa, we must remember that the total time separating crown Archaea from crown Bacteria is almost 10 billion years of evolution in total.

In order to help mitigate these issues, the next step would be to increase the sampling of Bacteria with our marker selection to ensure this was not the reason for our less well resolved bacterial clade. Another future goal would be to include eukaryotes, as they are a fundamental part of the tree of life, regardless of the complications including them can bring[54] due to the differences in genome architecture, the transfer of plastids and mitochondrial sequences to nuclear DNA[324], alternative splicing and the use of multiple isoforms[325] and the donation of bacterial genes found outside of Alphaproteobacteria or Cyanobacteria[326].

Another area to expand upon would be that of divergence time estimations. Our results highlight the issues with the lack of archaeal fossils - we should endeavour to either find such things, or use alternative methods of calibrating the molecular clock[247].

The big unsolved questions relate to how we can use all gene history to infer the tree species tree[210], potentially allowing us to map gene flow across deep phylogeny, but incorporating all the data from a genome in its own meaningful way. Other big questions relate to how these ancestral organisms lived. What was their metabolism and ecology like? Are they thermophiles as some have suggested[327] or did life have cooler origins[328]? Was it more similar to a diderm bacterium[24] or a monoderm archaeon[37]? Getting closer to the true species tree helps us to begin to answer these questions.

Chapter four shows us how we can use the phylogenetic method for more recent evolutionary questions. We find that Kinesin-1 is present across the choanozoan lineage, but surprisingly lost an important motif thought to bind to the cellular cargo. We see how the evolution of ZNF648 began at the common ancestor of Osteichthyes, but the gene itself and many of the component C_2H_2 motifs were lost independently multiple times over the course of its evolutionary history. We also examine the evolution of different components of the retromer complex, and its associated proteins. We find that although SNX27 was present in the common ancestor of Filozoa, it was not between the common ancestor of Choanozoa and Bilateria where the complex dance of interactions had fully emerged.

The tree of life will continue to grow, and although its leaves may flutter in the wind, we will continue to find where they truly belong.

REFERENCES

[1] King, L.W. *Enuma Elish: The Seven Tablets of Creation: Or the Babylonian and Assyrian Legends Concerning the Creation of the World and of Mankind.* Book Tree, 1999.

[2] Preiner, M., Xavier, J.C., Vieira, A.d.N., Kleinermanns, K., Allen, J.F., and Martin, W.F. Catalysts, autocatalysis and the origin of metabolism. *Interface focus*, 9(6):20190072, 2019.

[3] Bromham, L. *An introduction to molecular evolution and phylogenetics.* Oxford University Press, 2016.

[4] Lovejoy, A.O. et al. *The Great Chain of Being. A Study of the History of an Idea.* Harvard University Press, 1936.

[5] Linnaeus, C.v. et al. Systema naturae, vol. 1. *Systema naturae, Vol. 1*, 1758.

[6] Lamarck, J.B.d.M. *Philosophie zoologique...*, volume 1. F. Savy, 1873.

[7] Darwin, C. *On the Origin of Species by Means of Natural Selection.* Murray, London, 1859.

[8] Wallace, A.R. *Alfred Russel Wallace: Letters from the Malay Archipelago.* Oxford University Press, 2013.

[9] Mendel, G. Versuche über Pflanzenhybriden. *Verh. Nat.forsch. Ver. Brünn*, 4: 3–47, 1866.

[10] Haeckel, E. *The Evolution of Man*, volume 1. CK Paul & Company, 1879.

[11] Hossfeld, U. and Levit, G.S. 'Tree of life'took root 150 years ago. *Nature*, 540 (7631):38–38, 2016.

[12] Haeckel, E. *Generelle Morphologie der Organismen. Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Descendenz-Theorie, etc*, volume 1. Verlag Georg Reimer, 1866.

[13] Porter, J. Antony van Leeuwenhoek: tercentenary of his discovery of bacteria. *Bacteriological reviews*, 40(2):260, 1976.

[14] Gibbons, N.E. and Murray, R.G.E. Proposals Concerning the Higher Taxa of Bacteria. *International Journal of Systematic and Evolutionary Microbiology*, 28(1):1–6, 1978. ISSN 1466-5026.

[15] Satyanarayana, T., Raghukumar, C., and Shivaji, S. Extremophilic microbes: Diversity and perspectives. *Current Science*, 89(1):78–90, 2005. ISSN 00113891.

[16] Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G.A.D., Gasbarrini, A., and Mele, M.C. What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms*, 7(1), 2019. ISSN 2076-2607.

[17] Caccamo, P.D. and Brun, Y.V. The molecular basis of noncanonical bacterial morphology. *Trends in microbiology*, 26(3):191–208, 2018.

[18] Liu, Y., Li, B., and Feng, X.Q. Buckling of growing bacterial chains. *Journal of the Mechanics and Physics of Solids*, 145:104146, 2020. ISSN 0022-5096.

[19] Paula, A.J., Hwang, G., and Koo, H. Dynamics of bacterial population growth in biofilms resemble spatial and structural aspects of urbanization. *Nature communications*, 11(1):1–14, 2020.

[20] BLANK, C.E. and SÁNCHEZ-BARACALDO, P. Timing of morphological and ecological innovations in the cyanobacteria – a key to understanding the rise in atmospheric oxygen. *Geobiology*, 8(1):1–23, 2010.

[21] Koch, H., Lücker, S., Albertsen, M., Kitzinger, K., Herbold, C., Spieck, E., Nielsen, P.H., Wagner, M., and Daims, H. Expanded metabolic versatility of ubiquitous nitrite-oxidizing bacteria from the genus Nitrospira. *Proceedings of the National Academy of Sciences*, 112(36):11371–11376, 2015. doi: 10.1073/pnas.1506533112.

[22] Read, T.D., Brunham, R., Shen, C., Gill, S., Heidelberg, J., White, O., Hickey, E., Peterson, J., Utterback, T., Berry, K., et al. Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. *Nucleic acids research*, 28(6):1397–1406, 2000.

[23] Taib, N., Megrian, D., Witwinowski, J., Adam, P., Poppleton, D., Borrel, G., Beloin, C., and Gribaldo, S. Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nature ecology & evolution*, 4(12):1661–1672, 2020.

[24] Coleman, G.A., Davín, A.A., Mahendrarajah, T.A., Szánthó, L.L., Spang, A., Hugenholtz, P., Szöllősi, G.J., and Williams, T.A. A rooted phylogeny resolves early bacterial evolution. *Science*, 372(6542), 2021.

[25] Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., and Banfield, J.F. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, 523(7559):208–211, 2015.

[26] Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., and Hugenholtz, P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology*, 36(10):996–1004, 2018.

[27] Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. A new view of the tree of life. *Nature microbiology*, 1(5):1–6, 2016.

[28] Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature microbiology*, 2(11):1533–1542, 2017.

[29] Castelle, C.J. and Banfield, J.F. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell*, 172(6):1181–1197, 2018.

[30] Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J.G., Belda-Ferre, P., Al-Ghalith, G.A., Kopylova, E., McDonald, D., et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nature communications*, 10(1):1–14, 2019.

[31] Cavalier-Smith, T. Rooting the tree of life by transition analyses. *Biology direct*, 1(1):1–83, 2006.

[32] Lake, J.A., Skophammer, R.G., Herbold, C.W., and Servin, J.A. Genome beginnings: rooting the tree of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1527):2177–2185, 2009.

[33] Sagan, L. On the origin of mitosing cells. *Journal of theoretical biology*, 14(3):225–IN6, 1967.

[34] Muñoz-Gómez, S.A., Susko, E., Williamson, K., Eme, L., Slamovits, C.H., Moreira, D., López-García, P., and Roger, A.J. Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nature ecology & evolution*, pages 1–10, 2022.

[35] Martijn, J., Schön, M.E., Lind, A.E., Vosseberg, J., Williams, T.A., Spang, A., and Ettema, T.J. Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nature communications*, 11(1):1–14, 2020.

[36] Fan, L., Wu, D., Goremykin, V., Xiao, J., Xu, Y., Garg, S., Zhang, C., Martin, W.F., and Zhu, R. Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. *Nature ecology & evolution*, 4(9):1213–1219, 2020.

[37] Woese, C.R. and Fox, G.E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74 (11):5088–5090, 1977.

[38] Magrum, L.J., Luehrsen, K.R., and Woese, C.R. Are extreme halophiles actually "bacteria"? *Journal of Molecular Evolution*, 11(1):1–8, 1978.

[39] Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U., et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, 541(7637):353–358, 2017.

[40] Wang, H., Bier, R., Zgleszewski, L., Peipoch, M., Omondi, E., Mukherjee, A., Chen, F., Zhang, C., and Kan, J. Distinct Distribution of Archaea From Soil to Freshwater to Estuary: Implications of Archaeal Composition and Function in Different Environments. *Frontiers in Microbiology*, 11, 2020. ISSN 1664-302X.

[41] Moissl-Eichinger, C., Pausan, M., Taffner, J., Berg, G., Bang, C., and Schmitz, R.A. Archaea are interactive components of complex microbiomes. *Trends in microbiology*, 26(1):70–85, 2018.

[42] Cavicchioli, R., Curmi, P.M., Saunders, N., and Thomas, T. Pathogenic archaea: do they exist? *Bioessays*, 25(11):1119–1128, 2003.

[43] Woese, C.R., Kandler, O., and Wheelis, M.L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, 1990.

[44] Coleman, G.A., Pancost, R.D., and Williams, T.A. Investigating the Origins of Membrane Phospholipid Biosynthesis Genes Using Outgroup-Free Rooting. *Genome Biology and Evolution*, 11(3):883–898, 02 2019. ISSN 1759-6653.

[45] Klingl, A., Pickl, C., and Flechsler, J. *Archaeal Cell Walls*, pages 471–493.

Springer International Publishing, Cham, 2019. ISBN 978-3-030-18768-2.

[46] Korkhin, Y., Unligil, U.M., Littlefield, O., Nelson, P.J., Stuart, D.I., Sigler, P.B., Bell, S.D., and Abrescia, N.G.A. Evolution of Complex RNA Polymerases: The Complete Archaeal RNA Polymerase Structure. *PLOS Biology*, 7(5):1–10, 05 2009.

[47] Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459): 431–437, 2013.

[48] Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J., Frischkorn, K.R., Tringe, S.G., Singh, A., Markillie, L.M., et al. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Current biology*, 25(6):690–701, 2015.

[49] Bird, J.T., Baker, B.J., Probst, A.J., Podar, M., and Lloyd, K.G. Culture independent genomic comparisons reveal environmental adaptations for Altiarchaeales. *Frontiers in microbiology*, 7:1221, 2016.

[50] Cavalier-Smith, T., Ema, E., and Chao, Y. Multidomain ribosomal protein trees and the planctobacterial origin of neomura (eukaryotes, archaebacteria). *Protoplasma*, 257(3):621–753, 2020.

[51] Feng, Y., Neri, U., Gosselin, S., Louyakis, A.S., Papke, R.T., Gophna, U., and Gogarten, J.P. The evolutionary origins of extreme halophilic Archaeal lineages. *Genome biology and evolution*, 13(8):evab166, 2021.

[52] Williams, T.A., Szöllősi, G.J., Spang, A., Foster, P.G., Heaps, S.E., Boussau, B., Ettema, T.J., and Embley, T.M. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences*, 114(23):E4602–E4611, 2017.

[53] Dombrowski, N., Williams, T.A., Sun, J., Woodcroft, B.J., Lee, J.H., Minh, B.Q., Rinke, C., and Spang, A. Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nature communications*, 11(1):1–15, 2020.

[54] Williams, T.A., Cox, C.J., Foster, P.G., Szöllősi, G.J., and Embley, T.M. Phylogenomics provides robust support for a two-domains tree of life. *Nature*

ecology & evolution, 4(1):138–147, 2020.

[55] Aouad, M., Flandrois, J.P., Jauffrit, F., Gouy, M., Gribaldo, S., and Brochier-Armanet, C. A divide-and-conquer phylogenomic approach based on character supermatrices resolves early steps in the evolution of the Archaea. *BMC ecology and evolution*, 22(1):1–12, 2022.

[56] Imachi, H., Nobu, M.K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M., et al. Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature*, 577(7791):519–525, 2020.

[57] Miescher, F. Letter i; to wilhelm his; tübingen, february 26th, 1869. *Die histochemischen und physiologischen arbeiten von Friedrich Miescher-aus dem wissenschaftlichen Briefwechsel von F. Miescher*, 1:33–38, 1869.

[58] Fisher, R.A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2): 399–433, 1919.

[59] Griffith, F. The significance of pneumococcal types. *Epidemiology & Infection*, 27(2):113–159, 1928.

[60] Beadle, G.W. and Tatum, E.L. Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences*, 27(11):499–506, 1941. ISSN 0027-8424.

[61] Watson, J.D. and Crick, F.H. The structure of DNA. In *Cold Spring Harbor symposia on quantitative biology*, volume 18, pages 123–131. Cold Spring Harbor Laboratory Press, 1953.

[62] Hennig, W. et al. *Grundzuge einer Theorie der phylogenetischen Systematik*. Deutscher zentralverlag, 1950.

[63] Zuckerkandl, E. and Pauling, L. Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins*, pages 97–166. Elsevier, 1965.

[64] Sela, I., Ashkenazy, H., Katoh, K., and Pupko, T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic acids research*, 43(W1):W7–W14, 2015.

[65] Sokal, R.R. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, 38:1409–1438, 1958.

[66] Saitou, N. and Nei, M. The neighbor-joining method: a new method for

reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[67] Jukes, T.H. and Cantor, C.R. Evolution of protein molecules.(Munro HN ed.) Mammalian protein Metabolism, III. *New York Academic Press*, pages 21–132, 1969.

[68] Tavaré, S. et al. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86, 1986.

[69] Dayhoff, M., Schwartz, R., and Orcutt, B. 22 a model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 5:345–352, 1978.

[70] Jones, D.T., Taylor, W.R., and Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3):275–282, 1992.

[71] Le, S.Q. and Gascuel, O. An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25(7):1307–1320, 2008.

[72] Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., and Lanfear, R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5):1530–1534, 2020.

[73] Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

[74] Huelsenbeck, J.P. and Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

[75] Sayers, E.W., Cavanaugh, M., Clark, K., Pruitt, K.D., Schoch, C.L., Sherry, S.T., and Karsch-Mizrachi, I. GenBank. *Nucleic acids research*, 49(D1):D92–D96, 2021.

[76] Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372, 1996. ISSN 0169-5347.

[77] Lartillot, N. and Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109, 2004.

[78] Williams, T.A., Schrempf, D., Szöllősi, G.J., Cox, C.J., Foster, P.G., and

Embley, T.M. Inferring the deep past from molecular data. *Genome Biology and Evolution*, 13(5):evab067, 2021.

[79] Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*, 62(4):611–615, 2013.

[80] Si Quang, L., Gascuel, O., and Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24(20):2317–2323, 2008.

[81] Susko, E. and Roger, A.J. On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution*, 24(9):2139–2150, 2007.

[82] Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., Wörheide, G., and Pisani, D. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Current Biology*, 27(24):3864–3870, 2017.

[83] Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *evolution*, 39(4):783–791, 1985.

[84] Kishino, H. and Hasegawa, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of molecular evolution*, 29(2):170–179, 1989.

[85] Shimodaira, H. and Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16: 1114–1116, 1999.

[86] Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Systematic biology*, 51(3):492–508, 2002.

[87] Akaike, H. Theory and an extension of the maximum likelihood principal. In *International symposium on information theory. Budapest, Hungary: Akademiai Kaiado*, 1973.

[88] Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[89] Hurvich, C.M. and Tsai, C.L. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[90] Nylander, J.A., Ronquist, F., Huelsenbeck, J.P., and Nieves-Aldrey, J. Bayesian phylogenetic analysis of combined data. *Systematic biology*, 53(1):47–67, 2004.

[91] Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., Von Haeseler, A., and Jermiin, L.S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, 14(6):587–589, 2017.

[92] Avery, O.T., MacLeod, C.M., and McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *Journal of experimental medicine*, 79(2):137–158, 1944.

[93] Lederberg, J. and Tatum, E.L. Gene recombination in Escherichia coli. *Nature*, 158(4016):558, 1946.

[94] Tatum, E. and Lederberg, J. Gene recombination in the bacterium Escherichia coli. *Journal of bacteriology*, 53(6):673–684, 1947.

[95] Syvanen, M. Cross-species gene transfer; implications for a new theory of evolution. *Journal of theoretical Biology*, 112(2):333–343, 1985.

[96] Hilario, E., Gogarten, J.P., et al. Horizontal transfer of ATPase genes—the tree of life becomes a net of life. *Biosystems*, 31(2-3):111–119, 1993.

[97] Gupta, R.S. and Singh, B. Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. *Current Biology*, 4(12):1104–1114, 1994.

[98] Golding, G.B. and Gupta, R.S. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Molecular Biology and Evolution*, 12(1):1–6, 1995.

[99] Whitehouse, D.B., Tomkins, J., Lovegrove, J.U., Hopkinson, D.A., and McMillan, W.O. A phylogenetic approach to the identification of phosphoglucomutase genes. *Molecular Biology and Evolution*, 15(4):456–462, 1998.

[100] Davison, J. Genetic exchange between bacteria in the environment. *Plasmid*, 42(2):73–91, 1999.

[101] Struck, T.H. The impact of paralogy on phylogenomic studies–a case study on annelid relationships. *PloS one*, 8(5):e62892, 2013.

[102] Hedtke, S.M., Townsend, T.M., and Hillis, D.M. Resolution of Phylogenetic Conflict in Large Data Sets by Increased Taxon Sampling. *Systematic Biology*, 55(3):522–529, 06 2006. ISSN 1063-5157.

[103] Heath, T.A., Hedtke, S.M., and Hillis, D.M. Taxon sampling and the accuracy

of phylogenetic analyses. *Journal of Systematics and Evolution*, 46(3):239–257, 2008.

[104] Nabhan, A.R. and Sarkar, I.N. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in bioinformatics*, 13(1):122–134, 2012.

[105] Xia, X., Xie, Z., Salemi, M., Chen, L., and Wang, Y. An index of substitution saturation and its application. *Molecular phylogenetics and evolution*, 26(1): 1–7, 2003.

[106] Ho, S.Y. and Jermiin, L.S. Tracing the decay of the historical signal in biological sequence data. *Systematic biology*, 53(4):623–637, 2004.

[107] Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic zoology*, 27(4):401–410, 1978.

[108] Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. Heterotachy and long-branch attraction in phylogenetics. *BMC evolutionary biology*, 5(1):1–8, 2005.

[109] Woese, C.R. and Fox, G.E. The concept of cellular evolution. *Journal of molecular evolution*, 10(1):1–6, 1977.

[110] Margulis, L. *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.* Yale University Press, 1970.

[111] Lake, J.A., Henderson, E., Oakes, M., and Clark, M.W. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proceedings of the National Academy of Sciences*, 81(12):3786–3790, 1984.

[112] Lake, J.A. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature*, 331(6152):184–186, 1988.

[113] Lake, J.A. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution*, 4(2):167–191, 03 1987. ISSN 0737-4038.

[114] Lockhart, P.J., Steel, M.A., Hendy, M.D., and Penny, D. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*, 11(4):605–612, 1994.

[115] Fitch, W.M. Toward Defining the Course of Evolution: Minimum Change

for a Specific Tree Topology. *Systematic Zoology*, 20(4):406–416, 1971. ISSN 00397989.

[116] Farris, J.S. Methods for Computing Wagner Trees. *Systematic Biology*, 19(1): 83–92, 03 1970. ISSN 1063-5157.

[117] Gouy, M. and Li, W.H. Phylogenetic analysis based on rRNA sequences supports the archaebacterial rather than the eocyte tree. *Nature*, 339(6220):145–147, 1989.

[118] Iwabe, N., Kuma, K.i., Hasegawa, M., Osawa, S., and Miyata, T. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proceedings of the National Academy of Sciences*, 86(23):9355–9359, 1989.

[119] Pühler, G., Leffers, H., Gropp, F., Palm, P., Klenk, H.P., Lottspeich, F., Garrett, R.A., and Zillig, W. Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. *Proceedings of the National Academy of Sciences*, 86(12):4569–4573, 1989.

[120] Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J., Bowman, B.J., Manolson, M.F., Poole, R.J., Date, T., Oshima, T., et al. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proceedings of the National Academy of Sciences*, 86(17):6661–6665, 1989.

[121] Woese, C.R. Bacterial evolution. *Microbiological reviews*, 51(2):221–271, 1987.

[122] Forterre, P., Benachenhou-Lahfa, N., Confalonieri, F., Duguet, M., Elie, C., and Labedan, B. The nature of the last universal ancestor and the root of the tree of life, still open questions. *Biosystems*, 28(1-3):15–32, 1992.

[123] Doolittle, W.F. and Brown, J.R. Tempo, mode, the progenote, and the universal root. *Proceedings of the National Academy of Sciences*, 91(15):6721–6728, 1994.

[124] Mayr, E. Two empires or three? *Proceedings of the National Academy of Sciences*, 95(17):9720–9723, 1998.

[125] Pace, N.R. Time for a change. *Nature*, 441(7091):289–289, 2006.

[126] Gupta, R.S. What are archaebacteria: life's third domain or monoderm prokaryotes related to Gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. *Molecular microbiology*, 29(3):695–707, 1998.

[127] Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and

Chothia, C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of molecular biology*, 284(4):1201–1210, 1998.

[128] Koonin, E. and Galperin, M.Y. *Sequence—evolution—function: computational approaches in comparative genomics*. Springer Science & Business Media, 2002.

[129] Theobald, D.L. A formal test of the theory of universal common ancestry. *Nature*, 465(7295):219–222, 2010.

[130] Zillig, W. Comparative biochemistry of Archaea and Bacteria. *Current Opinion in Genetics & Development*, 1(4):544–551, 1991. ISSN 0959-437X.

[131] Cavalier-Smith, T. The Origin of Eukaryote and Archaebacterial Cells. *Annals of the New York Academy of Sciences*, 503(1):17–54, 1987.

[132] Cavalier-Smith, T. A revised six-kingdom system of life. *Biological Reviews*, 73 (3):203–266, 1998.

[133] Cavalier-Smith, T. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *International journal of systematic and evolutionary microbiology*, 52(1):7–76, 2002.

[134] Ribeiro, S. and Golding, G.B. The mosaic nature of the eukaryotic nucleus. *Molecular biology and evolution*, 15(7):779–788, 1998.

[135] Adachi, J. and Hasegawa, M. Protml: Maximum likelihood inference of protein phylogeny. *Tokyo: Computer Science Monographs of the Institute of Statistical Mathematics*, 1992.

[136] Cox, C.J., Foster, P.G., Hirt, R.P., Harris, S.R., and Embley, T.M. The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences*, 105(51):20356–20361, 2008.

[137] Foster, P.G., Cox, C.J., and Embley, T.M. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1527): 2197–2207, 2009.

[138] Williams, T.A., Foster, P.G., Cox, C.J., and Embley, T.M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, 504(7479): 231–236, 2013.

[139] Doolittle, W.F. Phylogenetic classification and the universal tree. *Science*, 284

(5423):2124–2128, 1999.

[140] Martin, W. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *BioEssays*, 21(2):99–104, 1999.

[141] Woese, C.R. On the evolution of cells. *Proceedings of the National Academy of Sciences*, 99(13):8742–8747, 2002.

[142] Jain, R., Rivera, M.C., Moore, J.E., and Lake, J.A. Horizontal gene transfer accelerates genome innovation and evolution. *Molecular Biology and Evolution*, 20(10):1598–1602, 2003.

[143] Allen, B.L. and Steel, M. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of combinatorics*, 5(1):1–15, 2001.

[144] Jain, R., Rivera, M.C., and Lake, J.A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences*, 96(7):3801–3806, 1999.

[145] Abby, S.S., Tannier, E., Gouy, M., and Daubin, V. Lateral gene transfer as a support for the tree of life. *Proceedings of the National Academy of Sciences*, 109(13):4962–4967, 2012.

[146] Brochier, C., Philippe, H., and Moreira, D. The evolutionary history of ribosomal protein RpS14:: horizontal gene transfer at the heart of the ribosome. *TRENDS in Genetics*, 16(12):529–533, 2000.

[147] Rivera, M.C. and Lake, J.A. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 431(7005):152–155, 2004.

[148] Dagan, T. and Martin, W. The tree of one percent. *Genome biology*, 7(10):118, 2006.

[149] Sorek, R., Zhu, Y., Creevey, C.J., Francino, M.P., Bork, P., and Rubin, E.M. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science*, 318(5855):1449–1452, 2007.

[150] Lartillot, N., Brinkmann, H., and Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC evolutionary biology*, 7(1):1–14, 2007.

[151] Lake, J.A., Servin, J.A., Herbold, C.W., and Skophammer, R.G. Evidence for a New Root of the Tree of Life. *Systematic Biology*, 57(6):835–843, 12 2008. ISSN 1063-5157.

[152] Pisani, D., Cotton, J.A., and McInerney, J.O. Supertrees Disentangle the Chimerical Origin of Eukaryotic Genomes. *Molecular Biology and Evolution*, 24 (8):1752–1760, 05 2007.

[153] Sánchez-Baracaldo, P., Raven, J.A., Pisani, D., and Knoll, A.H. Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proceedings of the National Academy of Sciences*, 114(37):E7737–E7745, 2017.

[154] Williams, T.A., Foster, P.G., Nye, T.M., Cox, C.J., and Embley, T.M. A congruent phylogenomic signal places eukaryotes within the Archaea. *Proceedings of the Royal Society B: Biological Sciences*, 279(1749):4870–4879, 2012.

[155] Konstantinidis, K.T. and Tiedje, J.M. Towards a genome-based taxonomy for prokaryotes. *Journal of bacteriology*, 187(18):6258–6264, 2005.

[156] Zeldovich, K.B., Berezovsky, I.N., and Shakhnovich, E.I. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS computational biology*, 3(1):e5, 2007.

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

[158] Posada, D. and Buckley, T.R. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5):793–808, 2004.

[159] Kass, R.E. and Raftery, A.E. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[160] Harris, J.K., Kelley, S.T., Spiegelman, G.B., and Pace, N.R. The genetic core of the universal ancestor. *Genome research*, 13(3):407–412, 2003.

[161] Ciccarelli, F.D., Doerks, T., Von Mering, C., Creevey, C.J., Snel, B., and Bork, P. Toward automatic reconstruction of a highly resolved tree of life. *science*, 311(5765):1283–1287, 2006.

[162] Yutin, N., Makarova, K.S., Mekhedov, S.L., Wolf, Y.I., and Koonin, E.V. The deep archaeal roots of eukaryotes. *Molecular Biology and Evolution*, 25(8): 1619–1630, 2008.

[163] Foster, P.G. Modeling compositional heterogeneity. *Systematic biology*, 53(3): 485–495, 2004.

[164] Guy, L. and Ettema, T.J. The archaeal 'TACK'superphylum and the origin of eukaryotes. *Trends in microbiology*, 19(12):580–587, 2011.

[165] Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21): 2688–2690, 08 2006. ISSN 1367-4803.

[166] Beiko, R.G., Doolittle, W.F., and Charlebois, R.L. The impact of reticulate evolution on genome phylogeny. *Systematic biology*, 57(6):844–856, 2008.

[167] Lasek-Nesselquist, E. and Gogarten, J.P. The effects of model choice and mitigating bias on the ribosomal tree of life. *Molecular phylogenetics and evolution*, 69(1):17–38, 2013.

[168] Petitjean, C., Deschamps, P., López-García, P., and Moreira, D. Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome biology and evolution*, 7(1):191–204, 2014.

[169] Spang, A., Saw, J.H., Jørgensen, S.L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A.E., Van Eijk, R., Schleper, C., Guy, L., and Ettema, T.J. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521 (7551):173–179, 2015.

[170] Rosenberg, M.S. and Kumar, S. Taxon sampling, bioinformatics, and phylogenomics. *Systematic Biology*, 52(1):119, 2003.

[171] Stamatakis, A. The RAxML v8. 2. X Manual. *Heidleberg Institute for Theoretical Studies. Available at: https://cme. h-its. org/exelixis/resource/download/NewManual. pdf*, 2016.

[172] Szöllősi, G.J., Rosikiewicz, W., Boussau, B., Tannier, E., and Daubin, V. Efficient exploration of the space of reconciled gene trees. *Systematic biology*, 62(6):901–912, 2013.

[173] Segata, N., Börnigen, D., Morgan, X.C., and Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications*, 4(1):1–11, 2013.

[174] Xavier, J.C., Gerhards, R.E., Wimmer, J.L., Brueckner, J., Tria, F.D., and Martin, W.F. The metabolic network of the last bacterial common ancestor. *Communications biology*, 4(1):1–10, 2021.

[175] Tria, F.D.K., Landan, G., and Dagan, T. Phylogenetic rooting using minimal

ancestor deviation. *Nature ecology & evolution*, 1(1):1–7, 2017.

[176] Moody, E.R.R., Mahendrarajah, T.A., Dombrowski, N., Clark, J.W., Petitjean, C., Offre, P., Szöllősi, G.J., Spang, A., and Williams, T.A. An estimate of the deepest branches of the tree of life from ancient vertically-evolving genes. *eLife*, 11:e66695, 2022.

[177] Fournier, G.P. and Gogarten, J.P. Rooting the ribosomal tree of life. *Molecular Biology and Evolution*, 27(8):1792–1801, 2010.

[178] Koonin, E.V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, 1(2):127–136, 2003.

[179] Mukherjee, S., Seshadri, R., Varghese, N.J., Eloe-Fadrosh, E.A., Meier-Kolthoff, J.P., Göker, M., Coates, R.C., Hadjithomas, M., Pavlopoulos, G.A., Paez-Espino, D., et al. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature biotechnology*, 35(7):676–683, 2017.

[180] Ramulu, H.G., Groussin, M., Talla, E., Planel, R., Daubin, V., and Brochier-Armanet, C. Ribosomal proteins: toward a next generation standard for prokaryotic systematics? *Molecular phylogenetics and evolution*, 75:103–117, 2014.

[181] Raymann, K., Brochier-Armanet, C., and Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proceedings of the National Academy of Sciences*, 112(21):6670–6675, 2015.

[182] Creevey, C.J., Doerks, T., Fitzpatrick, D.A., Raes, J., and Bork, P. Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PloS one*, 6(8):e22099, 2011.

[183] Puigbò, P., Wolf, Y.I., and Koonin, E.V. Search for a'Tree of Life'in the thicket of the phylogenetic forest. *Journal of biology*, 8(6):1–17, 2009.

[184] Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., and Warnow, T. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014.

[185] Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. Phylogenomics: the beginning of incongruence? *TRENDS in Genetics*, 22(4):225–231, 2006.

[186] Wang, H.C., Minh, B.Q., Susko, E., and Roger, A.J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic

estimation. *Systematic biology*, 67(2):216–235, 2018.

[187] Gouy, R., Baurain, D., and Philippe, H. Rooting the tree of life: the phylogenetic jury is still out. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140329, 2015.

[188] Tourasse, N.J. and Gouy, M. Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced from rRNA and protein-coding genes. *Molecular phylogenetics and evolution*, 13(1):159–168, 1999.

[189] Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30 (14):2068–2069, 2014.

[190] Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, 2020.

[191] Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. The Pfam protein families database. *Nucleic acids research*, 32(suppl_1):D138–D141, 2004.

[192] Haft, D.H., Selengut, J.D., and White, O. The TIGRFAMs database of protein families. *Nucleic acids research*, 31(1):371–373, 2003.

[193] Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic acids research*, 37(suppl_1):D233–D238, 2009.

[194] Rawlings, N.D., Barrett, A.J., and Finn, R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic acids research*, 44(D1):D343–D350, 2016.

[195] Saier Jr, M.H., Tran, C.V., and Barabote, R.D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic acids research*, 34(suppl_1):D181–D186, 2006.

[196] Søndergaard, D., Pedersen, C.N., and Greening, C. HydDB: a web tool for hydrogenase classification and analysis. *Scientific reports*, 6(1):1–8, 2016.

[197] Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. InterProScan 5: genome-scale

protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014.

[198] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[199] Finn, R.D., Clements, J., and Eddy, S.R. HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl_2):W29–W37, 2011.

[200] Buchfink, B., Xie, C., and Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 12(1):59–60, 2015.

[201] Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., and Vinh, L.S. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, 35(2):518–522, 2018.

[202] Kishino, H., Miyata, T., and Hasegawa, M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, 31(2):151–160, 1990.

[203] Chaumeil, P.A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database, 2020.

[204] Team, R.C. R: A Language and Environment for Statistical Computing. 2020.

[205] Team, R.C. R: A Language and Environment for Statistical Computing. 2021.

[206] Wickham, H. *ggplot2: Elegant graphics for data analysis.* Springer, 2016.

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

[208] Kapli, P., Yang, Z., and Telford, M.J. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21(7):428–444, 2020.

[209] Kapli, P., Flouri, T., and Telford, M.J. Systematic errors in phylogenetic trees. *Current Biology*, 31(2):R59–R64, 2021.

[210] Morel, B., Schade, P., Lutteropp, S., Williams, T.A., Szöllösi, G.J., and Stamatakis, A. SpeciesRax: A tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. *bioRxiv*, 2021.

[211] Zhang, C., Scornavacca, C., Molloy, E.K., and Mirarab, S. ASTRAL-Pro:

quartet-based species-tree inference despite paralogy. *Molecular Biology and Evolution*, 37(11):3292–3307, 2020.

[212] Criscuolo, A. and Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology*, 10(1):1–21, 2010.

[213] Valas, R.E. and Bourne, P.E. The origin of a derived superkingdom: how a gram-positive bacterium crossed the desert to become an archaeon. *Biology direct*, 6(1):1–33, 2011.

[214] Schrempf, D., Lartillot, N., and Szöllősi, G. Scalable empirical mixture models that account for across-site compositional heterogeneity. *Molecular Biology and Evolution*, 37(12):3616–3631, 2020.

[215] Sugitani, K., Mimura, K., Takeuchi, M., Lepot, K., Ito, S., and Javaux, E. Early evolution of large micro-organisms with cytological complexity revealed by microanalyses of 3.4 Ga organic-walled microfossils. *Geobiology*, 13(6):507–521, 2015.

[216] Betts, H.C., Puttick, M.N., Clark, J.W., Williams, T.A., Donoghue, P.C., and Pisani, D. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature ecology & evolution*, 2(10):1556–1562, 2018.

[217] Horita, J. and Berndt, M.E. Abiogenic methane formation and isotopic fractionation under hydrothermal conditions. *Science*, 285(5430):1055–1057, 1999.

[218] Lepland, A., Arrhenius, G., and Cornell, D. Apatite in early Archean Isua supracrustal rocks, southern West Greenland: its origin, association with graphite and potential as a biomarker. *Precambrian Research*, 118(3-4):221–241, 2002.

[219] Van Zuilen, M.A., Lepland, A., and Arrhenius, G. Reassessing the evidence for the earliest traces of life. *Nature*, 418(6898):627–630, 2002.

[220] Katoh, K. and Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics*, 9(4):286–298, 2008.

[221] Katoh, K. and Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

[222] Narrowe, A.B., Spang, A., Stairs, C.W., Caceres, E.F., Baker, B.J., Miller, C.S., and Ettema, T.J. Complex evolutionary history of translation elongation factor 2 and diphthamide biosynthesis in archaea and parabasalids. *Genome biology and evolution*, 10(9):2380–2393, 2018.

[223] Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

[224] Galperin, M.Y., Kristensen, D.M., Makarova, K.S., Wolf, Y.I., and Koonin, E.V. Microbial genome analysis: the COG approach. *Briefings in bioinformatics*, 20 (4):1063–1070, 2019.

[225] Liu, Y., Makarova, K.S., Huang, W.C., Wolf, Y.I., Nikolskaya, A.N., Zhang, X., Cai, M., Zhang, C.J., Xu, W., Luo, Z., et al. Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature*, 593(7860):553–557, 2021.

[226] Werner, F. and Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nature Reviews Microbiology*, 9(2):85–98, 2011.

[227] Wang, H.C., Li, K., Susko, E., and Roger, A.J. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC evolutionary biology*, 8(1):1–13, 2008.

[228] Martinez-Gutierrez, C.A. and Aylward, F.O. Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Molecular Biology and Evolution*, 38(12):5514–5527, 2021.

[229] Salichos, L. and Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449):327–331, 2013.

[230] Méheust, R., Burstein, D., Castelle, C.J., and Banfield, J.F. The distinction of CPR bacteria from other bacteria based on protein family content. *Nature communications*, 10(1):1–12, 2019.

[231] Sorokin, D.Y., Makarova, K.S., Abbas, B., Ferrer, M., Golyshin, P.N., Galinski, E.A., Ciordia, S., Mena, M.C., Merkel, A.Y., Wolf, Y.I., et al. Discovery of extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of methanogenesis. *Nature microbiology*, 2(8):1–11, 2017.

[232] Aouad, M., Borrel, G., Brochier-Armanet, C., and Gribaldo, S. Evolutionary placement of Methanonatronarchaeia. *Nature microbiology*, 4(4):558–559, 2019.

[233] Adam, P.S., Borrel, G., Brochier-Armanet, C., and Gribaldo, S. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *The ISME journal*, 11(11):2407–2425, 2017.

[234] Aouad, M., Taib, N., Oudart, A., Lecocq, M., Gouy, M., and Brochier-Armanet, C. Extreme halophilic archaea derive from two distinct methanogen Class II lineages. *Molecular phylogenetics and evolution*, 127:46–54, 2018.

[235] Dombrowski, N., Lee, J.H., Williams, T.A., Offre, P., and Spang, A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS microbiology letters*, 366(2):fnz008, 2019.

[236] Baker, B.J., De Anda, V., Seitz, K.W., Dombrowski, N., Santoro, A.E., and Lloyd, K.G. Diversity, ecology and evolution of Archaea. *Nature microbiology*, 5(7):887–900, 2020.

[237] Beam, J.P., Becraft, E.D., Brown, J.M., Schulz, F., Jarett, J.K., Bezuidt, O., Poulton, N.J., Clark, K., Dunfield, P.F., Ravin, N.V., et al. Ancestral absence of electron transport chains in Patescibacteria and DPANN. *Frontiers in microbiology*, 11:1848, 2020.

[238] Rinke, C., Chuvochina, M., Mussig, A.J., Chaumeil, P.A., Davín, A.A., Waite, D.W., Whitman, W.B., Parks, D.H., and Hugenholtz, P. A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nature Microbiology*, pages 1–14, 2021.

[239] Wacey, D. *Early life on earth: a practical guide*, volume 31. Springer Science & Business Media, 2009.

[240] Barboni, M., Boehnke, P., Keller, B., Kohl, I.E., Schoene, B., Young, E.D., and McKeegan, K.D. Early formation of the Moon 4.51 billion years ago. *Science advances*, 3(1):e1602365, 2017.

[241] Hanan, B. and Tilton, G. 60025: Relict of primitive lunar crust? *Earth and Planetary Science Letters*, 84(1):15–21, 1987.

[242] Kleine, T., Palme, H., Mezger, K., and Halliday, A.N. Hf-W chronometry of lunar metals and the age and early differentiation of the Moon. *Science*, 310 (5754):1671–1674, 2005.

[243] Collaboration, P., Aghanim, N., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A., Barreiro, R., Bartolo, N., et al. Planck

2018 results. VI. Cosmological parameters. 2020.

[244] Shih, P.M., Hemp, J., Ward, L.M., Matzke, N.J., and Fischer, W.W. Crown group Oxyphotobacteria postdate the rise of oxygen. *Geobiology*, 15(1):19–29, 2017.

[245] Drummond, A.J., Ho, S.Y.W., Phillips, M.J., and Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS biology*, 4(5):e88, 2006.

[246] Satkoski, A.M., Beukes, N.J., Li, W., Beard, B.L., and Johnson, C.M. A redox-stratified ocean 3.2 billion years ago. *Earth and Planetary Science Letters*, 430:43–53, 2015.

[247] Davín, A.A., Tannier, E., Williams, T.A., Boussau, B., Daubin, V., and Szöllősi, G.J. Gene transfers can date the tree of life. *Nature ecology & evolution*, 2(5): 904–909, 2018.

[248] Fournier, G., Moore, K., Rangel, L., Payette, J., Momper, L., and Bosak, T. The Archean origin of oxygenic photosynthesis and extant cyanobacterial lineages. *Proceedings of the Royal Society B*, 288(1959):20210675, 2021.

[249] Szöllősi, G.J., Höhna, S., Williams, T.A., Schrempf, D., Daubin, V., and Boussau, B. Relative time constraints improve molecular dating. *Systematic Biology*, 10 2021. ISSN 1063-5157. syab084.

[250] Wolfe, J.M. and Fournier, G.P. Horizontal gene transfer constrains the timing of methanogen evolution. *Nature ecology & evolution*, 2(5):897–903, 2018.

[251] Antón, Z., Weijman, J.F., Williams, C., Moody, E.R.R., Mantell, J., Yip, Y.Y., Cross, J.A., Williams, T.A., Steiner, R.A., Crump, M.P., Woolfson, D.N., and Dodding, M.P. Molecular mechanism for kinesin-1 direct membrane recognition. *Science Advances*, 7(31):eabg6636, 2021.

[252] Ferguson, D.C., Mokim, J.H., Meinders, M., Moody, E.R., Williams, T.A., Cooke, S., Trakarnsanga, K., Daniels, D.E., Ferrer-Vicens, I., Shoemark, D., Tipgomut, C., Macinnes, K.A., Wilson, M.C., Singleton, B.K., and Frayne, J. Characterization and evolutionary origin of novel C2H2 zinc finger protein (ZNF648) required for both erythroid and megakaryocyte differentiation in humans. *Haematologica*, 106(11):2859–2873, Oct. 2020.

[253] Simonetti, B., Guo, Q., Gimenez-Andres, M., Chen, K.E., Moody, E.R., Evans, A.J., Danson, C.M., Williams, T.A., Collins, B.M., and Cullen, P.J. Mechanistic

basis for SNX27-Retromer coupling to ESCPE-1 in promoting endosomal cargo recycling. *bioRxiv*, 2021.

[254] Malik, A.J., Poole, A.M., and Allison, J.R. Structural phylogenetics with confidence. *Molecular Biology and Evolution*, 37(9):2711–2726, 2020.

[255] Tedersoo, L., Sánchez-Ramírez, S., Kõljalg, U., Bahram, M., Döring, M., Schigel, D., May, T., Ryberg, M., and Abarenkov, K. High-level classification of the Fungi and a tool for evolutionary ecological analyses. *Fungal diversity*, 90(1): 135–159, 2018.

[256] Rapacciuolo, G., Graham, C.H., Marin, J., Behm, J.E., Costa, G.C., Hedges, S.B., Helmus, M.R., Radeloff, V.C., Young, B.E., and Brooks, T.M. Species diversity as a surrogate for conservation of phylogenetic and functional diversity in terrestrial vertebrates across the Americas. *Nature ecology & evolution*, 3(1): 53–61, 2019.

[257] Hu, Y., Thapa, A., Fan, H., Ma, T., Wu, Q., Ma, S., Zhang, D., Wang, B., Li, M., Yan, L., et al. Genomic evidence for two phylogenetic species and long-term population bottlenecks in red pandas. *Science advances*, 6(9):eaax5751, 2020.

[258] Penry, G.S., Hammond, P.S., Cockcroft, V.G., Best, P.B., Thornton, M., and Graves, J.A. Phylogenetic relationships in southern African Bryde's whales inferred from mitochondrial DNA: further support for subspecies delineation between the two allopatric populations. *Conservation Genetics*, 19(6):1349–1365, 2018.

[259] Faith, D.P. Conservation evaluation and phylogenetic diversity. *Biological conservation*, 61(1):1–10, 1992.

[260] Gumbs, R., Gray, C.L., Böhm, M., Hoffmann, M., Grenyer, R., Jetz, W., Meiri, S., Roll, U., Owen, N.R., and Rosindell, J. Global priorities for conservation of reptilian phylogenetic diversity in the face of human impacts. *Nature communications*, 11(1):1–13, 2020.

[261] Jäger, G. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific data*, 5(1):1–16, 2018.

[262] Teixidor-Toneu, I., Jordan, F.M., and Hawkins, J.A. Comparative phylogenetic methods and the cultural evolution of medicinal plant use. *Nature Plants*, 4 (10):754–761, 2018.

[263] Ringen, E.J., Duda, P., and Jaeggi, A.V. The evolution of daily food sharing: A Bayesian phylogenetic analysis. *Evolution and Human Behavior*, 40(4):375–384, 2019.

[264] Lei, D., Al Jabri, T., Teixidor-Toneu, I., Saslis-Lagoudakis, C.H., Ghazanfar, S.A., and Hawkins, J.A. Comparative analysis of four medicinal floras: Phylogenetic methods to identify cross-cultural patterns. *Plants, People, Planet*, 2 (6):614–626, 2020.

[265] Basava, K., Zhang, H., and Mace, R. A phylogenetic analysis of revolution and afterlife beliefs. *Nature Human Behaviour*, 5(5):604–611, 2021.

[266] Teixidor-Toneu, I., Kool, A., Greenhill, S.J., Kjesrud, K., Sandstedt, J.J., Manzanilla, V., and Jordan, F.M. Historical, archaeological and linguistic evidence test the phylogenetic inference of Viking-Age plant use. *Philosophical Transactions of the Royal Society B*, 376(1828):20200086, 2021.

[267] Lepage, T., Bryant, D., Philippe, H., and Lartillot, N. A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution*, 24(12): 2669–2680, 2007.

[268] Benton, M.J. and Donoghue, P.C. Paleontological evidence to date the tree of life. *Molecular Biology and Evolution*, 24(1):26–53, 2007.

[269] Balfourier, F., Bouchet, S., Robert, S., De Oliveira, R., Rimbert, H., Kitt, J., Choulet, F., Consortium, I.W.G.S., Consortium, B., and Paux, E. Worldwide phylogeography and history of wheat genetic diversity. *Science advances*, 5(5): eaav0536, 2019.

[270] Avise, J.C. Phylogeography: retrospect and prospect. *Journal of biogeography*, 36(1):3–15, 2009.

[271] Ou, C.Y., Ciesielski, C.A., Myers, G., Bandea, C.I., Luo, C.C., Korber, B.T., Mullins, J.I., Schochetman, G., Berkelman, R.L., Economou, A.N., et al. Molecular epidemiology of HIV transmission in a dental practice. *Science*, 256(5060): 1165–1171, 1992.

[272] Scaduto, D.I., Brown, J.M., Haaland, W.C., Zwickl, D.J., Hillis, D.M., and Metzker, M.L. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proceedings of the National Academy of Sciences*, 107(50):21242–21247, 2010.

[273] Zhang, J., Fan, Q., Luo, M., Yao, J., Pan, X., and Li, X. Phylogenetic evidence of HIV-1 transmission linkage between two men who have sex with men. *Virology Journal*, 18(1):1–8, 2021.

[274] Allyse, M.A., Robinson, D.H., Ferber, M.J., and Sharp, R.R. Direct-to-consumer testing 2.0: emerging models of direct-to-consumer genetic testing. In *Mayo clinic proceedings*, volume 93, pages 113–120. Elsevier, 2018.

[275] Zeberg, H. and Pääbo, S. A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[276] Holmes, E.C., Goldstein, S.A., Rasmussen, A.L., Robertson, D.L., Crits-Christoph, A., Wertheim, J.O., Anthony, S.J., Barclay, W.S., Boni, M.F., Doherty, P.C., et al. The origins of SARS-CoV-2: A critical review. *Cell*, 184 (19):4848–4856, 2021.

[277] Wang, L., Didelot, X., Yang, J., Wong, G., Shi, Y., Liu, W., Gao, G.F., and Bi, Y. Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. *Nature communications*, 11(1):1–6, 2020.

[278] Jacob, J.J., Vasudevan, K., Veeraraghavan, B., Iyadurai, R., and Gunasekaran, K. Genomic evolution of severe acute respiratory syndrome Coronavirus 2 in India and vaccine impact. *Indian journal of medical microbiology*, 38(2): 210–212, 2020.

[279] Jaimes, J.A., André, N.M., Chappie, J.S., Millet, J.K., and Whittaker, G.R. Phylogenetic analysis and structural modeling of SARS-CoV-2 spike protein reveals an evolutionary distinct and proteolytically sensitive activation loop. *Journal of molecular biology*, 432(10):3309–3325, 2020.

[280] Klinman, E. and Holzbaur, E.L. Walking forward with kinesin. *Trends in neurosciences*, 41(9):555–556, 2018.

[281] Vale, R.D., Reese, T.S., and Sheetz, M.P. Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility. *Cell*, 42(1): 39–50, 1985.

[282] Vale, R.D. The Molecular Motor Toolbox for Intracellular Transport. *Cell*, 112 (4):467–480, 2003.

[283] Cai, D., Hoppe, A.D., Swanson, J.A., and Verhey, K.J. Kinesin-1 structural organization and conformational changes revealed by FRET stoichiometry in live cells. *The Journal of cell biology*, 176(1):51–63, 2007.

[284] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[285] Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

[286] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. November 2015.

[287] Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M.R., Armean, I.M., Bennett, R., Bhai, J., Billis, K., Boddu, S., et al. Ensembl 2019. *Nucleic acids research*, 47(D1):D745–D751, 2019.

[288] Paponova, S., Chetverikov, P., Pautov, A., Yakovleva, O., Zukoff, S., Vishnyakov, A., Sukhareva, S., Krylova, E., Dodueva, I., and Lutova, L. Gall mite Fragariocoptes setiger (Eriophyoidea) changes leaf developmental program and regulates gene expression in the leaf tissues of Fragaria viridis (Rosaceae). *Annals of applied biology*, 172(1):33–46, 2018.

[289] Iuchi, S. Three classes of C2H2 zinc finger proteins. *Cellular and Molecular Life Sciences CMLS*, 58(4):625–635, 2001.

[290] Klug, A. The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation. *Quarterly reviews of biophysics*, 43(1):1–21, 2010.

[291] Miller, J., McLachlan, A., and Klug, A. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *The EMBO journal*, 4 (6):1609–1614, 1985.

[292] Elrod-Erickson, M., Benson, T.E., and Pabo, C.O. High-resolution structures of variant Zif268–DNA complexes: implications for understanding zinc finger–DNA recognition. *Structure*, 6(4):451–464, 1998.

[293] Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M., et al. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature biotechnology*, 33(5):555–562, 2015.

[294] Wolfe, S.A., Nekludova, L., and Pabo, C.O. DNA recognition by Cys2His2 zinc finger proteins. *Annual review of biophysics and biomolecular structure*, 29(1): 183–212, 2000.

[295] Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, 2009.

[296] O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

[297] De Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., and Hulo, N. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic acids research*, 34(suppl_2):W362–W365, 2006.

[298] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. BLAST+: architecture and applications. *BMC bioinformatics*, 10(1):1–9, 2009.

[299] Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 2015.

[300] Thybert, D., Roller, M., Navarro, F.C., Fiddes, I., Streeter, I., Feig, C., Martin-Galvez, D., Kolmogorov, M., Janoušek, V., Akanni, W., et al. Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes. *Genome research*, 28(4):448–459, 2018.

[301] Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, 10(6):845–858, 2015.

[302] Emerson, R.O. and Thomas, J.H. Adaptive evolution in zinc finger transcription

factors. *PLoS genetics*, 5(1):e1000325, 2009.

[303] Zhou, J., Oldfield, C.J., Yan, W., Shen, B., and Dunker, A.K. Intrinsically disordered domains: Sequence → disorder → function relationships. *Protein Science*, 28(9):1652–1663, 2019.

[304] Fago, A., Rohlfing, K., Petersen, E.E., Jendroszek, A., and Burmester, T. Functional diversification of sea lamprey globins in evolution and development. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1866(2):283–291, 2018.

[305] Goodman, M., Moore, G.W., and Matsuda, G. Darwinian evolution in the genealogy of haemoglobin. *Nature*, 253(5493):603–608, 1975.

[306] Hardison, R.C. Evolution of hemoglobin and its genes. *Cold Spring Harbor perspectives in medicine*, 2(12):a011627, 2012.

[307] Pillai, A.S., Chandler, S.A., Liu, Y., Signore, A.V., Cortez-Romero, C.R., Benesch, J.L., Laganowsky, A., Storz, J.F., Hochberg, G.K., and Thornton, J.W. Origin of complexity in haemoglobin evolution. *Nature*, 581(7809):480–485, 2020.

[308] Heucken, N. and Ivanov, R. The retromer, sorting nexins and the plant endomembrane protein trafficking. *Journal of cell science*, 131(2):jcs203695, 2018.

[309] Seaman, M.N. The retromer complex–endosomal protein recycling and beyond. *Journal of cell science*, 125(20):4693–4702, 2012.

[310] Chishti, A.H., Kim, A.C., Marfatia, S.M., Lutchman, M., Hanspal, M., Jindal, H., Liu, S.C., Low, P., Rouleau, G.A., Mohandas, N., et al. The FERM domain: a unique module involved in the linkage of cytoplasmic proteins to the membrane. *Trends in biochemical sciences*, 23(8):281–282, 1998.

[311] Ryan, J.F., Pang, K., Schnitzler, C.E., Nguyen, A.D., Moreland, R.T., Simmons, D.K., Koch, B.J., Francis, W.R., Havlak, P., Program, N.C.S., et al. The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. *Science*, 342(6164):1242592, 2013.

[312] Moreland, R.T., Nguyen, A.D., Ryan, J.F., Schnitzler, C.E., Koch, B.J., Siewert, K., Wolfsberg, T.G., and Baxevanis, A.D. A customized Web portal for the genome of the ctenophore Mnemiopsis leidyi. *BMC genomics*, 15(1):1–13, 2014.

[313] Moreland, R.T., Nguyen, A.D., Ryan, J.F., and Baxevanis, A.D. The Mnemiopsis Genome Project Portal: integrating new gene expression resources and improving data visualization. *Database*, 2020, 2020.

[314] Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular evolution*, 39(3):306–314, 1994.

[315] Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics*, 139(2):993–1005, 1995.

[316] Soubrier, J., Steel, M., Lee, M.S., Der Sarkissian, C., Guindon, S., Ho, S.Y., and Cooper, A. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Molecular Biology and Evolution*, 29(11):3345–3358, 2012.

[317] Strassert, J.F., Jamy, M., Mylnikov, A.P., Tikhonenkov, D.V., and Burki, F. New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. *Molecular Biology and Evolution*, 36(4):757–765, 2019.

[318] Cavalier-Smith, T. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biology letters*, 6(3):342–345, 2010.

[319] Gallon, M., Clairfeuille, T., Steinberg, F., Mas, C., Ghai, R., Sessions, R.B., Teasdale, R.D., Collins, B.M., and Cullen, P.J. A unique PDZ domain and arrestin-like fold interaction reveals mechanistic details of endocytic recycling by SNX27-retromer. *Proceedings of the National Academy of Sciences*, 111(35): E3604–E3613, 2014.

[320] Richter, D.J., Fozouni, P., Eisen, M.B., and King, N. Gene family innovation, conservation and loss on the animal stem lineage. *elife*, 7:e34226, 2018.

[321] López-Escardó, D., Grau-Bové, X., Guillaumet-Adkins, A., Gut, M., Sieracki, M.E., and Ruiz-Trillo, I. Reconstruction of protein domain evolution using single-cell amplified genomes of uncultured choanoflagellates sheds light on the origin of animals. *Philosophical Transactions of the Royal Society B*, 374(1786): 20190088, 2019.

[322] Da Cunha, V., Gaïa, M., and Forterre, P. The expanding Asgard archaea and their elusive relationships with Eukarya. *mLife*, 1(1):3–12, 2022.

[323] Superson, A.A., Phelan, D., Dekovich, A., and Battistuzzi, F.U. Choice of

species affects phylogenetic stability of deep nodes: an empirical example in Terrabacteria. *Bioinformatics*, 35(19):3608–3616, 2019.

[324] Marcet-Houben, M. and Gabaldón, T. Acquisition of prokaryotic genes by fungal genomes. *Trends in Genetics*, 26(1):5–8, 2010. ISSN 0168-9525.

[325] Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M.J. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature reviews Molecular cell biology*, 14(3):153–165, 2013.

[326] McDonald, T.R., Dietrich, F.S., and Lutzoni, F. Multiple horizontal gene transfers of ammonium transporters/ammonia permeases from prokaryotes to eukaryotes: toward a new functional and evolutionary classification. *Molecular Biology and Evolution*, 29(1):51–60, 2012.

[327] Weiss, M.C., Preiner, M., Xavier, J.C., Zimorski, V., and Martin, W.F. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS genetics*, 14(8):e1007518, 2018.

[328] Groussin, M., Boussau, B., Charles, S., Blanquart, S., and Gouy, M. The molecular signal for the adaptation to cold temperature during early life on Earth. *Biology Letters*, 9(5):20130608, 2013.

6 APPENDIX

Supplementary data-tables S1-S6 can be downloaded from https://github.com/Neo-sage/nicescripts/blob/master/Supplementary_Tables_1_to_6.zip

ID	Monophyly	Paralogy	Gene Name	Constrained tree AU	Delta LogLikelihood
p0010	yes	no	rpoB	0.52	2.344
p0003	yes	no	rpoC	0.449	9.2112
p0348	yes	no	atpG	0.574	13.696
p0076	yes	yes	miaB	0.563	16.547
p0229	yes	no	argJ	0.605	23.827
p0110	yes	no	pyrH	0.415	29.117
p0023	yes	no	infB	0.373	36.344
p0145	yes	no	Trad_0317	0.283	39.286
p0313	yes	no	argS	0.683	44.272
p0055	yes	yes	rpoD	0.752	54.482
p0339	yes	unclear	hemA	0.169	65.749
p0123	yes	yes	murF	0.303	71.013
p0067	yes	no	rpsC	0.232	76.556
p0389	yes	no	rpsG	0.214	79.83
p0141	yes	no	proA	0.169	101
p0056	yes	yes	rimI	0.198	120.88
p0242	yes	no	rnj	0.0923	137.63
p0038	yes	yes	rplB	0.15	145.48
p0080	yes	no	gatB	0.0573	185.96
p0365	yes	unclear	thyA	0.00672	318.09
p0037	yes	yes	groL	0.000613	368.59
p0337	yes	yes	pyrD	4.08E-05	431.66
p0124	no	yes	rho	0.531	2.224
p0090	no	yes	ftsY	0.495	2.2777
p0001	no	yes	fusA_1	0.485	2.5354
p0097	no	yes	fabF	0.521	3.0047
p0182	no	yes	gltB_1	0.534	4.7164
p0284	no	no	glgC_3	0.545	8.08
p0274	no	no	zwf	0.396	15.656
p0072	no	yes	pheT	0.546	15.87
p0094	no	yes	lysS	0.556	18.044
p0052	no	yes	ftsZ	0.433	18.852
p0121	no	unclear	hemB	0.426	30.253
p0383	no	no	serC	0.332	31.112
p0327	no	unclear	rumA	0.386	34.067
p0071	no	yes	polA	0.428	34.449
p0004	no	yes	tuf	0.388	38.735
p0048	no	yes	aspS	0.414	39.336
p0162	no	no	era	0.36	39.511
p0398	no	yes	Acid_7732	0.253	42.594
p0030	no	yes	pcrA1	0.379	43.548
p0180	no	yes	uppS	0.568	54.655
p0190	no	yes	yliG	0.351	54.846
p0027	no	yes	leuS	0.373	59.296
p0346	no	yes	deoA	0.322	59.349
p0138	no	no	lon	0.757	68.512
p0060	no	no	dnaB	0.301	69.485
p0100	no	yes	Ctha_0509	0.201	69.542
p0102	no	no	argH	0.302	78.553
p0163	no	no	mutL	0.219	81.033
p0236	no	no	gpmI	0.266	84.606
p0344	no	yes	mraY	0.776	86.724
p0084	no	no	prsA	0.186	88.846
p0379	no	no	Hbal_0903	0.0849	99.169
p0047	no	yes	Kfla_0407	0.852	100.15

p0297	no	no	hemD	0.158	118.01
p0241	no	no	proB	0.0911	124.27
p0285	no	yes	Strvi_8792	0.0343	134.42
p0176	no	yes	rluD	0.797	134.62
p0187	no	yes	sucD	0.0764	139.96
p0172	no	yes	HMPREF0078_1030	0.0473	140.1
p0170	no	no	Cthe_0253	0.0786	147.36
p0079	no	no	dnaA	0.0705	161.89
p0018	no	no	pgk	0.134	163.98
p0308	no	no	hemE	0.0493	165.72
p0268	no	no	pyrB	0.0886	166.66
p0109	no	no	mutS	0.127	167.25
p0396	no	yes	glgP	0.0339	168.37
p0142	no	no	queA	0.125	168.77
p0017	no	yes	gyrB	0.0737	183.46
p0183	no	yes	fruB	0.0582	185.34
p0247	no	yes	glgB_1	0.0332	190.06
p0249	no	no	pepA	0.021	190.18
p0128	no	unclear	purQ	0.0636	197.15
p0024	no	yes	epd	0.122	207.69
p0330	no	yes	glnS	0.0559	208.05
p0046	no	no	ruvB	0.0422	212.14
p0173	no	no	ilvC	0.0366	212.73
p0098	no	yes	HMPREF0868_0825	0.0123	215.61
p0011	no	yes	alaS	0.043	223.26
p0272	no	yes	PPSIR1_09635	0.00837	223.57
p0376	no	unclear	pncC	0.095	235.06
p0016	no	yes	OSCT_2440	0.0482	235.08
p0086	no	yes	atpA	0.0138	248.86
p0068	no	no	uvrC	0.0204	257.2
p0377	no	no	pckA	0.00148	269.34
p0059	no	no	pheS	0.011	270.8
p0191	no	yes	glmU	0.0248	271
p0026	no	yes	dnaJ	0.0103	271.51
p0140	no	yes	soj_4	0.007	274.51
p0049	no	yes	obg	0.0232	274.77
p0372	no	no	metE	7.53E-06	276.83
p0198	no	yes	sucC	0.0047	284.06
p0150	no	no	smc	0.000323	285.25
p0082	no	yes	tgt	0.0338	288.78
p0169	no	yes	coaBC	0.0328	294.32
p0058	no	yes	carA	0.00599	306.94
p0233	no	yes	lysC	0.000628	323.34
p0280	no	unclear	Ccur_10940	0.00403	330.08
p0317	no	yes	dinB	0.000792	331.46
p0366	no	yes	thrC	0.000123	333.02
p0279	no	yes	glgA	0.00326	337.14
p0168	no	yes	thiC	0.0012	340.18
p0021	no	no	uvrB	0.00379	343.78
p0202	no	yes	proS	0.00596	346.81
p0373	no	yes	purK	5.75E-12	351.08
p0101	no	yes	pstB	0.00445	354.1
p0008	no	no	serS	0.00413	363.02
p0029	no	yes	MYPE3150	0.003	364.92
p0214	no	no	AXF14_00940	2.72E-62	365.32
p0288	no	no	trpD	0.00861	374.35
p0209	no	no	rpe	0.012	379.48
p0221	no	yes	Theba_1838	0.000326	393.74
p0319	no	yes	kdsA	0.000602	403.66
p0292	no	yes	aroB	0.00183	403.69
p0273	no	yes	nuoH_2	2.12E-05	405
p0300	no	yes	adk	0.00286	408.18

p0256	no	no	fhs	0.000422	426.64
p0205	no	yes	lipA	0.000385	429.35
p0314	no	no	Francci3_0256	7.90999999999999E-110	432.06
p0259	no	yes	PM8797T_05425	6.87E-05	435.51
p0326	no	yes	fba	1.51E-06	438.36
p0286	no	yes	upp	2.5E-05	440.56
p0283	no	yes	rph	0.00607	448.24
p0207	no	yes	asnS	0.00118	448.67
p0051	no	no	purA2	0.000752	453.44
p0164	no	yes	glpK	0.00048	464
p0035	no	no	ligA	0.000277	467.71
p0323	no	no	DPCES_3526	0.0037	476.62
p0091	no	yes	gatA	0.0018	479.18
p0020	no	yes	thrS	0.00126	481.08
p0009	no	yes	RESH_00663	1.17E-09	484.4
p0312	no	yes	PTO1304	1.86E-62	485.66
p0007	no	no	eno	0.00403	490.1
p0034	no	no	metK	3.07E-05	501.79
p0239	no	yes	dus	0.000166	501.82
p0248	no	no	Plabr_1363	4.46E-07	505.21
p0358	no	yes	ung	0.00018	517.14
p0208	no	no	CIY_31170	8.20999999999999E-137	518.18
p0154	no	unclear	argG	1.03E-05	518.83
p0333	no	yes	GCWU000322_00769	2.07E-07	519.75
p0085	no	no	murA	2.13E-05	522.21
p0361	no	yes	thiE	0.000117	529
p0012	no	yes	guaA	0.000366	543.17
p0271	no	yes	argB	1.69E-14	543.92
p0108	no	yes	hflX	5.39E-05	548.08
p0390	no	yes	aroA	0.000539	555.82
p0132	no	yes	metAP1b	0.00022	561.98
p0117	no	yes	purH	3.63E-05	566.55
p0188	no	no	asd-1	1.88E-08	575.26
p0069	no	no	folD	0.000163	581.2
p0028	no	no	glmS	0.000131	593.49
p0148	no	yes	csd	3.79E-05	594.41
p0193	no	no	aroC	0.000835	594.9
p0305	no	yes	ald	2.69E-67	596.98
p0015	no	yes	Mevan_0093	2.31E-05	599.18
p0384	no	yes	gnd2	0.000146	601.28
p0166	no	yes	rnr	1.41E-25	603.36
p0006	no	yes	cysS	1.4E-06	605.98
p0196	no	yes	Mvol_0256	2.57E-06	610.22
p0252	no	yes	htpG	3.03E-55	613.67
p0031	no	yes	Dtur_1796	2.08E-05	620.52
p0087	no	no	purF	9.11E-05	624.58
p0260	no	no	ribD	0.000123	626.27
p0195	no	yes	pyrC	0.000143	628.52
p0267	no	yes	RED65_04790	2.21E-06	641.73
p0368	no	yes	purK	1.61E-08	641.74
p0224	no	no	xseA	1.14E-05	644.3
p0106	no	no	hisD	2.73E-08	645
p0107	no	yes	hisA	0.000132	645.97
p0192	no	no	Hbut_0523	9.36E-09	652.76
p0013	no	no	pyrG	1.54E-06	658.99
p0328	no	yes	DSM3645_16345	1.55E-06	675.65
p0261	no	yes	kuste3185	6.58E-65	681.43
p0210	no	yes	def	0.000245	684.54
p0355	no	yes	hrcN	1.36E-59	700.29
p0014	no	yes	glyA	0.000344	702.95
p0092	no	yes	Calni_0332	2.34E-09	703.44
p0360	no	yes	wecB	3.74E-06	705.71

p0175	no	yes	Hbal_0460	1.69E-44	708.57
p0320	no	yes	Desca_1656	4.22E-07	717.4
p0130	no	yes	dxs	1.56E-07	726.94
p0039	no	yes	clpB	3.12E-05	730.31
p0126	no	yes	Aaci_1220	1.67E-08	731.71
p0153	no	yes	fabHA	2.36E-38	739.34
p0000	no	yes	oppF-valS	0.000494	747.24
p0380	no	yes	Cpin_1604	2.22E-05	749.61
p0151	no	yes	EubceDRAFT1_0090	3.43E-09	750.45
p0341	no	yes	ABC3966	2.55E-41	756.67
p0277	no	yes	agcS	1.51E-64	768.4
p0231	no	yes	Haur_2544	8.51999999999999E-71	792.08
p0316	no	no	argS	9.61E-31	794.85
p0160	no	yes	glmM	2.72E-132	816.4
p0367	no	yes	Ava_3988	1.93E-44	832.18
p0155	no	yes	BN938_1133	1.67E-36	840.69
p0129	no	yes	ribB	2.01E-05	851.6
p0357	no	yes	nuoN	7.98E-11	863.35
p0393	no	no	sod	1.21E-05	868.49
p0041	no	yes	purD	4.09E-06	871.29
p0147	no	yes	HMPREF1281_00721	5.67E-87	879.86
p0143	no	no	SRU_1858	4.18E-72	890.72
p0243	no	yes	hisRS	8.55999999999999E-81	898.01
p0250	no	yes	gcvT	2.21E-66	906.79
p0216	no	yes	gcvP	2.03E-51	907.58
p0246	no	yes	Rhom172_1875	0.0008	917.95
p0237	no	yes	I545_5322	4.76E-09	930.65
p0032	no	yes	topA	8.22E-58	934.14
p0223	no	yes	gpsA	1.63E-38	942.94
p0269	no	no	gltA	7.12999999999999E-72	944.13
p0136	no	yes	DJ66_1127	9.44E-05	964.33
p0120	no	no	B5G27_07165	2.5E-37	984.24
p0262	no	yes	STAUR_5897	2.1E-55	996.78
p0321	no	yes	Dret_0471	3.1E-54	1000.6
p0222	no	yes	Tter_1539	2.09E-23	1012.4
p0371	no	yes	HMPREF0972_00374	5.11E-05	1027.4
p0217	no	yes	ppk1	1.24E-74	1038.5
p0033	no	yes	HMPREF0889_1213	0.000239	1042.3
p0002	no	yes	ftsH	6.04E-54	1045.8
p0225	no	unclear	B739_1093	0.00053	1057.9
p0318	no	yes	RED65_14647	1.04E-10	1069.5
p0089	no	yes	ilvD	2.99E-06	1077.1
p0381	no	yes	Dtox_0978	2.26E-62	1083.8
p0199	no	yes	HMPREF9087_3025	1.82E-35	1086.7
p0157	no	yes	dapA	2.37E-05	1091
p0215	no	no	ahcY	9.16E-10	1096.1
p0077	no	yes	pox2	3.53E-35	1104.9
p0351	no	yes	cmpD_1	1.77E-20	1141.3
p0255	no	unclear	galU	9.75E-08	1161.2
p0197	no	yes	rfbA	4.33E-26	1173.7
p0219	no	yes	amt	2.36E-10	1179.2
p0114	no	yes	SAMN05421810_102418	0.000659	1193.2
p0105	no	no	trpB	3.18E-07	1196.8
p0306	no	yes	dppB	1.06E-47	1250.8
p0005	no	yes	ileS	8.33E-09	1253.8
p0211	no	yes	ilvA	1.01E-05	1266.9
p0075	no	yes	AciX9_1167	7.19E-08	1285.6
p0115	no	unclear	Mlg_1083	1.98E-35	1291.6
p0095	no	yes	Emin_0286	3.51E-09	1294
p0074	no	unclear	rhlE	4.17E-05	1309
p0343	no	yes	Nther_2414	0.00217	1314.3
p0293	no	yes	MA_2961	1.06E-71	1318.7

p0289	no	yes	pcaB_2	3.15E-55	1350.1
p0347	no	yes	csbB	2.78E-12	1360.5
p0325	no	yes	appF	6.55E-09	1387.1
p0340	no	yes	uraA	0.000313	1403.4
p0167	no	yes	TK1612	1.45E-05	1423.6
p0265	no	yes	bcd3	3.83E-58	1431.2
p0338	no	yes	Ppha_0635	7.84E-35	1437.1
p0171	no	yes	Mettu_3986	1.11E-09	1468.6
p0392	no	yes	SACT1_5091	2.1E-48	1505.1
p0135	no	yes	Swol_1850	9.60999999999999E-83	1521.2
p0073	no	yes	Hbut_0873	0.000111	1524
p0298	no	yes	FM106_13590	0.0007	1534
p0307	no	yes	BSCH_01633c	5.65E-42	1538.7
p0096	no	yes	lpdA	0.000191	1540.9
p0184	no	yes	sufB	1.82E-58	1547.4
p0226	no	unclear	MJ1054	3.28E-46	1551.6
p0201	no	yes	nuoCD	1.53E-86	1562.8
p0278	no	yes	acnA	2.56E-51	1593
p0375	no	yes	yheH_3	2.88E-06	1598.3
p0185	no	yes	tktA_1	2.62E-60	1636.3
p0227	no	yes	Vdis_1728	1.26E-54	1637.3
p0263	no	yes	lysA	1.82E-44	1640.5
p0181	no	yes	R2A130_0550	2.26E-48	1646.9
p0019	no	no	FSU_1023	3.71E-11	1652.2
p0352	no	yes	VDG1235_506	1.04E-29	1667.2
p0395	no	yes	kdpB	9.97E-15	1744.4
p0394	no	yes	hutU	1.47E-51	1750.9
p0270	no	yes	NEIELOOT_00858	5.66999999999999E-121	1768.9
p0287	no	yes	yheH	6.37E-07	1799.3
p0290	no	yes	Caur_2134	1.75E-06	1803
p0382	no	yes	osmV	4.94E-39	1835.6
p0362	no	yes	TPSD3_14250	1.96E-43	1839.9
p0131	no	yes	potA_1	1.01E-10	1863.7
p0388	no	unclear	katA	3.06E-53	1883.5
p0364	no	yes	livG2	5.28E-99	1895.8
p0152	no	yes	glnA	3.17E-55	1902.5
p0391	no	yes	bdhA	9.72999999999999E-72	1908.9
p0122	no	yes	PFL_2138	2.07E-73	1925.8
p0334	no	yes	Maeo_0301	2.01E-06	1931.4
p0332	no	yes	Spirs_1356	9.58E-64	2028.9
p0295	no	yes	bioA	7.63E-66	2050.5
p0228	no	unclear	Sfum_2516	0.000267	2154.7
p0387	no	yes	gpmA	1.02E-58	2194.5
p0350	no	yes	metN	2.85E-48	2245.8
p0244	no	yes	gdh	1.11E-33	2281.1
p0177	no	no	SY1_10840	2.52E-10	2391.7
p0093	no	yes	leuC2	6.06E-36	2397.8
p0025	no	yes	BIF_01143	1.07E-38	2397.9
p0134	no	yes	Metig_0177	8.53E-09	2400.6
p0036	no	yes	carB	2.98E-51	2427.1
p0139	no	yes	sdhA	0.00189	2547.2
p0386	no	yes	putA	7.32E-50	2917.4
p0022	no	yes	mdlB	1.58E-07	3146
p0220	no	unclear	nuoL1	6.74E-73	3273.6
p0235	no	yes	fadD	0.00116	3293.1
p0174	no	yes	HMPREF9464_01212	7.56E-05	3544.3
p0203	no	no	Dace_1181	2.43E-48	3574
p0291	no	yes	Dde_2080	7.87E-06	3863.1
p0149	no	yes	Dalk_3586	3.86E-08	4329.8
p0234	no	yes	BTH_I1153	2.3E-31	5007.2
p0159	no	yes	FRAAL4315	2.47E-55	5964.2
p0088	N/A	yes	atpD	N/A	N/A

p0397	N/A	yes	rsmI	N/A	N/A
p0042	N/A	no	secA	N/A	N/A
p0043	N/A	no	prfA	N/A	N/A
p0044	N/A	no	recA	N/A	N/A
p0057	N/A	no	der	N/A	N/A
p0061	N/A	no	secY	N/A	N/A
p0062	N/A	no	rplA	N/A	N/A
p0063	N/A	no	dnaG	N/A	N/A
p0078	N/A	no	clpX	N/A	N/A
p0081	N/A	no	rpoA	N/A	N/A
p0083	N/A	no	clpP	N/A	N/A
p0103	N/A	no	nusA	N/A	N/A
p0104	N/A	no	spoT	N/A	N/A
p0112	N/A	no	typA	N/A	N/A
p0113	N/A	no	rlmN_1	N/A	N/A
p0118	N/A	no	gidA	N/A	N/A
p0119	N/A	no	radA	N/A	N/A
p0125	N/A	no	Hoch_2472	N/A	N/A
p0127	N/A	no	recG	N/A	N/A
p0137	N/A	no	mnmE	N/A	N/A
p0144	N/A	no	priA	N/A	N/A
p0156	N/A	no	dxr	N/A	N/A
p0178	N/A	no	trmD	N/A	N/A
p0189	N/A	no	TherJR_1205	N/A	N/A
p0194	N/A	no	CfE428DRAFT_0212	N/A	N/A
p0200	N/A	no	rsmH	N/A	N/A
p0206	N/A	no	fabD	N/A	N/A
p0213	N/A	no	ispG	N/A	N/A
p0238	N/A	no	tpiA	N/A	N/A
p0258	N/A	no	tyrS	N/A	N/A
p0266	N/A	no	prfC	N/A	N/A
p0294	N/A	no	glyS	N/A	N/A
p0309	N/A	no	hslU	N/A	N/A
p0311	N/A	no	plsX	N/A	N/A
p0322	N/A	no	glyS	N/A	N/A
p0369	N/A	no	Ctha_0180	N/A	N/A
p0370	N/A	no	CSE45_4173	N/A	N/A
p0399	N/A	no	murG	N/A	N/A
p0116	N/A	unclear	nifU/mnmA	N/A	N/A
p0264	N/A	unclear	panC/cmk	N/A	N/A
p0299	N/A	unclear	murD	N/A	N/A
p0303	N/A	unclear	mutS	N/A	N/A
p0335	N/A	unclear	BURPS1106A_0496	N/A	N/A
p0345	N/A	unclear	Isop_1369	N/A	N/A
p0040	N/A	yes	TaqDRAFT_4602	N/A	N/A
p0045	N/A	yes	lepA	N/A	N/A
p0050	N/A	yes	ychF	N/A	N/A
p0054	N/A	yes	gltX	N/A	N/A
p0064	N/A	yes	mfd	N/A	N/A
p0065	N/A	yes	rpsB	N/A	N/A
p0066	N/A	yes	prfB	N/A	N/A
p0070	N/A	yes	pnp	N/A	N/A
p0099	N/A	yes	rplE	N/A	N/A
p0111	N/A	yes	ispH	N/A	N/A
p0133	N/A	yes	BACPEC_01518	N/A	N/A
p0146	N/A	yes	ybeZ	N/A	N/A
p0161	N/A	yes	ackA	N/A	N/A
p0165	N/A	yes	murC/ddl	N/A	N/A
p0179	N/A	yes	EubceDRAFT1_1969	N/A	N/A
p0186	N/A	yes	PRU_0318	N/A	N/A
p0204	N/A	yes	purL	N/A	N/A
p0230	N/A	yes	accD	N/A	N/A

p0232	N/A	yes	CfE428DRAFT_1143	N/A	N/A
p0253	N/A	yes	accD	N/A	N/A
p0254	N/A	yes	rny	N/A	N/A
p0257	N/A	yes	gspE	N/A	N/A
p0282	N/A	yes	pta	N/A	N/A
p0301	N/A	yes	clpV1_1	N/A	N/A
p0304	N/A	yes	ftsA	N/A	N/A
p0310	N/A	yes	trpS	N/A	N/A
p0315	N/A	yes	yheS	N/A	N/A
p0324	N/A	yes	Despr_1545	N/A	N/A
p0342	N/A	yes	pilT	N/A	N/A
p0354	N/A	yes	miaA2	N/A	N/A
p0356	N/A	yes	Psta_0845	N/A	N/A
p0363	N/A	yes	DSM3645_29701	N/A	N/A
p0378	N/A	yes	Haur_0879	N/A	N/A
p0385	N/A	yes	hisB/murF	N/A	N/A
p0088	N/A	yes	atpD	N/A	N/A
p0397	N/A	yes	rsmI	N/A	N/A

Table 7 – A summary table of expanded marker gene set, ranked by monophyly and Δ Log-Likelihood. Those with N/A only had one or no representatives from Archaea in the 1000 species subsample and therefore were dropped from the downstream analysis.

COG ID	COG Category	Gene	Non-Ribosomal	Core	Bacterial	Overlap
COG0001	H	HemL	NR_M179R_b	-	-	NR_M179R_b
COG0008	J	GlnS	-	R_51	-	R_51
COG0012	J	GTP1	-	-	B_AB8	B_AB8
COG0015	F	PurB	NR_M157R_b	-	-	NR_M157R_b
COG0016	J	PheS	NR_M078R_b	-	-	NR_M078R_b
COG0017	J	AsnS	-	R_30	-	R_30
COG0024	J	Map	-	R_32	-	R_32
COG0034	F	PurF	NR_M161R_b	-	-	NR_M161R_b
COG0043	H	UbiD	NR_M175R_b	-	-	NR_M175R_b
COG0046	F	PurL1	NR_M156R_b	-	-	NR_M156R_b
COG0048	J	RpsL	-	R_6	-	R_6
COG0049	J	RpsG	-	R_19	B_AB19	R_19;B_AB19
COG0050	J	TufA	-	-	B_AB5	B_AB5
COG0051	J	RpsJ	-	R_14	B_AB30	R_14;B_AB30
COG0052	J	RpsB	-	R_21	B_AB17	R_21;B_AB17
COG0055	C	AtpD	-	-	B_AB4	B_AB4
COG0064	J	GatB	NR_M064R_b	-	-	NR_M064R_b
COG0068	O	HypF	NR_MA32R_b	-	-	NR_MA32R_b
COG0072	J	PheT	-	R_25	-	R_25
COG0080	J	RplK	-	R_33	-	R_33
COG0081	J	RplA	-	R_27	B_AB21	R_27;B_AB21
COG0083	E	ThrB	NR_M142R_b	-	-	NR_M142R_b
COG0085	K	RpoB	-	R_7	B_AB1	R_7;B_AB1
COG0086	K	RpoC	-	R_18R_36	-	R_18R_36
COG0087	J	RplC	-	-	B_AB11	B_AB11
COG0090	J	RplB	-	R_29	B_AB14	R_29;B_AB14
COG0091	J	RplV	-	R_38	-	R_38
COG0092	J	RpsC	-	R_3	-	R_3
COG0093	J	RplN	-	R_12	B_AB27	R_12;B_AB27
COG0094	J	RplE	-	R_2	B_AB24	R_2;B_AB24
COG0096	J	RpsH	-	R_9	B_AB29	R_9;B_AB29
COG0097	J	RplF	-	R_43	B_AB23	R_43;B_AB23
COG0098	J	RpsE	-	R_24	B_AB15	R_24;B_AB15
COG0099	J	RpsM	-	R_23	B_AB20	R_23;B_AB20
COG0100	J	RpsK	-	R_17	B_AB26	R_17;B_AB26
COG0102	J	RplM	-	R_13	B_AB25	R_13;B_AB25
COG0103	J	RpsI	-	R_1	-	R_1
COG0124	J	HisS	NR_MA19R_b	-	-	NR_MA19R_b
COG0126	G	Pgk	NR_M165R_b	-	-	NR_M165R_b
COG0127	F	RdgB	NR_M158R_b	-	-	NR_M158R_b
COG0130	J	TruB	-	R_37	-	R_37
COG0137	E	ArgG	NR_M138R_b	-	-	NR_M138R_b
COG0149	G	TpiA	-	-	B_AB18	B_AB18
COG0150	F	PurM	NR_M151R_b	-	-	NR_M151R_b
COG0164	L	RnhB	-	R_45	-	R_45
COG0180	J	TrpS	-	R_49	-	R_49
COG0181	H	HemC	NR_M176R_b	-	-	NR_M176R_b
COG0185	J	RpsS	-	R_11	B_AB28	R_11;B_AB28
COG0186	J	RpsQ	-	R_40	-	R_40
COG0197	J	RplP	-	R_41	B_AB22	R_41;B_AB22
COG0198	J	RplX	-	R_39	-	R_39
COG0201	U	SecY	-	R_8	B_AB7	R_8;B_AB7
COG0202	K	RpoA	-	R_42	-	R_42
COG0214	H	PdxS	NR_M171R_b	-	-	NR_M171R_b
COG0255	J	RpmC	-	R_35	-	R_35
COG0258	L	ExoIX	-	-	B_AB10	B_AB10
COG0315	H	MoaC	NR_MA53R_b	-	-	NR_MA53R_b
COG0343	J	Tgt	NR_M061R_b	-	-	NR_M061R_b
COG0391	GH	CofD	NR_M208R_b	-	-	NR_M208R_b
COG0409	O	HypD	NR_MA33R_b	-	-	NR_MA33R_b
COG0442	J	ProS	NR_MA17R_b	-	-	NR_MA17R_b

COG0452	H	CoaBC	NR_M170R_b	-	-	NR_M170R_b
COG0459	O	GroEL	-	R_34	-	R_34
COG0460	E	ThrA	NR_M143R_b	-	-	NR_M143R_b
COG0468	L	RecA	-	R_28	B_AB9	R_28;B_AB9
COG0470	L	HolB	-	R_31	-	R_31
COG0480	J	FusA	-	R_4	B_AB2	R_4;B_AB2
COG0495	J	LeuS	-	R_50	-	R_50
COG0496	L	SurE	NR_MA42R_b	-	-	NR_MA42R_b
COG0504	F	PyrG	-	-	B_AB6	B_AB6
COG0522	J	RpsD	-	R_44	-	R_44
COG0525	J	ValS	-	-	B_AB3	B_AB3
COG0527	E	MetL1	NR_M146R_b	-	-	NR_M146R_b
COG0528	F	PyrH	NR_M153R_b	-	-	NR_M153R_b
COG0532	J	InfB	-	R_26	-	R_26
COG0533	J	TsaD	NR_M010R_b	R_10	-	NR_M010R_b;R_10
COG0540	F	PyrB	NR_M149R_b	-	-	NR_M149R_b
COG0541	U	Ffh	NR_M028R_b	R_5	-	NR_M028R_b;R_5
COG0552	U	FtsY	NR_MA23R_b	R_16	-	NR_MA23R_b;R_16
COG0587	L	DnaE	-	-	B_AB16	B_AB16
COG0615	M	TagD	NR_M029R_b	-	-	NR_M029R_b
COG0621	J	MiaB	NR_M071R_b	-	-	NR_M071R_b
COG0689	J	Rph	NR_M066R_b	-	-	NR_M066R_b
COG0750	OK	RseP	NR_M004R_b	R_46	-	NR_M004R_b;R_46
COG1155	C	NtpA	-	R_47	-	R_47
COG1156	C	NtpB	-	R_20	-	R_20
COG1185	J	Pnp	-	-	B_AB13	B_AB13
COG1236	J	YSH1	NR_M032R_b	-	-	NR_M032R_b
COG1245	J	Rli1	-	R_15	-	R_15
COG2255	L	RuvB	-	-	B_AB12	B_AB12
COG2262	J	HflX	NR_M195R_b	-	-	NR_M195R_b
COG2511	J	GatE	NR_M041R_b	-	-	NR_M041R_b
COG3425	I	PksG	NR_MA47R_b	-	-	NR_MA47R_b
COG5256	J	TEF1	-	R_22	-	R_22
COG5257	J	GCD11	-	R_48	-	R_48

Table 8 – A table showing the COG ID and categories for each of the genes from the non-ribosomal, core and bacterial marker sets. This table also displays the overlap between all three marker sets.

COG ID	**Gene**	**COG Category**
COG0001	HemL	H
COG0008	GlnS	J
COG0012	GTP1	J
COG0015	PurB	F
COG0016	PheS	J
COG0017	AsnS	J
COG0024	Map	J
COG0034	PurF	F
COG0043	UbiD	H
COG0046	PurL1	F
COG0048	RpsL	J
COG0049	RpsG	J
COG0050	TufA	J
COG0051	RpsJ	J
COG0052	RpsB	J
COG0055	AtpD	C
COG0064	GatB	J
COG0068	HypF	O
COG0072	PheT	J
COG0080	RplK	J
COG0081	RplA	J
COG0083	ThrB	E
COG0085	RpoB	K
COG0086	RpoC	K
COG0087	RplC	J
COG0090	RplB	J
COG0091	RplV	J
COG0092	RpsC	J
COG0093	RplN	J

COG0094	RplE	J
COG0096	RpsH	J
COG0097	RplF	J
COG0098	RpsE	J
COG0099	RpsM	J
COG0100	RpsK	J
COG0102	RplM	J
COG0103	RpsI	J
COG0124	HisS	J
COG0126	Pgk	G
COG0127	RdgB	F
COG0130	TruB	J
COG0137	ArgG	E
COG0149	TpiA	G
COG0150	PurM	F
COG0164	RnhB	L
COG0180	TrpS	J
COG0181	HemC	H
COG0185	RpsS	J
COG0186	RpsQ	J
COG0197	RplP	J
COG0198	RplX	J
COG0201	SecY	U
COG0202	RpoA	K
COG0214	PdxS	H
COG0255	RpmC	J
COG0258	ExoIX	L
COG0315	MoaC	H
COG0343	Tgt	J
COG0391	CofD	GH

COG0409	HypD	O
COG0442	ProS	J
COG0452	CoaBC	H
COG0459	GroEL	O
COG0460	ThrA	E
COG0468	RecA	L
COG0470	HolB	L
COG0480	FusA	J
COG0495	LeuS	J
COG0496	SurE	L
COG0504	PyrG	F
COG0522	RpsD	J
COG0525	ValS	J
COG0527	MetL1	E
COG0528	PyrH	F
COG0532	InfB	J
COG0533	TsaD	J
COG0540	PyrB	F
COG0541	Ffh	U
COG0552	FtsY	U
COG0587	DnaE	L
COG0615	TagD	M
COG0621	MiaB	J
COG0689	Rph	J
COG0750	RseP	OK
COG1155	NtpA	C
COG1156	NtpB	C
COG1185	Pnp	J
COG1236	YSH1	J
COG1245	Rli1	J

COG2255	RuvB	L
COG2262	HflX	J
COG2511	GatE	J
COG3425	PksG	I
COG5256	TEF1	J
COG5257	GCD11	J

Table 9 – The 95 unique markers kept after annotation and subsequent dropping of overlapping markers, with gene names and COG IDs, and COG functional classification.

COG ID	COG Category	Split score	AB length	Split score rank
COG5257	J	26.65	3.204915487	1
COG0541	U	28.65	0.63620907	2
COG0086	K	32.2	5.158469751	3
COG0552	U	32.25	0.604674859	4
COG0085	K	32.55	3.053615739	5
COG0201	U	33.95	4.818092579	6
COG0098	J	34.35	0.907386016	7
COG0532	J	34.45	2.352006521	8
COG0049	J	34.95	2.344799935	9
COG0052	J	35.05	1.958947513	10
COG0092	J	35.7	1.01593372	11
COG0081	J	36.1	1.221888596	12
COG0087	J	37.45	1.517220152	13
COG0051	J	37.5	2.008952041	14
COG0090	J	37.75	2.417253919	15
COG0533	J	38.2	1.067065203	16
COG0096	J	38.6	1.667529664	17
COG0072	J	38.65	0.299714432	18
COG0093	J	38.8	2.11788577	19
COG0480	J	38.8	2.810971325	20
COG0258	L	38.9	3.136049078	21
COG0094	J	39.4	1.592210615	22
COG0016	J	40.1	0.836448193	23
COG0064	J	40.65	0.291170918	24
COG0103	J	41	1.314334774	25
COG0202	K	41.2	1.968302227	26
COG0099	J	41.4	0.581422422	27
COG0468	L	41.65	1.544423353	28
COG0100	J	41.85	1.372515424	29

COG0008	J	42.3	1.888008928	30
COG0391	GH	42.45	3.4128605	31
COG0024	J	43	0.627301423	32
COG0197	J	43.1	2.126925952	33
COG0343	J	43.1	0.107253687	34
COG0097	J	43.95	1.070513497	35
COG0080	J	44.5	1.318345858	36
COG0185	J	44.65	0.767745145	37
COG0452	H	44.75	0.310454753	38
COG0012	J	44.8	0.881637769	39
COG0214	H	44.85	0.071167345	40
COG0522	J	45.15	2.13931248	41
COG0186	J	46	0.551488415	42
COG0091	J	46.35	1.071432756	43
COG0130	J	46.45	0.516130468	44
COG0048	J	46.7	2.135231873	45
COG0149	G	47.05	9.018230748	46
COG0164	L	49.35	1.323344483	47
COG0102	J	49.6	0.433652014	48
COG0495	J	49.7	0.067886761	49
COG0127	F	49.75	0.47519301	50
COG0315	H	50.1	0.428416894	51
COG0180	J	55.25	0.95687083	52
COG0198	J	57.8	0.770032463	53
COG0409	O	60.6	0.067672975	54

Table 10 – Summary table of the 54 marker genes ordered via within-domain split score and rank. The first 27 markers are the consensus set of vertically evolving marker genes.