12-16-2016

Short Branch Attraction, the Fundamental Bipartition in Cellular Life, and Eukaryogenesis

Amanda A. Dick PhD
University of Connecticut, amanda.dick@uconn.edu

Short Branch Attraction, the Fundamental Bipartition of Cellular Life, and Eukaryogenesis

Amanda A. Dick, PhD

University of Connecticut, 2016

Short Branch Attraction is a phenomenon that occurs when BLAST searches are used as a surrogate method for phylogenetic analysis. This results from branch length heterogeneity, but it is the short branches, not the long, that are attracting.

The root of the cellular tree of life is on the bacterial branch, meaning the Archaea and eukaryotic nucleocytoplasm form a clade. Because this split is the first in the cellular tree of life, it represents a taxonomic ranking higher than the domain, the realm. I name the clade containing the Archaea and eukaryotic nucleocytoplasm the Ibisii based on shared characteristics having to do with information processing and translation. The Bacteria are the only known members of the other realm, which I call the Bacterii.

Eukaryogenesis is the study of how the Eukarya emerged from a prokaryotic state. The beginning state of the process is represented by the relationship between Eukarya and their closest relative, the Archaea. The ending state is represented by the location of the root within the Eukarya. Neither is known. I attempt to use Eukaryal stem branch length (ESBL) to inform on the relationship between Archaea and Eukarya. The long ESBL found, shows a great deal of evolution, due to a product of time and rate of evolution. This suggests that Eukarya have not evolved from Archaea, but is inconclusive. I propose a new model for evolution, the Watershed of Life. Instead of the data representing a strictly bifurcating tree, it represents two signals: vertical descent and horizontal gene transfer, accounting for opposing signals present in our data. Both the data that has been claimed to show the Eukarya evolved from within the Archaea and

data that shows that the Eukarya did not, are consistent with only a watershed in which the Eukarya did not evolve from within the Archaea, but shared genes with the Archaea. Finally, an attempt is made to root the Eukarya using paralogs duplicated along the eukaryal stem branch, because said paralogs provide a closer outgroup. Unfortunately, the paralog datasets did not contain enough data to resolve the placement of the root.

Short Branch Attraction, the Fundamental Bipartition of Cellular Life, and Eukaryogenesis

Amanda A. Dick

B.S., University of California, Irvine 2005

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2016

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Short Branch Attraction, the Fundamental Bipartition of Cellular Life, and Eukaryogenesis

Presented by

Amanda April Dick

Major Advisor _____
J. Peter Gogarten

Associate Advisor _____
Paul Lewis

Associate Advisor _____
Kenneth Noll

University of Connecticut
2016

Acknowledgments:

I would like to acknowledge the professors whose mentorship has made this thesis possible. Foremost, is Dr. J. Peter Gogarten, who gave me the freedom to develop my own ideas and hypotheses and the support needed to carry them through. Secondly, is Dr. Paul Lewis, who gave valuable input with regards to experimental design. Additionally, the remaining members of my committee gave valuable input or provided enlightening scientific discussion on the topics of my thesis; namely Dr. Ken Noll, Dr. R. Thane Papke, Dr. David Benson, Dr. Andrew Pask, and Dr. Jonathan Klassen. Prior to entering the PhD program, I was lucky enough to be mentored by Dr. C. E. Jones, who gave me the freedom to discover which scientific areas interested me the most. Dr. Luis Villarreal was the instructor of the only writing class that ever made the least bit of sense to me. I credit Dr. Villarreal with teaching me how to write and teaching me how to teach myself. Dr. Douglas Eernisse and Dr. Brandon Gaut were the first to teach me about phylogenetics; a subject that has fascinated me since discovering what the term meant. Dr Gogarten, Dr. Villarreal, Dr. Nilay Patel, and Dr. Arthur Weis were essential to my employment during this entire learning process. Finally, Dr. Arthur Weis was the first to allow me into his laboratory and teach me how to do science and encourage my intellectual development; without him, my dreams of becoming a biologist may never have come true.

Two of my colleagues provided useful scientific discussion and aided my learning to code; namely Dr. David Williams and Timothy Harlow.

I would also like to thank my family. My father, Carl Dick, has been by my side from the beginning and was for a long time, the only logical adult I knew. My mother, Sondra Dickens, prompted my first scientific hypothesis at the age of four; a hypothesis that was proven right, when my successful experiment forced her to admit she lied. My brother, Andrew Dick, was the

only sane child I knew growing up and told me not to worry about the rampant insanity amongst the majority of our species. My brother also consistently pushed me to hurry up and finish my PhD already. My sister, Cyndi Dickens, is probably the reason I survived to adulthood, because she was the only one to seek medical treatment for the pneumonia I could not recover from. While on the subject of medical treatment, my ex-boyfriend, Tabashi Price, arranged medical treatment for a case of Bronchitis that I was too ill to seek treatment for myself. And most importantly, my daughter, Zaalayka Dick-Price, has provided the reason for me to keep going.

Table of Contents

Chapter 1: Introduction

1.1 Overview of Thesis

The foci of this thesis are the early evolution of life and the origin of eukaryotes. I cover both, the approaches and pitfalls in inferring evolution from the molecular record, and the consequences for taxonomy, i.e., the naming of groups of organisms. This thesis covers work on Short Branch Attraction, a reconstruction artifact that sometimes is found when BLAST searches are used for phylogenetic inference, recent changes in taxonomy, and Eukaryogenesis, the study of how the Eukarya originated. Research on reconstruction artifacts will be presented first, because an advanced understanding of the issues faced with bioinformatic analyses is vital to doing the complicated type of reconstructions that was undertaken in later parts of this thesis. Chapter 2 covers Short Branch Attraction (SBA), a feature of BLAST searches, caused by slowly evolving sequences.

Chapter 3 changes course, to discuss the necessary changes that must be made to taxonomy and systematics. Having appropriate terms and definitions is critical in scientific communication. Unfortunately, this aspect of science tends to lag behind the primary discoveries that support them. Names and ranks for critically important clades must be assigned.

In Chapter 4, eukaryal stem branch length is used in an attempt to answer the question of the relationship between Archaea and Eukarya, because if Eukarya evolved from within Archaea and are no more than 1.2 billion years old (Berney and Pawlowski, 2006; Cavalier-Smith, 2002; Douzery et al., 2004; Eme et al., 2014), as many researchers have suggested, then the amount of time allowed for Eukaryogenesis to occur is short and the eukaryal stem branch length should be

correspondingly short, implying a rapid radiation; a somewhat strange, but thoroughly prokaryotic, Archaea practically overnight underwent a series of changes that led to LECA (Last Eukaryal Common Ancestor), which quickly radiated into Eukarya as we know them today. But this is at odds with the invariable finding, no matter which dataset one studies, of a long eukaryal stem branch (Fournier et al., 2011; Pace et al., 1986). The long eukaryal stem branch shows that there has been a great deal of evolution, reflecting the product of time and rate of evolution, that needs to be accounted for. This suggests a monophyletic Archaea or a polytomy between Euryarchaea, Crenarchaea, and Eukarya. At the very least, Eukarya must be somewhat older than 1.2 billion years, the age of the oldest clearly Eukaryal fossils (Butterfield et al., 1985).

However, because the rate of evolution cannot be distinguished from the time of evolution when measuring stem branch length (T. H. Jukes and Cantor, 1969), my findings on the long eukaryal stem branch are inconclusive. Overall, even though more and more scientific evidence is amassed each year (Cox et al., 2008; Lasek-Nesselquist and Gogarten, 2013a; Nasir et al., 2015; Williams et al., 2013), the exact relationship between Archaea and Eukarya remains inconclusive. That is why Chapter 5 proposes that the data is not the problem, but the model of evolution is. If instead of looking at the data as representing a strictly bifurcation tree, the data is viewed as representing the two signals of the watershed of life, then the two opposing signals present in the data have been accounted for. The only problem is determining which of the signals represents vertical descent, i.e. universally present with a consistency in divergence times, and which represents horizontal gene transfer, i.e. patchily distributed with inconsistence divergence times. The Eocyte data is consistent only with the pattern seen with horizontally transferred genes in the 3 domain tree, while the 3 domain data is consistent with the vertically

inherited genes in the 3 domain trees. At present, there is no data supporting either of the signals expected if the Eukarya evolved from within the Archaea.

After examining Eukaryogenesis from the bottom up, in Chapter 6 I switch to looking at the problem from the top down and the question of where the root within the Eukarya is located. I attempt to root the Eukarya using paralogs duplicated on the Eukaryal stem branch. The problem with rooting the Eukarya is that the closest living outgroup, the Archaea, is too far away to inform on the ancestral state of the group, because of the long Eukaryal stem branch (Fournier et al., 2011). Genes that were duplicated on that stem branch provide a way around this problem, because they represent closer outgroups than the Archaea. These paralogs can then be used to root each other, and thereby the Eukarya, as was first done when the ATP synthase paralogs were used to root the tree of life (Gogarten et al., 1989).

1.2 Short Branch Attraction

In 1999, twenty-five percent of the *Thermotoga maritima* genome was thought to be horizontally transferred from Archaea (Nelson et al., 1999a), because unidirectional best BLAST hits were used as a proxy for phylogenetic analysis in Horizontal Gene Transfer detection. But years later when this analysis was repeated, using the exact same genome sequence as query, that number decreased first to 11%, then 10% (Swithers, 2012), showing that BLAST results change over time, even though the sequences being used as query are unchanged. The only change is the size of the non-redundant database, which grows exponentially over time (Kulikova et al., 2007). Dr. Gogarten hypothesized that ancient horizontal gene transfers were at fault, because transferred genes were originally finding the donor as best match, before a closely related species that also inherited the transferred gene was sequenced. But upon preliminary examination, far

more evidence of heterogeneous branch lengths was found as the cause. That led to the hypothesis that short branches cause this artifact in BLAST searches. This hypothesis was tested using simulations that showed short branches return other short branches as top scoring BLAST hits over more closely related sequences.

1.3 Molecular Data and a Natural Taxonomy

Modern taxonomy started with Linnaeus who gave each species two names and assigned them to a series of ranks (Linnaeus, 1758). But back in Linnaeus' time, Darwin's principle of evolution through natural selection was not known, so groups were not necessarily based on shared descent. The highest rank was the split between plants, animals, and minerals, because fungi, protists, Bacteria, Archaea, and virus had yet to be discovered or recognized as distinct groups. As these discoveries were made, taxonomy was adjusted to accommodate this new knowledge.

Robert Whittaker proposed the five kingdom system that was widely taught in American high schools (Whittaker, 1969) and championed by Ernst Mayr (Mayr, 1974), but although he won favor with teachers and book writers, it was Willi Hennig's principles of taxonomy (Hennig, 1965a), written while being held as a prisoner of war by the Allies, that won favor with scientists. The advantage of Hennig's system was two-fold: the phylogenetic tree was imbedded in the taxonomy, so that the tree could be recreated from the taxonomy, and precise definitions were given, allowing for clearer communication. Precision in vocabulary is one of the prerequisites of good science, because it aids communication between scientists. That is why Hennig's system triumphed, despite the disadvantages of being on the wrong side of the war and dying young.

In accordance with Hennig's principles, modern taxonomy needs a name to represent the Eukarya/Archaea clade. If recent studies finding the Eukarya branch from within the Archaea turn out to be accurate (Nasir et al., 2015; Williams et al., 2013), the Archaea will be a paraphyletic grade (Huxley, 1959a) and therefore cannot be a taxonomic domain. Also, groups linked by horizontal gene transfer are deserving of a name. Finally, there is the issue of a species definition that is applicable for Bacteria and Archaea. Chapter three tackles all of these issues.

1.4. Eukaryogenesis

1.4.1 Eukaryogenesis And What Remains Unanswered.

Eukaryogenesis is the study of how Eukarya originated, likely from a simple anucleate state to the complexity that is seen in all existing members. Although it has been suggested that life arose in a complex Eukarya-like state and evolved through reduction to produce the other two domains (Patterson and Sogin, 1994; Forterre and Philippe, 1999; Hartman and Fedorov, 2002; Kurland et al., 2006), the consensus is that the Last Universal Cellular Ancestor (LUCA) was prokaryotic in nature (Gogarten and Taiz, 1992; Delaye et al., 2005; Martin et al., 2007).

There are two important yet currently missing pieces of evidence with regards to Eukaryogenesis that prevent progress regarding the evolution of Eukarya. First, there is the bottom up approach, which asks what is the nature of their common ancestor with their closest living relative, the Archaea. Secondly, there is the top down approach, which asks what is the nature of the common ancestor within the Eukarya. If we know the answers to these two

questions, then we know where the process of Eukaryogenesis started and where it ended. Only with these two pieces of information in hand, can future researchers hope to tease apart the process that led to the revolution in cellular structure that is Eukaryogenesis.

The main hypotheses regarding Eukarya's common ancestor with Archaea fall into two categories: **Paraphyletic Archaea hypotheses**, illustrated in Figure 1, specify that the last common ancestor shared between Eukarya and Archaea was already an Archaeon. For example, an Archaeon ingested a Bacterium and developed an endosymbiosis with it to form the mitochondria, creating Eukarya. An example of this is the Eocyte hypothesis (Cox et al., 2008). **Monophyletic Archaea hypotheses**, shown in Figure 2, on the other hand, specify that the last common ancestor was not an Archaeon, but an ancestor shared by the two domains, *i.e.* the organism that ingested the Bacterium was not an Archaeon. The Archaeozoa hypothesis is an example (Cox et al., 2008).

There is a third possibility and that is that the split in question is a polytomy, with the Eukarya branching off from the Archaea at the same time that the archaeal domain diverged into its kingdoms, as shown in Figure 3. There is conflicting evidence supporting both monophyletic and paraphyletic hypotheses, leaving, in effect, a polytomy. Whether this polytomy is soft, meaning that there is a meaningful bipartition pattern, but the evidence for it has not been or cannot be identified, or whether this polytomy is hard, meaning that the Eukarya, Crenarchaea, and Euryarchaea all three diverged simultaneously, will require additional evidence to determine.

With regards to the second major question of Eukaryogenesis: where is the root within the Eukarya? The answer is truly unknown. Various research studies have come forth favoring a myriad of possibilities. The first rooting used rRNA and produced a tree where plants, animals,

and fungi were at the top of the crown of the tree, with a great many single celled Eukarya branching off first (Pace et al., 1986; Sogin, 1991; Sogin et al., 1989, 1986; Woese et al., 1990). This reinforced the idea that multicellular organisms are more evolved than single celled organisms and should therefore be higher up on the tree; a modern misconception that shows Plato's Great Chain of Being (Lovejoy, n.d.) is still impacting thought processes in science. This rooting of the tree has fallen out of favor, because it appears to be the result of Long Branch Attraction (Baldauf et al., 2000).

A popular rooting of the eukaryal tree is the unikont/bikont rooting, which splits Eukarya based on the morphology of possessing one flagellum (unikont) or two (bikont) (Derelle and Lang, 2012; Richards and Cavalier-Smith, 2005; Stechmann and Cavalier-Smith, 2002). The unikonts include Opisthokonta, animals and fungi, and Amoebozoa. The bikonts include everything else. This rooting, unlike other rootings, has two sets of physiological traits that determine which of the two super groups a given Eukaryon belongs. But this distinction does not necessarily coincide with the first split in the Eukarya; a scenario which would mean that the Last Eukaryal Common Ancestor (LECA) gave rise to one clade that went on to be composed only of members with one flagella and another clade composed only of members with two. When determining the probability that this rooting is correct, it would help to know the ancestral state of flagella in Eukarya, because many have argued that the root actually belongs *within* the bikonts (Rogozin et al., 2009) or within the unikonts (Stechmann and Cavalier-Smith, 2002). Unfortunately, Archaea, the closest living outgroup, do not have flagella of the eukaryal kind (Peabody et al., 2003) and so are neither unikonts nor bikonts and therefore unable to inform on the ancestral state.

Excavates, a weird group of single-celled Eukarya, have also been suggested to be the location of the root within the Eukarya (Arisue et al., 2005; He et al., 2014). Excavates are a highly varied group, united by a ventral feeding groove (Baldauf, 2003). Excavates commonly diverge first from molecular trees, and indeed were among the first to diverge from the rRNA tree, so heavily criticized for Long Branch Attraction (Sogin, 1991). The excavates include many quickly evolving parasites, such as Giardia (Sogin et al., 1989), that lead many to ignore them in analyses completely, in favor of finding a rooting that cannot be an artifact. The problem is exacerbated by lack of sequencing of more slowly evolving members and that monophyly of this group is unproven (Simpson, 2003). Many trees put the root not *on* the excavates, but *within* them (Cavalier-Smith and Chao, 2010), making them paraphyletic. Still others suggest multiple, polyphyletic origins for the feeding groove.

The other possibilities are not popular or common, but have all been supported by one dataset or another: rootings within or on the amoeba (Stiller et al., 1998); within or on the Archaeaplastida (Rogozin et al., 2009); on a supergroup containing the animals, fungi, amoeba, and excavates (Wideman et al., 2013). The lack of consensus is so profound that analyses attempting to determine gene presence/absence in LECA generally repeat the analyses given three different rootings, to ensure results are robust against the location of the root. But even with three rootings, there is still a very real chance that the true rooting is not among the ones used. Furthermore, the only features that are robust to all possible rootings, are ones that are almost universally present in all Eukarya, because they were present in LECA; conversely features that were not present in LECA cannot be ruled out without knowing the root. Therefore, it is of the upmost importance to know the location of the root going forward.

To evaluate a rooting of a molecular tree, it helps to be aware of certain already established supergroups within the Eukarya (Derelle and Lang, 2012). Numerous studies support the Opisthokonta clade (Huang et al., 2005), so any tree that breaks this clade up, is suffering from one of the many known issues with phylogenetic reconstruction, and would therefore not be useful in placing the root within the Eukarya. Additionally, the SAR, Stramenophiles, Alveolates, and Rhizaria, are another well-supported group (Derelle and Lang, 2012). Archaeplastida, a clade defined as a group of organisms that all had primary plastid endosymbionts, which includes plants, green algae, red algae, and glaucophytes (Bhattacharya and Medlin, 1995; Cavalier-Smith, 1982; Huang and Gogarten, 2007), should also be recovered in any tree purporting to root the Eukarya, unless this group is in fact polyphyletic (Bhattacharya and Medlin, 1995).

A few years ago, there were gaps in whole genome sequencing for a great many of these groups. Archaeplastida are known for their giant genomes, which are prohibitive to genome sequencing, so that there were only whole genome sequences available for two members of this group when this project started. There were no Rhizaria genomes available at all, because of their complex and repetitive genomes. Amoebas, with their humongous genomes, were represented by only 3 species. And, as mentioned previously, the highly divergent parasitic excavates were the first sequenced, leaving a hole in our knowledge of their non-parasitic more normal relatives. These sequencing gaps limited possible attempts at rooting the Eukarya to the minimal sequence information available, such as rRNA and ribosomal proteins data sets. But recent years have seen advancement in genome sequencing and a concerted effort to sequence

the diversity of the eukaryal tree. Now that all of these sequences are available, they will be used in this thesis when attempting to root the eukaryal tree.

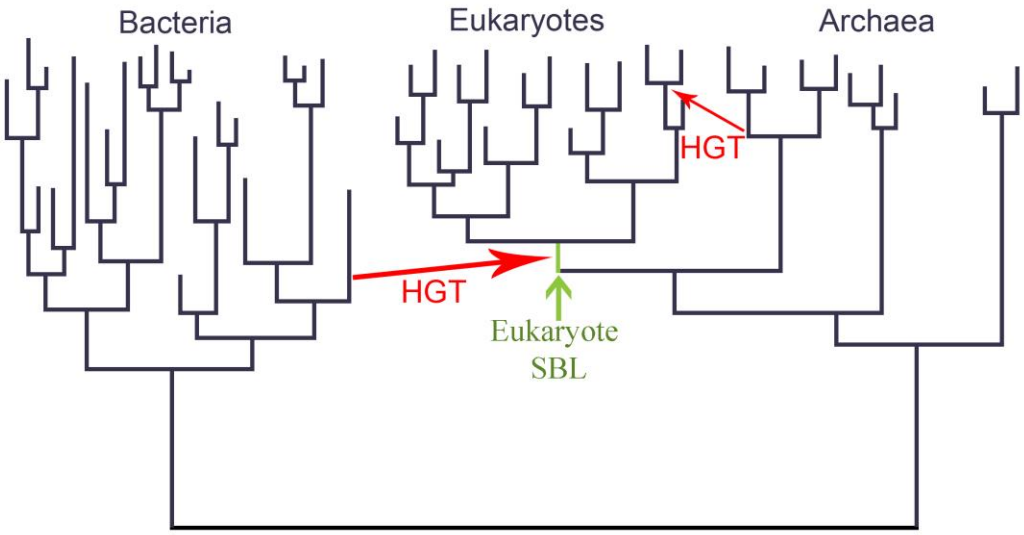


Figure 1: Paraphyletic Archaea hypotheses. In this sketch, the Eukarya evolved from within the archaeal domain (the Archaea would be paraphyletic). The tree is rooted on the bacterial branch. Red arrows indicate acquisition of an alpha proteobacterial endosymbiont (precursor of the mitochondria), and the transfer of tyrosyl RS to the ancestor of the opisthokonts from a relative of Haloarcula. The eukaryal Stem Branch Length (SBL), highlighted in green, is short.

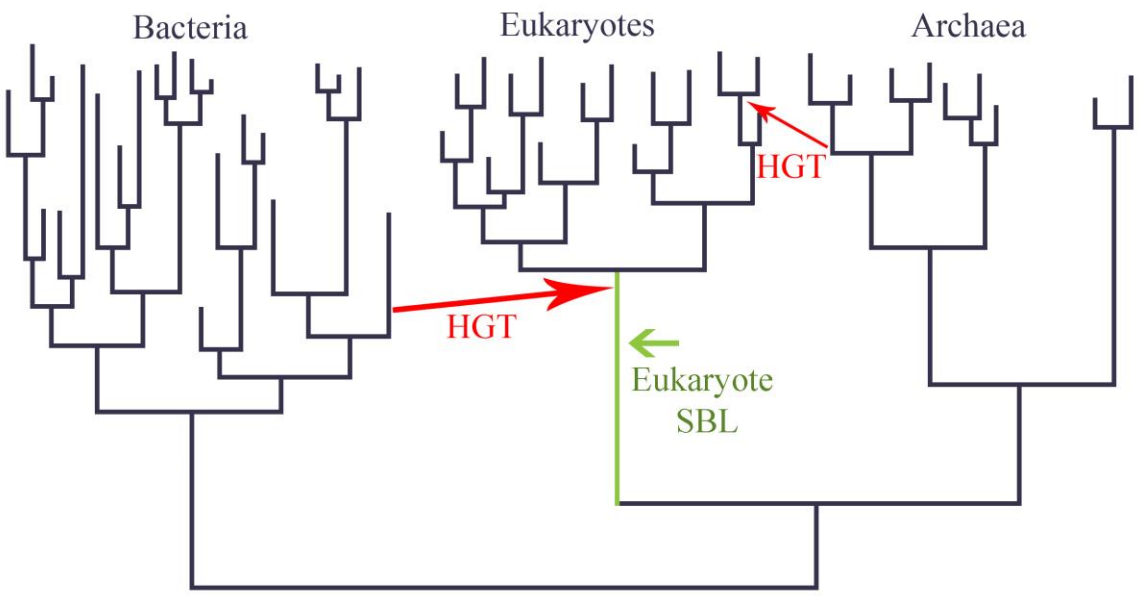


Figure 2: Monophyletic Archaea hypotheses. Eukarya and Archaea are sister-taxa. A long independent evolutionary history separates LECA (Last Eukaryal Common Ancestor) from the common ancestor with the Archaea. The eukaryal SBL, highlighted in green, is now much longer.

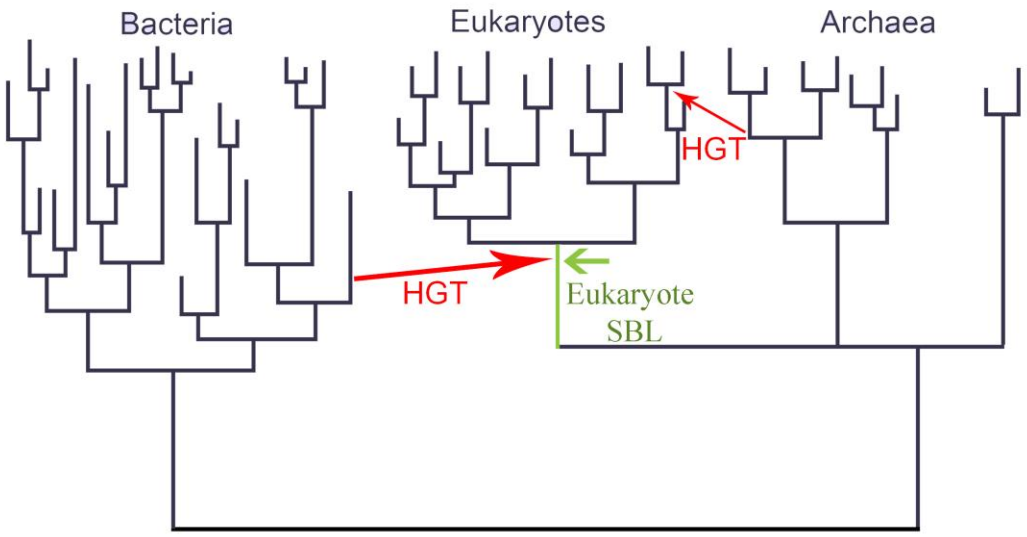


Figure 3: Archaeal and eukaryal polytomy. Eukarya and Archaea diverged during the same time period that the entire archaeal domain diverged. The Archaea and Eukarya together form a clade to the exclusion of Bacteria, but it is impossible to tell which split occurred first.

1.4.2. The Current Scientific Evidence

Now that the knowledge needed to move forward has been established, it is time to review what is already known from all available sources.

The geological record provides some information on the dates of the emergence of the three domains of life. It has been proposed that LUCA, the last common ancestor of cellular life, lived before the late heavy bombardment 3.8 bya, based on trees with the molecular clock calibrated with the fossil record (Boussau and Gouy, 2012). 3.45 billion year old stromatolites were most probably the product of microbial mat formation (Allwood et al., 2009). Cellular microfossils of sulfate-reducers, presumably bacterial in nature, were found in fossils from 3.4 billion years ago (bya) (Wacey et al., 2011). But Bacteria and Archaea are not morphologically distinct (Woese and Fox, 1977), so it is difficult to tell when the Archaea first appeared in the fossil record.

The first unambiguous sign of Archaea are 2.8 bya carbon isotopes signatures from methanogenesis ("The Proterozoic Biosphere," n.d.). Ancient eukaryal fossils are highly debated, with estimates for the oldest Eukarya as low as 0.850 bya (Cavalier-Smith, 2002) and as high as 3.2 bya (Javaux et al., 2010). The estimate of 3.2 bya comes from morphologically indistinct Acritarchs; the same authors who suggest that these fossils may be ancient Eukarya also suggest that the fossils may be large Bacteria, although there is a third possibility: these fossils may represent the ancestor of the Archaea and the Eukarya. Other fossil Acritarchs, from 2.7 bya, have been ascribed to the Eukarya (Javaux et al., 2001). Carbon isotope data suggests that Eukarya were present by 2.0 bya (Des Marais, 1997). And finally, molecular clock analyses give 1.866-1.679 bya as the best estimate for the root of LECA (Parfrey et al., 2011), which agrees

with the finding of fossilized red algae at 1.2 bya (Butterfield et al., 1985). Red algae are primary endosymbionts that engulfed cyanobacteria (Stiller and Hall, 1997), which had to occur after LECA, implying the passage of time on the eukaryal side of the tree prior to 1.2 bya.

Endosymbioses and HGTs restrict the Eukaryogenesis timeline. Eukarya have mitochondria (Roger, 1999), which group within and have evolved from alpha-proteobacteria (Roberts, 1996). Since all present-day Eukarya have evolved from a common ancestor that had mitochondria, alpha-proteobacteria diversified before Eukarya (Roberts, 1996; Roger, 1999). Some Eukarya have primary chloroplasts (Bhattacharya and Medlin, 1995; Cavalier-Smith, 1992), which group within Cyanobacteria in 16s rRNA phylogenies (Roberts, 1996), so Cyanobacteria diversified prior to the diversification of Archaeplastida. Horizontal gene transfer from Haloarcula to Opisthokonta (Huang et al., 2005) and from Chlamydia to Archaeplastida (Huang and Gogarten, 2007), indicate that Haloarcula and Chlamydia diversified before Opisthokonta and Archaeplastida, respectively.

Protein trees, such as one made from concatenating 45 conserved proteins (Cox et al., 2008), weakly support archaeal paraphyly. But trees rooted using ancient gene duplications tend to favor archaeal monophyly. The elongation factors Tu and G (Baldauf et al., 1996a), the signal recognition particle (SRP) and SRP receptor protein (SR) (Gribaldo and Cammarano, 1998), and multiple tRNA synthetases (Brown et al., 1997; Brown and Doolittle, 1995), all show support for archaeal monophyly, although support is often dependent on reconstruction algorithm used. Simple models tend to favor archaeal monophyly, while those allowing for rate heterogeneity tend to favor archaeal paraphyly (Cox et al., 2008).

1.4.3. Addressing Major Questions in the Field of Eukaryogenesis

With conflicting evidence, questions remain. Do the Eukarya have a unique history or are they simply Archaea with endosymbionts? Are the Archaea monophyletic or paraphyletic? And where is the root within the eukaryal domain?

One of the problems previous analyses have had in answering these questions is due to the extremely long branch separating the Eukarya from their closest living relatives, the Archaea (Derelle and Lang, 2012). An analogy used to describe this problem is throwing a dart at a phylogenetic tree across a football field and expecting it to find the root. Long branches create artifacts in phylogenetic analyses and long branch outgroups fail to do the very thing outgroups are used for: inform on the ancestral state. Therefore, an outgroup with a long branch is liable to attach at random to the ingroup of the tree, accounting for the seemingly random collection of rootings that are available in the literature.

A dataset with an eukaryal outgroup that is closer than the Archaea would be invaluable. If instead of throwing a dart across a football field, we could throw it across the room, the dart would have a much better chance at aiming true. The closer our dart to our dartboard, i.e. the closer the outgroup to the ingroup, the better the chances of correctly rooting the ingroup. Our ability to root the Eukarya would greatly increase. Unfortunately, there are no surviving lineages that meet this criteria and a time machine cannot be used to go back in time to obtain these sequences. But what can be used are ancient paralogs that were duplicated on the eukaryal stem branch.

Paralogs that were present in LECA and diverged from each other along the eukaryal stem branch, can be used as outgroups for each other, to break up the incredibly long eukaryal

branch. In this way, a closer outgroup for living Eukarya can be found within those same Eukarya than it is possible to find elsewhere. There is the additional benefit that many of these paralogs work in concert with each other in one macromolecular complex, so that they are more likely to share the same evolutionary history than random proteins. This minimizes the problem with concatenation that using sequences with differing histories creates trees that average the signal, representing the history of none of the individual genes. Being able to concatenate sequences has a huge advantage when it comes to recovering signal, because it adds the signal present in each individual gene and on many occasions, has led to the recovery of strong support for a split that the individual genes were not able to support (Bapteste et al., 2005).

Genes that have undergone duplication along the eukaryal stem branch and that function together in one complex will be used for these analyses. Three gene families that meet these criteria are the Tubulin, Histone, and Proteasome families. In the Tubulin family, gamma, alpha, and beta tubulin are present in all modern Eukarya (Fournier et al., 2011). In the Histone family, histone 2A, histone 2B, histone 3, and histone 4 are universally present paralogs in Eukarya that duplicated along the eukaryal stem branch (Bell and White, 2010; Thatcher and Gorovsky, 1994). Analyses done in Chapter 6 indicate that in the Proteasome family, there are 14 subunits present in LECA. It is possible that the first duplication within this family, that between the 7 alpha and the 7 beta subunits, occurred before the Eukarya split from Archaea, because there are a handful of unrelated Archaea that possess a proteasome composed of 7 identical alpha and 7 identical beta subunits (possible, because the patchy distribution of the proteasome in Archaea indicates it is also possible that these Archaea acquired this primitive proteasome through horizontal gene transfer with early stem branch Eukarya). But after the alpha subunit diverged

from the beta subunit, each went on to duplicate into 7 more subunits; duplications that undoubtedly occurred along the eukaryal stem branch. Because of the uncertainty regarding where in the tree the split between the alpha and beta subunits occurred, the desire to keep the outgroups as close to the eukaryal root as possible, and the fact that both alpha and beta types have 7 paralogous subunits to work with, they will be separated into two independent datasets.

All 4 of these datasets possess closer outgroups than the Archaea and multiple sequences with the same evolutionary history, perfect for concatenation (modified from (Hilario and Gogarten, 1993). With 4 paralogs, the histones contain 4 sequences that can be concatenated, increasing resolution power. Because the 4 subunits are homologous, 4 different concatenations can be made, placing the root of the Eukarya on the tree 4 times, giving 4 chances of finding the root. The same is true of the 3 Tubulin subunits, with 3 sequences available for concatenation and 3 chances to find the root. For both the alpha set and beta set of proteasome subunits, each has 7 sequences to concatenate and 7 chances to find the root.

But before going after the root within the Eukarya, an attempt will be made to use these datasets to evaluate the likelihood of the various relationships between Archaea and Eukarya. This will be done using stem branch length. Long branches leading to the eukaryal line and divergent sequences within Eukarya in protein phylogenies would indicate copious eukaryal evolution in isolation from the Archaea, as first proposed in Chapter 4.2. This would favor the monophyletic Archaea scenarios.

Datasets with their many paralogs contain a portion of the eukaryal stem branch multiple times, allowing for multiple measures of partial eukaryal stem branch length to constrain the time of Eukaryogenesis. These measurements will necessarily be underestimates of eukaryal

stem branch length, because the subunits were duplicated somewhere along this branch, not necessarily at the beginning of this branch. But even so, if these measurements are still found to be long, then that is all the more evidence for copious evolution on the eukaryal stem branch.

For measuring eukaryal stem branch length, only the Histone and Tubulin datasets will be used, because of the uncertainty of the evolutionary history within the Proteasome family when it comes to the Archaea. If the alpha and beta proteasome subunits present in Archaea are there as the result of horizontal gene transfer, then there in not a true archaeal stem branch for comparison.

Chapter 2: **Short branches lead to systematic artifacts when BLAST searches are used as surrogate for phylogenetic reconstruction.**

[This chapter was published in modified form in Molecular Phylogenetics and Evolution, 2016 doi: 10.1016/j.ympev.2016.11.016. [Epub ahead of print]]

2.1 Abstract

Long Branch Attraction (LBA) is a well-known artifact in phylogenetic reconstruction when dealing with branch length heterogeneity. Here we show another phenomenon, Short Branch Attraction (SBA), which occurs when BLAST searches, a 19honetic analysis, are used as a surrogate method for phylogenetic analysis. This error also results from branch length heterogeneity, but this time it is the short branches that are attracting. The SBA artifact is reciprocal and can be returned 100% of the time when multiple branches differ in length by a factor of more than two. SBA is an intended feature of BLAST searches, but becomes an issue, when top scoring BLAST hit analyses are used to infer Horizontal Gene Transfers (HGT)s, assign taxonomic category with environmental sequence data in phylotyping, or gather homologous sequences for building gene families. SBA can lead researchers to believe that there has been a HGT event when only vertical descent has occurred, cause slowly evolving taxa to be over-represented and quickly evolving taxa to be under-represented in phylotyping, or systematically exclude quickly evolving taxa from analyses. SBA also contributes to the changing results of top scoring BLAST hit analyses as the database grows, because more slowly evolving taxa, or short branches, are added over time, introducing more potential for SBA. SBA can be detected by examining reciprocal best BLAST hits among a larger group of taxa, including the known closest phylogenetic neighbors. Therefore, one should look for this

phenomenon when conducting best BLAST hit analyses as a surrogate method to identify HGTs, in phylotyping, or when using BLAST to gather homologous sequences.

2.2 Introduction

BLAST stands for Basic Local Alignment Search Tool (Altschul et al., 1997) and is freely available from the NCBI. It is currently the most frequently used bioinformatics tool, because of its speed and ease of use. The user needs to supply a query sequence and choose a database and with a few simple clicks, BLAST will search the database for matches in sequence similarity to the query. This is incredibly useful for a number of purposes. For example, if an unknown gene is sequenced and a close match is present in the database, BLAST results will indicate what type of gene it is and what species it came from. If only distant relatives are in the database, BLAST can still provide information on the function of the sequence by identifying functional domains present. This works, to a point, even if the domains are used in a novel way. BLAST is also used in identifying genes in whole genome sequencing and is used in the most effective gene calling programs.

BLAST can do all of these things because it is a tool for detecting sequence similarity. If what a researcher wants to do is detect sequence similarity, BLAST is a great program to use. A problem only arises, when researchers overreach and try to use BLAST for more than identifying sequence similarity. In recent years, BLAST has been used with increasing frequency to inform on phylogeny. This first started by using BLAST to generate datasets with appropriate orthologs from various sequenced genomes, for further analysis with phylogenetic reconstruction programs. The unspoken assumption in this case was that the later appropriate phylogenetic reconstruction would erase any bias from using BLAST to choose the sequences included in the

dataset. As sequencing ability increased, ability to analyze data lagged. In the beginning, due to the infeasibility of whole genome phylogenetic reconstruction, the easiest analysis to preform was to BLAST the entire genome. This led to BLAST being used to detect Horizontal Gene Transfer, by assuming the top hit was phylogenetically informative. As analyses needed to be done on larger and larger scales, BLAST was put to new and unintended uses.

But has the usefulness of BLAST overreached when it comes to phylogenetic inference? What can a program built to determine sequence similarity tell the user about the underlying phylogeny of the sequences? What is the link between phylogeny and sequence similarity? In this paper, we propose that that link is total branch length. Furthermore, it is the total amount of substitution between any two homologous sequences that determines the strength of a match when it comes to sequence similarity. This has implications for branch length heterogeneity interfering with phylogenetic inference by sequence similarity detection algorithms, such as BLAST. The following experiments were done with the aim of illustrating total branch length as the link between sequence similarity and phylogeny and to clarify the situations in which sequence similarity can be used as a proxy for phylogeny. To explore how BLAST results vary with respect to branch length, we use simulated sequence evolution along a family of trees depicted in Figure 1 Panel H.

Many studies have screened for horizontally transferred genes based on top scoring BLAST hit analyses (McNulty et al., 2012; Nelson et al., 1999b; Ruepp et al., 2000; Worning et al., 2000; Zhaxybayeva et al., 2009b). When unidirectional analyses were shown to be phylogenetically uninformative (Koski and Golding, 2001), a switch was made to the more restrictive reciprocal top scoring BLAST hit approach (e.g., (Dagan et al., 2010)). Reciprocal

Top Scoring BLAST hits are also used in gene family reconstruction to determine homologous gene families (Zhaxybayeva et al., 2006). In addition, BLAST is used in phylotyping or ribotyping analyses where environmental sequences are assigned a taxonomic affiliation based on top scoring BLAST hit (Basaran et al., 2001; Maloy et al., 2009). In doing these types of analyses one assumes that the top scoring BLAST hit will be the sister taxon, especially if the hit is reciprocal. BLAST was developed to find regions of homology and pick out the most similar sequence from a database, not to inform on phylogeny (Altschul et al., 1990). It is well established that the top scoring BLAST hit does not necessarily represent the nearest phylogenetic neighbor (Eisen, 2000; Koski and Golding, 2001) and this is indeed expected, because the nearest neighbor is not always the most similar. However, the question of how useful BLAST searches might be in matters of phylogeny is still left unanswered. Specifically branch-length space has never been mapped with respect to the ability of BLAST to recapitulate phylogeny.

BLAST searches work by comparing two sequences at a time and looking for local alignments using the Maximal Segment Pair (MSP) score (Altschul et al., 1990). In a gapped BLAST search (Altschul et al., 1997), the current standard for BLAST, it first matches identical residues between two sequences and stores this information. Continuous regions with high scoring segments are identified, using a threshold score as a cutoff to speed up the process, and are then extended, using a similarity criterion. The highest scoring regions are joined using dynamic programming. The best match made in the pairwise alignments is reported as the top scoring BLAST hit and is a factor of the length and identity of the match. Longer matches are considered good measures of homology and these matches can include indels and small regions of mismatch. Fewer mismatched positions result in a higher percent identity and a stronger

match, as measured by E-value, which gives the expected number of matches of a given quality due to chance alone. It has been shown that the phylogenetic nearest neighbor does not always produce the top scoring BLAST hit (Koski and Golding, 2001), but given the pairwise nature of the BLAST algorithm, might it be that the taxon with the shortest total evolutionary distance will be the top scoring hit?

With a normal BLAST search (blastn or blastp), an entire database is searched, but the other sequences in the database do not influence the results of an individual search, except that E-value is proportional to the size of the databank (Altschul et al., 1990). This differs from phylogenetic reconstruction programs, in which other sequences are used to give weight to conserved residues and calculate the evolutionary model when calculating the tree (Guindon and Gascuel, 2003; Ronquist and Huelsenbeck, 2003). Inclusion of an outgroup can allow for rooting the tree and identification of synapomorphies and symplesiomorphies. Without knowledge of the outgroup, it is impossible to distinguish synapomorphies from symplesiomorphies and the two categories of matches are therefore given equal weight in the BLAST search. The importance in using synapomorphies to the exclusion of symplesiomorphies in phylogenetic reconstruction traces back to Willi Hennig, who first expounded on their importance in forming monophyletic groups (Hennig, 1965b, 1975a). If monophyletic groups are not used, then there is increased ambiguity in detection of HGT. With phenetic programs, such as BLAST, where unrooted trees are used, there is no way of differentiating a synapomorphy from a symplesiomorphy. The remaining question is at which sets of branch lengths will this cause a problem?

If time that passed is proportional to branch lengths, then we expect that the length of interior branches should ensure that the shortest total branch length is between sister taxa, as

shown in Figure 1 panel A. If some branch lengths are shorter, the shortest total branch length may be between the two short branches. For the example tree given in Fig 1 B, the condition for non-sister taxa to be connected by the shortest overall branch is given by Equation 1.

Equation 1: $Y + X > Z + 2X$

Where Y is the length of the invariable branches (0.1 substitutions per site in Figure 1 panels A, B, and H), Z is the length of the central branch, and X is the length of the variable branches. Equation 1 reduces to Equation 2. Because the relationship between sequence similarity and phylogenetic reconstruction using BLAST is dependent on total branch lengths, these equations mean that Short Branch Attraction is likely to occur, if the combined lengths of one of the short branches and the central branch are shorter than the length of the invariable branch, as seen in Figure 1 panel B.

Equation 2: $Y > Z + X$

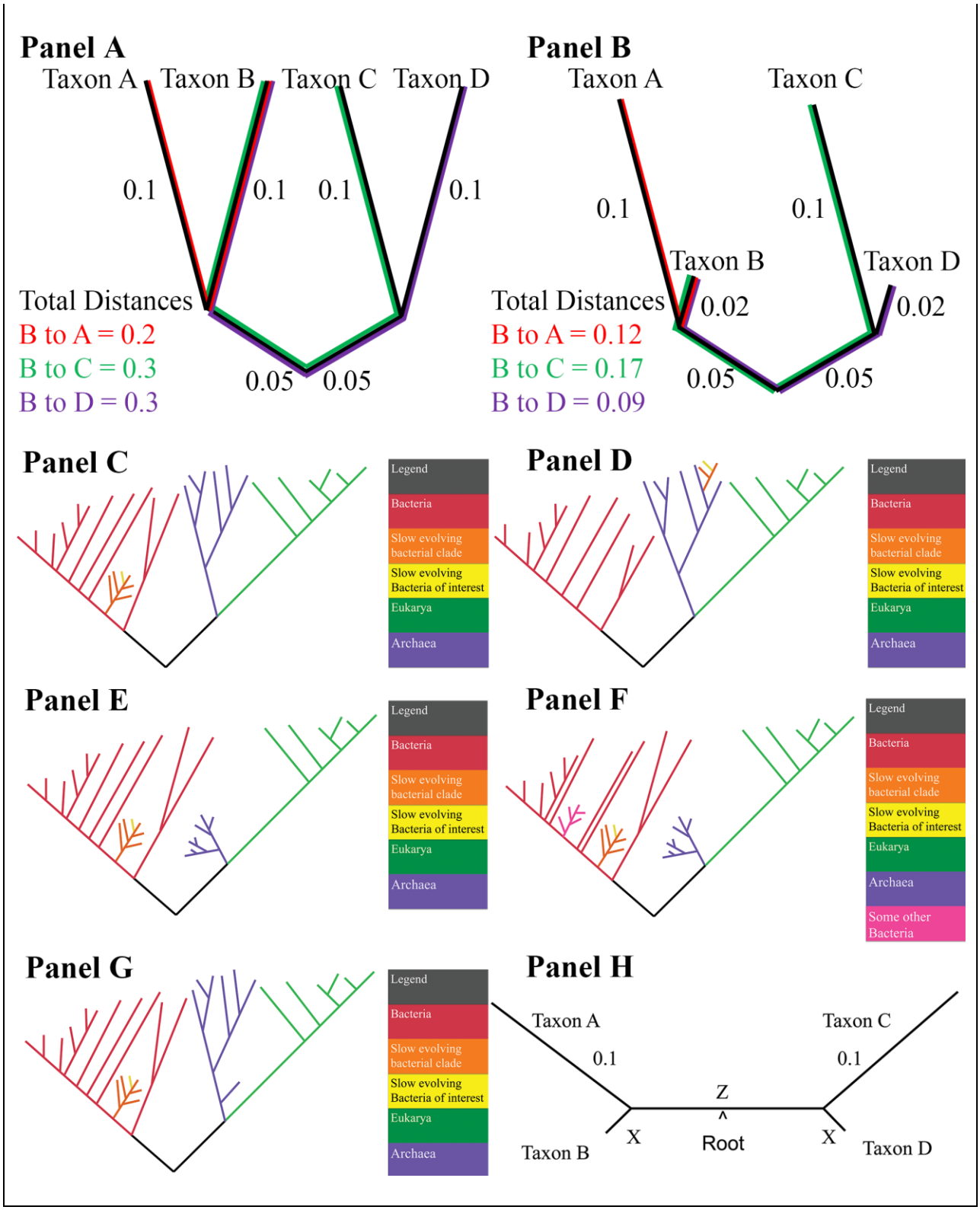
Short Branch Attraction differs from the well-known Long Branch Attraction (LBA), because Long Branch Attraction is a phenomenon of phylogenetic reconstruction. Short Branch Attraction is a phenomenon of phenetic reconstruction. In cases of Short Branch Attraction, it is the lack of evolution between slowly evolving taxa that causes the artifact. Conversely, LBA has been demonstrated in phylogenetic reconstruction and is caused by numerous convergent apomorphies resulting in homoplasies and the attraction of long branches, while the short branches behave normally (Felsenstein, 1978; Huelsenbeck and Hillis, 1993; Swofford et al., 2001).

As the number of taxa included in the database grows, and the phylogenetic relationship between them increases in complexity, the behavior of the BLAST search becomes more difficult to predict. If only one taxon is on a short branch, as shown in Figure 1 panel C, then the

shortest pairwise distances are still between sister taxa. In these cases, top scoring BLAST hit analyses will be phylogenetically informative and should minimally return a top scoring BLAST hit to a taxon within the same domain as the query. Then top scoring BLAST hits to another domain really would represent Horizontal Gene Transfers (HGTs), or even ancestral HGTs, as shown in figure 1 panel D. In the case of ancestral HGTs, a closely related group of organisms share, through vertical inheritance, a HGT. BLAST would then return the other members of this clade before returning members of the domain from where the transferred gene originated. It causes a problem when a second clade is evolving as slowly as the one of interest, as shown in Figure 1 panel E with the Archaea on short branches. Here the top scoring BLAST hit will return the Archaea when given a slowly evolving bacterial query, even in cases where there is no HGT. In this tree, the red clade on the left is still the sister clade, but the shortest total branch lengths are to the Archaea, thus the Archaea will be returned as the top scoring BLAST hit.

Short Branch Attraction explains some instances of why BLAST searches return different results, in terms of HGTs, as the non-redundant database grows. The apparent HGTs observed from the short bacterial branch attracting to the short archaeal domain, is no longer present when other slowly evolving Bacteria are sequenced, as shown in Figure 1 panel F. This is because the shortest total branch lengths are now between the two slowly evolving Bacteria, creating a new stronger Short Branch Attraction. This phenomenon can also work the other way around, if before there were no slowly evolving taxa from other domains sequenced and then one is sequenced and added to the database, as shown in Figure 1 panel G. Here Short Branch Attraction is created when before there was none.

Figure 1: Relating branch lengths in phylogenetic trees to expected BLAST results. Panel A shows a homogeneous tree with the shortest total pairwise distances between sister taxa. The top scoring BLAST hit is expected to be the sister taxon. Panel B shows a heterogeneous tree where the shortest total pairwise distance is between the two slowly evolving taxa. BLAST is expected to return Short Branch Attraction in this case. Panel C shows a hypothetical three domain tree with only one slowly evolving clade on short branches, where the shortest pairwise distances are still between sister taxa. The top scoring hit is not expected to be out of domain. In cases where the out of domain top hit has branches of normal length, as shown in Panel D, we can conclude that there really was a horizontal gene transfer event. But if the Archaea are all on short branches, as in Panel E, then the top scoring BLAST hit will return them even in cases where there is no HGT. Panel F shows how Short Branch Attraction might be the cause of discrepancies in top scoring BLAST hit analyses when the database grows, as more slowly evolving sequences are added, creating new, closer, Short Branch Attractions. Panel G shows how adding new slowly evolving species of Archaea to the database could introduce Short Branch Attraction when before there was none. Panel H shows the true tree used for the simulations. The root is indicated by the ^ and is at the midpoint along the central branch. Taxa B and D are on branches of length X. The length of the central branch is represented by Z. clade before returning members of the domain from where the transferred gene originated. Problems arise when a second clade is evolving as slowly as the one of interest, as shown in Figure 1 panel E with the Archaea on short branches. Here the top scoring BLAST hit will return the Archaea when given a slowly evolving bacterial query, even in cases where there is no HGT. In this tree, the red clade on the left is still the sister clade, but the shortest total branch lengths are to the Archaea, thus the Archaea will be returned as the top scoring BLAST hit.


Panel A
Taxon A
Taxon B
Taxon C
Taxon D
0.1
0.1
0.1
0.1
0.05
0.05
Total Distances
B to A = 0.2
B to C = 0.3
B to D = 0.3
Panel B
Taxon A
Taxon C
0.1
0.1
Taxon B
Taxon D
0.02
0.02
0.05
0.05
Total Distances
B to A = 0.12
B to C = 0.17
B to D = 0.09
Panel C
Panel D
Panel E
Panel F
Panel G
Legend
Bacteria
Slow evolving bacterial clade
Slow evolving Bacteria of interest
Eukarya
Archaea
Some other Bacteria
Panel H
Taxon A
Taxon C
0.1
0.1
Z
X
X
Root
Taxon B
Taxon D

With gene family reconstruction, Reciprocal Top Scoring BLAST Hit analysis is normally a stepping stone to further phylogenetic analysis and not a proxy for phylogenetic analysis itself. Since homology is an all or nothing trait, technically there are no degrees of homology and any homolog included into the family is appropriate for gene family reconstruction. Inappropriate results are only possible when homologs are excluded. A bias in exclusion introduced at this step will likely be perpetuated further down the line, with certain homologs being left out of their appropriate larger gene family. There is currently no way to measure the rate of false negatives with a BLAST search and if these false negatives were subject to a systematic bias, then it would be very difficult to determine that it was happening at all. Similarities shared with an ancestral sequence, due to Short Branch Attraction, could reasonably account for a systematic bias in gene family reconstruction.

To study the effect of Short Branch Attraction on BLAST searches, simulations of nucleotide sequence evolution were performed using various branch lengths to determine where in branch length space BLAST searches are subject to Short Branch Attraction and where the top scoring BLAST hit represents the sister taxon. Three sequence lengths, 10000, 1000, and 200 nucleotides, were simulated, to demonstrate the effect of sequence length on branch length space. With increasing sequence length, there is more data available for the program to use when deciding between two matches and the transition zone between returning the phylogenetically informative results and the Short Branch Attraction results is expected to be sharper. 1,000 nucleotides is a common sequence length and should demonstrate what the BLAST program can be expected to do in the case of typical real gene sequence data. 200 nucleotides is the length of typical environmental sequences and will be applicable to phylotyping or ribotyping analyses. A

short sequence length should provide less data and therefore less information when choosing between the various possible matches. Thus the shorter sequence length is expected to show a more gradual transition between the Short Branch Attraction and the phylogenetically informative states.

2.3 Methods

Nucleotide datasets were simulated using Seq-gen (Rambaut and Grassly, 1997) using four taxa trees, as shown in Figure 1 panel H. Central branch lengths vary from 0.004 to 0.1 substitutions per site and variable branch lengths varying from 0.0001 to 0.1 substitutions per site. The lengths of the 2 invariable branches were held at 0.1 substitutions per site for all simulations. These trees were rooted midway along the central branch, as indicated by the ^, to ensure symmetry. Three replicates of 100 simulated datasets were performed for each tree. Each dataset contained four sequences and was generated using the Jukes Cantor 69 substitution model, because of its simplicity (T. Jukes and Cantor, 1969). Three analyses were performed that vary with respect to sequence length: 10000, 1000, and 200 nucleotides. Each of the four taxa of each dataset were used as query against a database comprised of the other three taxa in the dataset to determine the top scoring BLAST hit in all four possible directions. Classic nucleotide BLAST, as implemented in blastall 2.2.19 (Altschul et al., 1997), was used. The results for the 100 datasets in a replicate were totaled to produce the percent of time the top scoring BLAST hit returned the sister taxa, which was graphed using rgl, a 3-dimensional graphing package, in R (Adler and Murdoch, 2012; R Developmental Core Team, 2008). Perl scripts were used to perform repetitive tasks en masse.

2.4 Results and Discussion

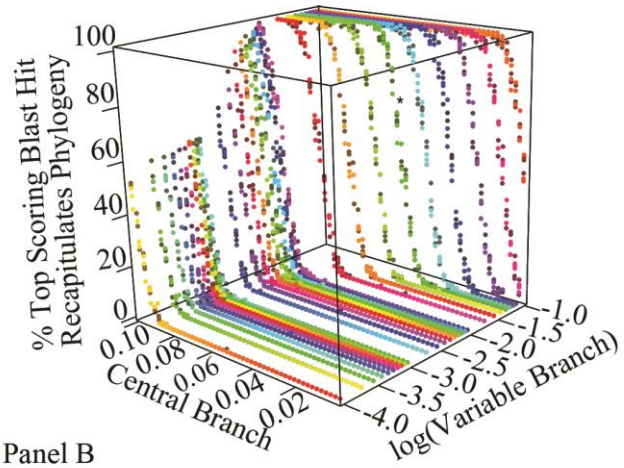
Figure 2 Panel A shows the results of the top scoring BLAST hit analysis using the simulated datasets of 10000 nucleotides in length and the mapping of branch-length space with respect to BLAST result. This graph is a 2-dimensional projection of a 3-dimensional plot. There are three taxa in each database for the query to possibly match, but BLAST almost always returns either the sister taxa or the Short Branch Attraction taxa. Whenever the sequence belonging to the sister taxon is not returned, the sequence belonging to the slowly evolving taxa is returned. When the sister taxon is returned in 0% of the simulations, then the Short Branch Attraction artifact occurs in 100% of the simulations.

Branch lengths are equal in the rear right corner of Figure 2 Panel A and as predicted, the top scoring BLAST hit did return the sister taxon with 100% success in these simulations. As branch lengths shorten, either along the central branch or along the peripheral branches, Short Branch Attraction plays an increasing role until the Short Branch Attraction taxon is returned in 100% of cases. There is a very sharp transition between the two states, as expected with such long sequences. Short Branch Attraction is a major problem when either the central or peripheral branches are one-tenth the length of the invariable branches. It is interesting to note that the effect of short branches is additive so that the effect of decreasing the length of the central and variable branches together is combined (*cf.* equation 1). When the central and variable branches are both half the length of the invariable branches, indicated in the graph with an asterisk, the Short Branch Attraction taxon is already returned in approximately 50% of simulations; branches that are the smallest amount shorter return the Short Branch Attraction artifact 100% of the time. This is cause for concern, because branches that differ by a factor of 2 are not normally considered so heterogeneous as to be a problem. Slightly heterogeneous branch lengths like these are commonplace in real data where both evolutionary rate and time between nodes, can vary.

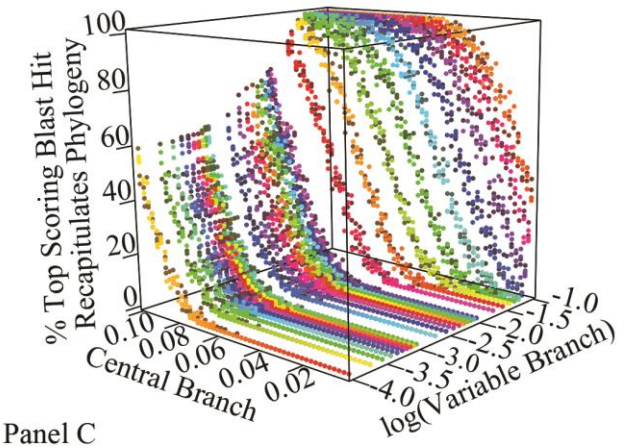
Figure 2: Branch length versus the percent of time the top scoring BLAST hit recapitulates phylogeny. Panel A shows the results of the simulations of 10000 nucleotides in length. The percent of time the top scoring BLAST hit returns the sister taxon is graphed on the Y axis, with 100% meaning that the BLAST search always returns the sister taxon. At 0% the BLAST search always returns Short Branch Attraction; the BLAST search almost never returned the third possibility. Each dot represents the result of 100 simulations. The Z axis gives the length of the central branch, and the X axis the log base 10 of the length of the variable branches. Both the results of the analyses performed using Taxon B as query and those with Taxon D as query are shown, with Taxon B in the bright rainbow colors and D in the darker rainbow colors. The shape of the surface formed by the results is identical for the two analyses, showing that the BLAST hits are reciprocal. For the majority of branch length space, the BLAST search is subject to Short Branch Attraction. It is only free from this phenomenon when the branches are relatively homogeneous in length. Panel B shows the result of using 1000 nucleotide long sequences in the simulations and demonstrates a more gradual transition between the two states and a smaller region in which the sister taxon is returned at a high frequency. Panel C shows the result of using sequences of 200 nucleotide in length. A much greater proportion of branch length space is now occupied by the transition zone, while the area where the sister taxon is returned the majority of the time is now virtually non-existent.

Branch Length versus % of Time Top Scoring Blast Hit Recapitulates Phylogeny

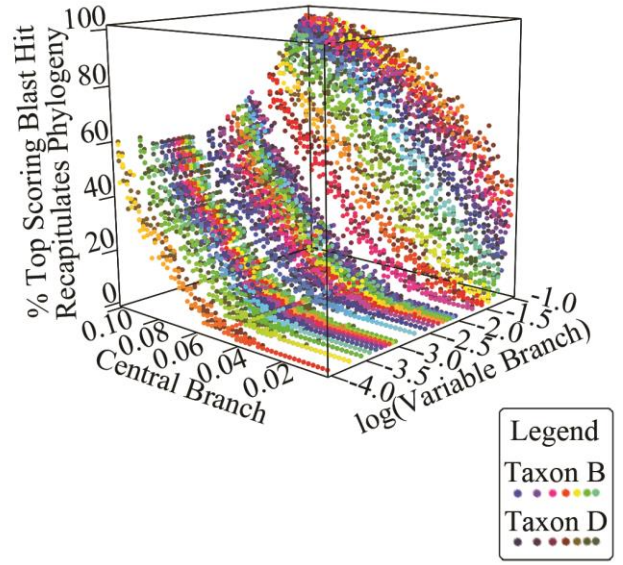
Panel A



Panel B



Panel C



The top scoring BLAST hit analyses were reciprocal, down to the individual simulations, meaning that either Taxa B and D both returned the phylogenetic sister or they both return the Short Branch Attraction taxon. The observed reciprocity results in the curve of the planes for the two analyses having identical shapes and colocalization in branch length space, as shown in figure 2 Panel A.

When the nucleotide sequences are average in length, 1000 nucleotides per sequence, as shown in Figure 2 Panel B, the transition is less distinct between returning the closely related sequence and the slowly evolving sequence. There is still a region in the rear right of the graph, where branch lengths are equal and BLAST returns the sister taxon 100% of the time, however, this region occupies a slightly smaller portion of branch length space. There is an extended region where BLAST transitions between the two possible answers, where any simulation stands a chance of returning the Short Branch Attraction taxon or the sister taxon, but not the third possible choice. And there is an extensive region of branch length space where Short Branch Attraction reigns as the most probable top scoring BLAST hit result.

As sequence length decreases even further to 200 nucleotides, as shown in Figure 2 Panel C, the transition between sister taxon and the slowly evolving taxon becomes even less distinct. There is no region within the coordinates of this graph where the sister taxon is returned in 100% of the simulations and the region where BLAST returns the sister taxa in the majority of simulations is quite small. There is still a region of branch length space where Short Branch Attraction is returned in 100% of simulations, but this region is less than that seen for the other sequence lengths. There is a much greater proportion of branch length space where it is possible to return either the slowly evolving taxa or the sister taxa as the top hit. This means that shorter sequence lengths are less likely to suffer from Short Branch Attraction, but are also less likely to

return the sister taxa. Because of the minuscule amount of information contained in such short sequences, the top BLAST hit of short sequences are increasingly likely to be a result of random chance. The top BLAST hit cannot return any sequence, however, because Taxon C, the third possibility, does not occur even with this short sequence length, when BLAST results seem so intensely ruled by chance, as seen in Figure 3.

It is possible to provide a mathematic approximation to indicate where in branch length space Short Branch Attraction is likely to be a problem, as shown with the solid curved line in Figure 4, panel A and Equation 2, where Y is set at 0.1 substitutions per site. This mathematical description closely coincides with the observed simulation data in figure 2, as shown by the Xs. Light colored Xs indicate that the sister taxa is returned in greater than 80% of simulations, while dark Xs indicate that the Short Branch Attraction artifact is returned in greater than 80% of cases. This relationship between branch lengths can be used to alert one to the dangers of Short Branch Attraction. When the length of the variable branch plus the length of the central branch is less than the length of the invariable branch, then the shortest evolutionary distance in the tree is between the two short branches and Short Branch Attraction is likely to be a problem.

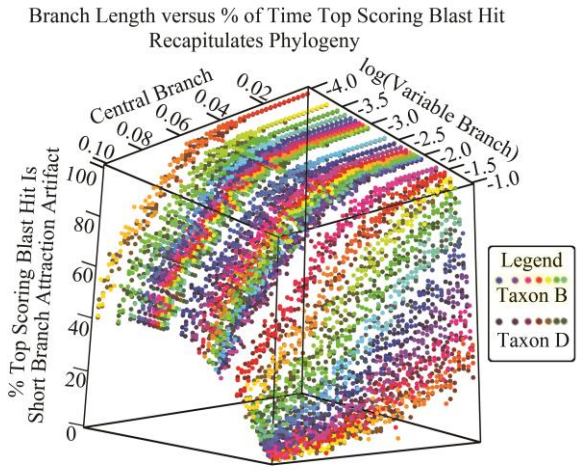
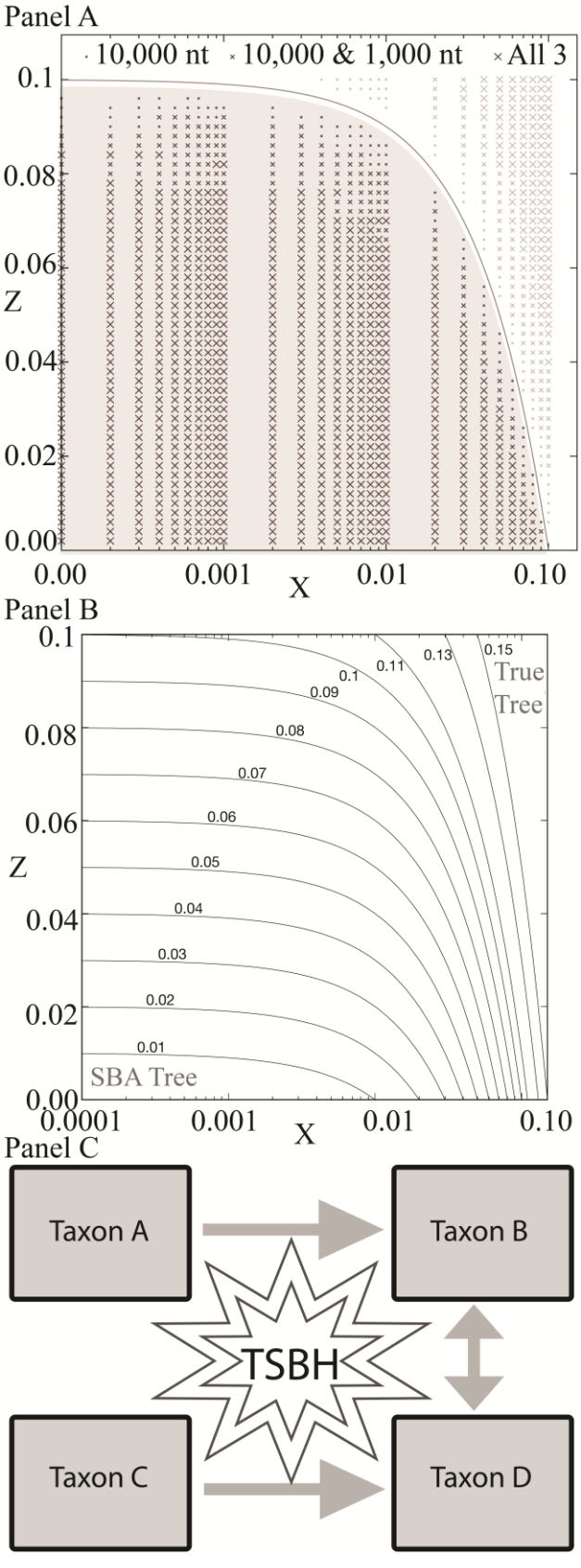


Figure 3: Percent of time Short Branch Attraction is return versus branch length for sequences of 200 nucleotides in length. At 100%, Short Branch Attraction is always returned and occurs when branches are short. The sister is never returned in 100% of simulations, but when branch lengths are equal, the sister taxon is returned the majority of the time, illustrated by the minimum point on this graph. This graph is the inverse of the graph in Figure 2 Panel C, where the percent of time the sister taxa is returned, is graphed.

Figure 4 panel B graphs Equation 2 using values of the length of the invariable branch, Y, ranging from 0.01 to 0.15. The invariable branch is the sister to the variable branch, so when this value is small, the shortest evolutionary distance is to the sister taxon in much of branch length space. As the length of the invariable branch increases, the proportion of branch length space where the shortest total branch length is between the two variable branches, increases. Therefore, Short Branch Attraction dominates branch length space when the invariable branch is long, but not when the invariable branch is short.

LBA is a well-known phenomenon in which heterogeneous branch lengths introduce artifacts into phylogenetic reconstructions, but this is not what is happening here in this phenetic reconstruction. The longest branches are not saturated with substitutions and they are behaving as expected in analyses when used as query. In 100% of cases where Short Branch Attraction is shown grouping the two short branches together, the two longer invariable branches are not subject to this phenomenon and instead return the sister taxon as the top scoring BLAST hit (data not shown). This confirms that LBA is not the cause of the phenomenon. This observation also suggests a means to detect Short Branch Attraction. In the four taxa tree shown in Figure 1, Panel H, the two taxa picking each other as top scoring BLAST hit as a result of Short Branch Attraction are each the top scoring BLAST hit of their true sister taxon, as shown in Figure 4 Panel C. Therefore, using all taxa included in the analysis as queries will result in non-reciprocal best BLAST hit relationships with the taxa giving rise to Short Branch Attraction. If the sister taxon also returns the taxon of interest as the top scoring BLAST hit, then this reveals possible Short Branch Attraction and a phylogenetic tree should be constructed to elucidate the evolutionary history of the gene of interest. It is interesting to note that the program DarkHorse is already attempting to promote matches to more closely related taxa above those to more distantly related ones (Podell and Gaasterland, 2007). Doing so will likely reduce the effect of Short Branch Attraction in BLAST-based analyses.

Figure 4: Ways to determine if Short Branch Attraction is a problem in a particular BLAST search. The solid line in Panel A plots Equation 2: $Y = Z + X$, where Y is set to 0.1 substitutions per site. The region filled with dark colored Xs shows where in branch length space Short Branch Attraction occurs when the invariable branches are of length 0.1 substitutions per site. The region filled with the light colored Xs shows where in branch length space the sister taxon is returned. The Largest Xs represent areas where all three simulations return the result greater than 80% of the time. The medium size Xs represent areas where only the simulations of 10000 and 1000 nucleotides in length return the result greater than 80% of the time. The smallest Xs represent areas where only the simulations of 10000 nucleotides in length return the result in greater than 80% of cases. If data falls into the region of branch length space with the dark colored Xs, top BLAST hit should not be used to inform on phylogeny. Panel B plots Equation 2 for values of Y between 0.01 and 0.15 substitutions per site. The area under the curve represents where Short Branch Attraction is expected with BLAST searches. The area above the curve represents where BLAST is expected to return the sister taxon as the top scoring hit. As the length of the invariable branch, Y, is increased, an increasing proportion of branch length space is dominated by Short Branch Attraction. In Panel C, the sister taxon indicates when Short Branch Attraction might be occurring. In instances of Short Branch Attraction, both the sister taxon and the other slowly evolving taxon return the taxon of interest as the top scoring BLAST hit.


Panel A
10,000 nt
10,000 & 1,000 nt
All 3
0.1
0.08
0.06
Z
0.04
0.02
0.00
0.00
0.001
X
0.01
0.10
Panel B
0.1
0.08
0.06
Z
0.04
0.02
0.00
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.11
0.13
0.15
True
Tree
SBA Tree
0.0001
0.001
X
0.01
0.10
Panel C
Taxon A
Taxon B
TSBH
Taxon C
Taxon D

2.5 Conclusion

If BLAST is used only for its original purposes of detecting homology and finding the most similar sequence from a database, then no issues result from returning a sequence that is not the most closely related one. When using BLAST, or any other measure of similarity, as a proxy for phylogeny, slowly evolving taxa attract each other as top scoring hits, creating Short Branch Attraction. Erroneous detection of HGT events or incorrect phylotyping can result from this phenomenon. Slowly evolving paralogs due to an increase in rate heterogeneity, but a similar overall rate of substitution, are more likely to be assembled into gene families. Paralogs with a more homogeneous substitution rate or an overall increase in substitution rate might be systematically left out of gene family assemblies.

Short Branch Attraction, a phenetic artifact, is different from LBA, a phylogenetic artifact and is the result of slowly evolving sequences retaining the same character state. Short Branch Attraction can occur with as little as a factor of two in terms of branch length heterogeneity and contributes to changing results in the detection of putative gene transfers as the non-redundant database grows. To test for Short Branch Attraction one should use all target sequences as queries, or at least the homolog from the expected sister taxon. If the sequence of interest is the best (non-reciprocal) BLAST hit from another homolog, further analyses need to be performed to elucidate the phylogenetic neighbor.

Chapter 3: The Fundamental Bipartition of Cellular Life and its Implications for Systematics and Taxonomy

3.1. Abstract

In 1989, the root of the cellular tree of life was placed on the bacterial branch using ancient duplicated genes, meaning that the Archaea and eukaryotic nucleocytoplasm form a clade. A quarter century of further investigation has upheld this finding. Because this split is the first one in the cellular tree of life, it represents a taxonomic ranking higher than that of the domain, which we here term the realm. There are two realms of life, just as there are two sides of the cellular tree of life. We name the clade containing the Archaea and eukaryotic nucleocytoplasm the Ibisii based on a number of shared characteristics having to do with information processing and translation. The Bacteria are the only known members of the other realm, which we call the Bacterii.

The taxonomy at the highest levels of the cellular tree of life has been muddled in recent years by the discovery of new major lineages and by rampant horizontal gene transfer between lineages. Here an attempt is made to extend Hennig's principles of taxonomy to apply to the entire cellular tree of life. We call attention to the importance of paraphyletic grades and federations (groups joined by frequent horizontal gene transfer) to evolution. Finally, we revisit the domain level and suggest adjustments to the taxonomy that may be necessary, if the Archaea prove to be paraphyletic and in fact represent a paraphyletic grade of life.

3.2. Principles of Taxonomic Classification

In 1965 Hennig established the gold standard of modern systematic classification as one that reflects phylogeny, names only monophyletic groups, provides a name for the clades

resulting from every bifurcation in the tree, and names taxonomic rankings that contain clades approximately equal in age (Hennig, 1965c). These principles were developed in insects, but work well with most multicellular Eukarya. However, Horizontal Gene Transfer (HGT), fusion of lineages, and the absence of polarized characters complicate matters when trying to classify all of cellular life, leading some to question the usefulness of a bifurcating tree of life (Doolittle and Brunet, 2016). To work around the first two complications, it has been suggested that primacy should be given to one gene or a small subset of genes (Ciccarelli et al., 2006). But using only one gene ignores the contribution of every other gene in the genome and often makes resolution difficult. Using a concatenation still ignores the majority of the genome and has the added disadvantage that if genes with different histories are used, the resulting tree may reflect neither the history of the organisms nor the history of the genes (Gogarten and Townsend, 2005). In an ideal approach, any taxonomic unit should be comprised of a group of organisms that have been traveling together through shared ancestry and all forms of gene exchange (sex and HGT, which is often mediated by the mobilome and at the whim of evolutionary forces playing out in the virosphere (Villarreal, 2016)) for a long enough period of time so that they separate themselves from other organisms in gene content, way of life, and exhibit loose reciprocal monophyly, i.e. for any given gene, accepted taxa are monophyletic and do not violate the monophyly of other accepted taxa, excepting where it is reasonable that these violations occurred due to HGT. Which parts and how much of the genome reflects the organismal history will necessarily vary over the tree of life and will be up to the discretion of each field to determine. This is similar to the bacterial species definition given by Dykhuizen and Green in 1991 (Dykhuizen and Green, 1991), in that HGT replaces sex as the mechanism for gene exchange in

some organisms. But while Dykhuizen and Green restricted HGT to primarily being of importance within a species, we acknowledge that it is a major force driving evolution that may affect a large fraction of the genome and having affected all levels of the tree of life, is not restricted to species. As such, a true representation of evolutionary history of life must take HGT into account. However, because HGT is more frequent between close relatives, it tends to reinforce relationships, building them up, not tearing them down (Andam et al., 2010a; Pace et al., 2012)

HGT functions at two levels. At the species level, mentioned above, it allows close relatives to share genes in a manner analogous to sex. The second, broader level, is across distantly related groups in close physical contact, often due to shared niche or close ecological association, creating highways of gene sharing (Beiko et al., 2005). With our modified taxonomic principles, highways of gene sharing are allowed without any addition to the taxonomy, violating strict reciprocal gene monophyly, up until the point where the distinction between the two contributing taxa is eroded. This is necessarily a subjective criteria and will change as science progresses and the evolutionary history of life on Earth is better understood. Only when gene exchange is so prevalent that the true organismal history cannot be detangled do Hennig's taxonomic principles break down and our extensions are needed.

Often over time, this intense gene exchange will wane or the taxa will become so thoroughly mixed that they become one organism (Timmis et al., 2004), so that loose reciprocal monophyly is restored and our taxonomic principles hold again, as with Eukarya after incorporation of the mitochondria. For an depth discussion on how to treat such groupings in phylogenetics, see (Mindell, 1992).

For example, the number of genes horizontally transferred between Thermotogae and Archaea through a highway of gene sharing is on the order of 10% of the genome (Zhaxybayeva et al., 2009c) and gene phylogenies place these genes inside the Archaea. Nonetheless, these HGT events are not sufficient to erode the distinguishing bacterial characteristics of the taxon (Zhaxybayeva et al., 2009c). Therefore, they have not violated loose reciprocal monophyly and are still distinct taxa.

On the other hand, with our same Thermotoga example, HGT with Clostridiae was so rampant that the majority of Thermotogae genes are placed among clostridial homologs. One could use special genes, such as genes encoding the translation machinery that are only infrequently transferred between divergent organisms to provide a reference tree relative to which gene transfer events can be mapped (Williams et al., 2011). However, it seems premature to give one cellular process primacy in defining taxonomy. Rather than arbitrarily assigning primacy to one gene or a small subset of genes, we prefer to remain agnostic, leaving the split between Thermotogae, Aquificae, and Clostridiae unresolved.

Two similar cases were described as highways of gene sharing overwhelming the phylogenetic signal retained in the genome: Aquificae (Boussau et al., 2008) and Sphaerochaeta (Caro-Quintero et al., 2012). These highways of gene sharing between different bacterial phyla make it difficult to determine some between phyla relationships in Bacteria, and we consider it prudent to leave these relationships unresolved at present.

When a group has spread out into multiple niches and begun to differentiate into multiple groups, but does not yet meet the criteria of loose reciprocal monophyly exhibiting sufficient phylogenetic distinction to tease the two apart, then they are to be considered sub-taxa of a taxon

in the process of diverging into two or more; conversely, this divergence may be the prelude to future condensation back into one single taxon through frequent gene exchange (Sheppard et al., 2008; Zhaxybayeva et al., 2009a) or through extinction of groups with derived characters. If the sub-taxa continue evolving independently, then they will eventually resolve into loosely reciprocally monophyletic groups. This applies to all taxonomic rankings, so that any level of ranking in which taxon amalgamation results in unbreakable polytomies is not a valid ranking worth delineating. But, the conglomeration is in some ways a unit and as such, equally deserving of a name as paraphyletic grades, because they represent a lineage fusion event in life's evolutionary history. To ignore them is to brush aside a significant part of evolution on Earth.

Clades are groups that share synapomorphies. Paraphyletic grades are groups that share Sympleisomorphies. But what are groups that share horizontally transferred genes through highways of gene sharing (including endosymbiosis followed by endosymbiotic gene transfer)? We propose to call them federations, because they bring together and link formerly separate groups, just like formerly independent states are joined politically in a federation of states. A monophyletic federations is thus composed of a composite ancestor, itself resulting from a reticulation in the tree of life, and all of its descendants (Mindell, 1992). This is an extension of the term used for *Prochlorococcus*, a group of bacteria united by horizontal gene transfer (Biller et al., 2015).

The Thermotogae and Clostridia form a federation linked by a massive highway of HGT so great that it has all but erased the underlying vertical branching pattern of the organisms that engaged in these HGT events. Therefore, it is useful to name this federation and others that we

can observe in nature. Lichen is a federation of algae and fungi (de Bary, 1879). The Eukarya are a federation composed of the Bacteria endosymbiont that gave rise to the mitochondria and a pre-mitochondriate stem-group Eukaryeon (Sagan, 1967). The list goes on and on. However, if the reticulation in the tree is easily teased apart through current phylogenetics and the partners are not interdependent, then it is not necessary to create a federation.

3.3. The Clade Comprising Archaea and the Eukaryotic Nucleocytoplasm

When discussing Eukarya, we recognize that Bacteria (in the form of mitochondria and plastids, and through individually transferred genes(Pittis and Gabaldón, 2016)) made important contribution to the modern eukaryal cell, and that the uptake of the mitochondrial symbionts may have triggered key events in Eukaryogenesis. Summarizing both the symbiont and host contributions to the eukaryotic cell, a rooted net undoubtedly is an appropriate representation of gene flow (Margulis, 1995; Rivera and Lake, 2004; Williams et al., 2011). Nevertheless, Bacteria derived cell organelles remain clearly distinct from the nucleocytoplasmic component of eukaryal cells and genes of bacterial origin are usually distinguishable from those of host origin. Hence, hereafter only the nucleocytoplasmic component will be referred to as Eukarya.

In 1989, the tree of life was first rooted on the bacterial branch using the ancient gene duplication of the catalytic and non-catalytic subunits of the ATPsynthase (Gogarten et al., 1989). This split was suggested a decade prior when it appeared in a phylogenetic tree of the 5S subunit of the ribosome (Hori et al., 1982; Hori and Osawa, 1987, 1979) and exhibited shared primary sequence in a well known loop in region A and the base-paired region B of the 5S rRNA (Hori et al., 1982). Similarities between Archaea and Eukarya were also observed in the

sequences of L7/L12-type acidic ribosomal protein, ferredoxins, RNA polymerase, and the sequence of initiator tRNA ((Hori et al., 1982) and references therein). Similarities are also present in the Signal Recognition Particle (SRP), the exosome, a number of ribosomal proteins, and DNA replication, including helicase ((Gribaldo and Brochier-Armanet, 2006) and references therein). Additionlly, Archaea and Eukarya share DNA polymerase delta (Olsen and Woese, 1997). Unfortunately at the time these molecular trees remained unrooted, so it was not until the ancient gene duplication in the ATPsynthase was used that the divide of cellular life into Bacteria on one side and Eukarya and Archaea on the other was recognized as a proper clade (Gogarten et al., 1989). The Eukarya and the Archaea were shown to share two synapomorphies in the ATPsynthase: the decay of the walker motif in the Rossmann fold of the non-catalytic subunit (Zhaxybayeva and Gogarten, 2007), and a large insertion in the catalytic subunit (Gogarten et al., 1989). The 1989 paper had far-reaching implications in evolutionary thought throughout biology. A problem is that the authors, including one of the authors of the current manuscript (JPG), did not name their discovery and we have forever since been relegated to describing this aboriginal bifurcation every time we wish to mention it. Given that this is the first split of all cellular life that resulted in forms surviving to present day, we often find ourselves with need to mention this bifurcation.

The rooting of the tree of life was confirmed by analysis of additional ancient gene duplication events, starting with the elongation factors Tu and G (Baldauf et al., 1996a; Iwabe et al., 1989), the signal recognition particle (SRP) and the SRP receptor protein (SR) (Gribaldo and Cammarano, 1998), and different aminoacyl tRNA synthetases. When the isoleucyl-tRNA synthetase was rooted using two of its ancient paralogs, the valyl- and leucyl-tRNA synthetases,

the Bacteria were again shown to diverge first, leaving Archaea and Eukarya as sister taxa (Brown and Doolittle, 1995). The tryptophanyl- and tyrosyl-tRNA synthetases were reciprocally rooted and both exhibited a rooting along the bacterial branch (Brown et al., 1997).

Fournier and Gogarten (Fournier and Gogarten, 2010) used amino acid composition bias in ribosomal proteins to root the ribosomal tree of life and again the root was squarely placed along the bacterial branch, and again the taxon formed by Eukarya and Archaea went without a name. To be able to discuss a phylogenetic tree, all monophyletic groups, clades, must have a name (Hennig, 1966). Even if the monophyly of each individual domain is still in question, the monophyly of the clade comprising the Archaea and Eukarya is almost universally agreed upon, as are a number of shared characters, including the use of TATA-binding proteins (Marsh et al., 1994) and the RNA polymerase (Huet et al., 1983).

In 1987, T. Cavalier-Smith first gave name to a clade containing the Eukarya and Archaea, two years before Gogarten *et al*. and Iwabe *et al*. rooted the tree of life. Cavalier-Smith called them the Neomura, or new-wall, based on an almost unanimously rejected hypothesis that the root of life is within the Gram Negative Bacteria and that Eukarya and Archaea therefore share the derived trait of having lost muramic acid from their cell walls. There is no evidence to suggest that having the ability to synthesize muramic acid is the ancestral state of life, which was a significant factor in Neomura's inability to stick as a name; the other factor being Cavalier-Smith's controversial Gram Negative bacterial root of the tree of life (Cavalier-Smith, 1987).

29 years later google can produce hits to only 160 articles that contain the word Neomura, 20 by Cavalier-Smith himself. Even the man who coined the term acknowledges that it has not caught on (Cavalier-Smith, 2014). The support for placing the root of the tree of life on

the bacterial branch falsifies Neomura as a hypothesis. We thus consider Neomura inappropriate as a name and submit that it is time for a new name. A name with a hypothesis behind it that is scientifically supported by data; a name that fits and has a chance of taking off into the memetosphere.

27 years after the fact, we are putting forth this manuscript in an effort to finally undo the injustice that was done to the clade comprising the Archaea and the Eukarya by offering them a proper, fitting name. Recently there has been a movement to do away with naming the rankings in taxonomy (Group, 1998), because rankings are supposed to indicate clades of a similar age (Hennig, 1966, 1965c), but often do not. But because this is the first split in cellular life, involving only two clades of similar age, we indeed have a rank deserving of a name. Therefore, we propose that this taxonomic level be called the realm, because it is composed of many kingdoms and falls between empire and domain in taxonomic classification schemes. There are two realms of life, one of which is composed of the Bacteria and the other the Eukarya and Archaea. Finally, we have chosen the suffix -ii for all proper taxonomic names belonging to the realm ranking.

When choosing a name for the realm containing the Archaea and the Eukarya, we focus on traits present in all members of this clade and in the clade's common ancestor, and that are exclusive to this clade. Not many features fit these criteria, but the majority that do have to do with the machinery of replication, transcription, and translation. This machinery allows the cells to process the information stored in DNA into functional RNA and protein molecules. Synonyms for information processing include cognition, knowledge, comprehension, and reasoning. The Archaea and Eukarya share the baroque machinery used in what is in essence their cognition at

the cellular level. It is not that their machinery is necessarily better than that of the Bacteria, and in many ways the complexity seems gratuitous, but that it is shared by all members of this clade, was present in the clade's common ancestor, and is not found outside the clade. Undoubtedly some of the individual genes that make up this shared machinery were present in LUCA and lost in Bacteria. Currently we cannot determine which of the many shared traits fall into this category, but others will certainly prove to be synapomorphies. By choosing a name that encompasses the entire cognition of these cells, we are resting our name on a collection of genes and not just a single gene. But finding a name based on cellular cognition is a challenge.

We looked to ancient Egypt for a name, because it is the first recorded realm of human civilization, just like the realms of life represent the first bipartition in cellular life. The Egyptian god of cognition, knowledge, comprehension, reasoning, and wisdom is Dhwty, a god named after the ibis. Owing to the difficulty of pronouncing the name of this ancient Egyptian god, the realm suffix -ii was added to Ibis, giving us Ibisii. Thus we propose to name the realm containing the Archaea and the Eukarya the Ibisii. It is interesting to note that the bill of the ibis, the symbol of Dhwty, is the same shape as the TATA Binding Protein (TBP), which binds to the promoter of DNA and recruits transcription factors and the RNA polymerase in all Ibisii. TBP curves around the DNA, like the bill of an ibis, to initiate transcription. Also, the word transcription comes from the word scribe, or writing. The Ancient Egyptians are famous for their writing, the god of which is Dhwty. This leads us back to Ibisii and the wise ibis as the perfect symbol for a clade based on information processing and writing.

Finally, we propose that the other half of life, comprising only the domain Bacteria, be called the Bacterii. It is possible that some as yet unknown domain that groups on the bacterial

side of the tree may be hiding in the remote regions of the biosphere, as suggested by (Hug, et al. 2016) or that future studies of existing Bacteria may result in splitting Bacteria into multiple domains, possibly confirming previously reported deep splits (Battistuzzi and Hedges, 2009; Boussau et al., 2008). If either occurs, then these additional domains would too belong to the Bacterii, just like if a hidden domain is discovered that is not an Eukaryeon and not an Archeon, but branches on the ibisial side of the tree, would too belong to the Ibisii, because it is the split between Bacteria and Archaea/Eukarya that we are delineating. It is conceivable that there will come a time when additional levels of classification are needed between realm and domain to describe additional splits in the tree of life.

3.4. The Domain Level

While the question of how many realms of life there are was solved by Gogarten *et al.* (Gogarten et al., 1989) and Iwabe *et al.* (Iwabe et al., 1989) back in 1989, the question of how many domains of life (*sensu* (Woese et al., 1990)) there are is still under debate. Woese *et al.* distinguished and named three domains, shown in Figure 1, Panel A, based on unrooted 16S rRNA grouped by S_{AB} similarity criteria. In doing so they found a group of non-nucleated cells that was as different from typical Bacteria as Eukarya. The authors hypothesized that this taxon was one of three aboriginal descendants of a primitive ancestor called the progenote, explicitly rejecting the idea that Archaea could be considered "Proto-Eukaryotes" (Woese et al., 1978); this was superseded by Gogarten et al. (Gogarten et al., 1989) and Iwabe *et al.* (Iwabe et al., 1989) when Archaea were shown to be the sister group to Eukarya. The validity of Woese's findings were first called into question on the basis of computer simulations that suggested that S_{AB}

values might be virtually meaningless (Hori et al., 1982) and change in substitution rates in the 16S exaggerating the difference between Archaea and Eukarya (*Evolution of Life - Fossils, Molecules and Culture | Syozo Osawa | Springer*, n.d.), but the 5S ribosomal phylogeny also found this non-nucleated taxon to be distinct from Bacteria (Hori et al., 1982; Hori and Osawa, 1987, 1979) and showed it to be monophyletic (Hori and Osawa, 1987). Hori named this group the metabacteria (Hori and Osawa, 1979), but that name did not catch on. Woese had historical precedence on his side when he first named them the Archaebacteria. He later changed it to Archaea, which is most popular today. Both of Woese's names were based on the later rejected hypothesis that Archaea directly emerged from an ancient paraphyletic grade of life, the progenote, that also gave rise to both the Bacteria and the Eukarya (Kandler, 1995; Woese et al., 1978). Hori turned out to be correct when it came to Archaea not being old, but the inappropriate name remains to this day.

In 1990, Woese wrote, "The archaebacteria are called Archaea to denote their apparent primitive nature (vis a vis the eukaryotes in particular)." While there is no way of knowing precisely what Woese meant by this, one interpretation is that the term Archaea describes a paraphyletic group defined by shared primitive characters. Given this, if Archaea are found to be paraphyletic as has been argued in recent years (Cox et al., 2008; Foster et al., 2009; Spang et al., 2015), then the fact that Woese had an inkling that the term describes a paraphyletic group lends strong support for the term to revert to the name of a paraphyletic grade. On a side note, paraphyletic grades are not proper taxonomic units and should not be used as such; systematics *sensu* Hennig utilizes only proper clades in assigning taxonomic names (Hennig, 1975b, 1965d). Paraphyletic grades are given non-taxonomic names.

Recently several studies (Cox et al., 2008; Foster et al., 2009; Lasek-Nesselquist and Gogarten, 2013b; Spang et al., 2015; Williams and Embley, 2014) have called into question whether or not the Archaea are monophyletic, or whether the Eukarya actually group within the Archaea, as shown in Figure 1, Panels B, C, and D. If the Archaea are paraphyletic, then this is not a proper taxonomic group, and a four (Figure 1 Panel B), five (Figure 1 Panel C), six (Figure 1 Panel D), or more domain system would be needed. Depending on which particular tree wins out, Bacteria, Crenarchaea or Proteoarchaea, Euryarchaea, Eukarya, Lokiarchaea, DPANN, and possibly any number of the newly sequenced Archaea may all be at the domain level, resulting in three or more ibisial domains. The name Archaea as a level of taxonomy would have to go, but would persist as designation for a paraphyletic grade within the Ibisii.

Trees showing paraphyletic Archaea have been described as 2 Domain trees (Embley and Williams, 2015; Gribaldo et al., 2010; Koonin, 2015; Williams et al., 2013). This in essence suffers from the same problems that were faced back when the domain level was created (Woese et al., 1990). With only 2 Domains of life, Eukarya are demoted to the level of kingdom, or possibly a subdomain. Animals and plants are then demoted from their status as kingdoms, which they have held for centuries. Long recognized phyla would also be demoted and the shifting of taxonomic categories would continue throughout the entire ranking system. Assigning Eukarya to the subdomain category is also problematic, because of the bipartitions between Eukarya and the classical kingdoms of plants, animals, and fungi that are currently being delineated. There just are not enough intermediate levels to satisfy the bipartition pattern. Therefore, we prefer to retain Eukarya as a domain and have 4 or more domains, depending on the precise topology of the Ibisii. The question is not two or three domains, but three or more.

Certain 2-domain papers have gone so far as to suggest the Eukarya are unfit to hold the ranking of domain, because they result from a fusion between two lineages and are a federation (Williams et al., 2013). We reject this idea for two reasons. Firstly, lineage fusions may occur throughout the tree of life (Lake et al., 2015) and such level of pervasiveness combined with throwing out all resulting taxonomic units would necessitate throwing out the entire tree. Secondly, at some point after the fusion, gene exchange with both parental populations was severed, resulting in a monophyletic eukaryal clade that has been evolving independently for over a billion years. Once loose reciprocal monophyly has been restored, as is the case with the Eukarya, then taxonomic efforts should proceed as normal, as described in the **Principles of Taxonomic Classification**, delineated earlier in this work.

There is also data and scientific evidence suggesting that the Archaea are monophyletic and group as sister taxa to the Eukarya (Baldauf et al., 1996b; Brown et al., 1997; Brown and Doolittle, 1995; Gribaldo and Cammarano, 1998; Hori and Osawa, 1987). If this proves to be correct, then the name Archaea would remain, since it would be a proper monophyletic group. Woese would still have first claim to their discovery and the three domain system would hold. This, however, is a separate issue from that of the realm.

3.5. The Paraphyletic Grade Archaea

Mayr's concept of ecofunctional adaptiogenetics of paraphyletic groups is no longer acceptable for naming taxonomic groups (Hennig, 1975b), and rightly so, but has been repurposed as the term evolutionary grade (Huxley, 1959b). There are two types of grades, polyphyletic and paraphyletic. Polyphyletic grades are virtually useless, because they are joined by convergent traits and therefore have no predictive power. Paraphyletic grades are useful when

they define an ancestral state of a group, because knowing the ancestral state has its value in determining evolutionary history. If a group is proven to be paraphyletic, then the exclusive traits they share, symplesiomorphies, are ancestral to the group.

The question of what it means to be an Archaeon is a complex question that depends upon the topology of the tree of life. In the 3-domain tree, shown in Figure 1 Panel A, Archaea are a domain. With the 4-domain tree shown in Figure 1 Panel B, Archaea do not form a clade and are therefore not eligible as a name for any taxonomic unit. Just like the term amniote replaced the term reptile as a taxonomic term when reptiles were shown to be a paraphyletic grade of life rather than a clade, the term Archaea must be replaced in taxonomic classification. We do, however, still mean something when we use the term Archaea. Archaea is still the name that describes the Crenarchaea, the Euryarchaea, and ancestor of the two, but not the Eukarya. Archaea possess a number of similarities to each other that they do not share with Eukarya, such as ether-linked lipid membranes and derived tRNA and rRNAs (Woese et al., 1978). Under this topology, these similarities are either sympleisomorphies, meaning that the shared traits represent the ancestral state, or products of HGT, which means the traits can be shared between domains; both scenarios are intriguing, because the traits in question are so rarely changed throughout the tree of life. In conclusion, under the condition that the Archaea prove to be paraphyletic, the name Archaea stays only with the paraphyletic grade of organisms it was originally assigned to.

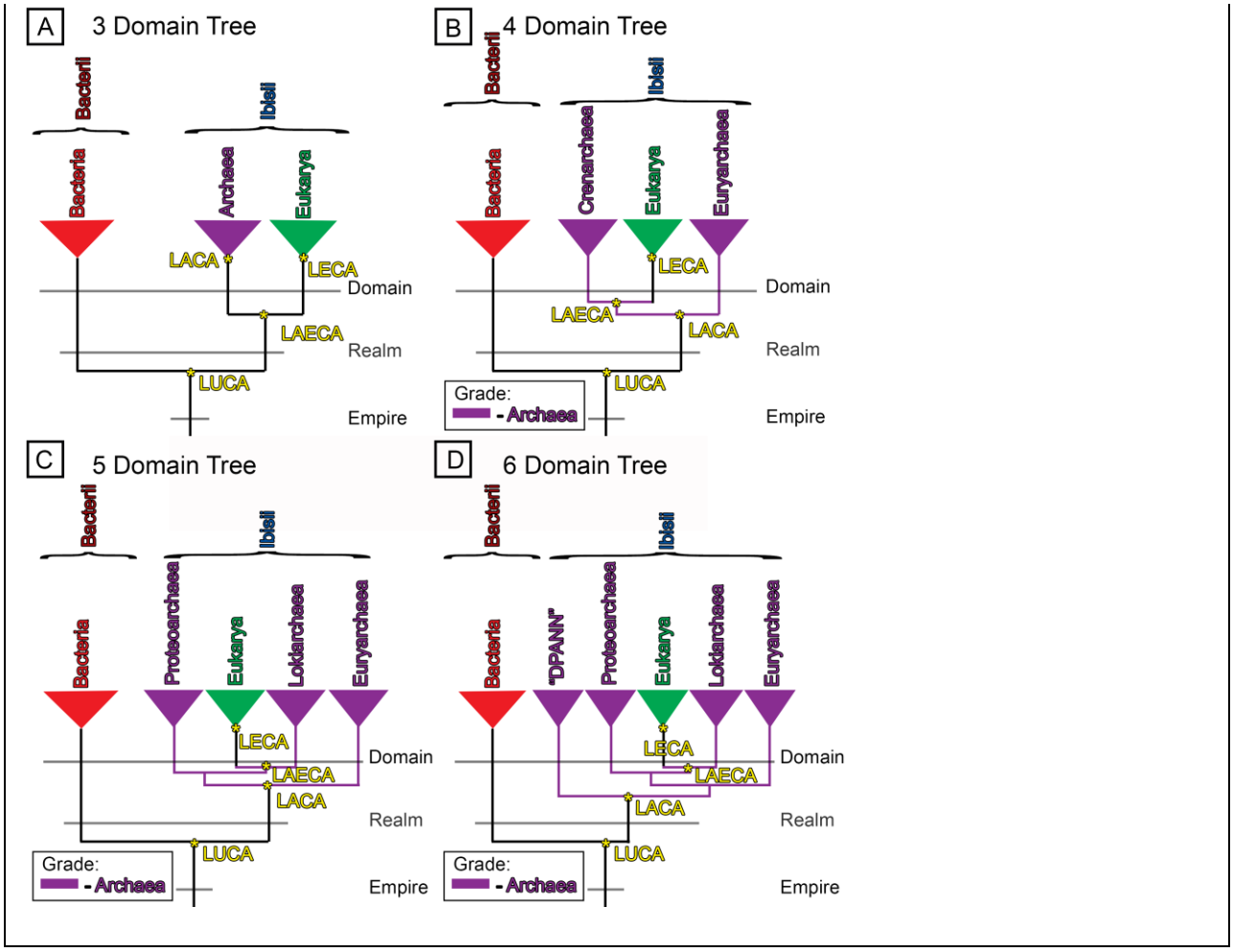
3.6. Summary:

The first bipartition in the cellular tree of life defines two realms of life. These are the Bacterii, presently composed of only the Bacteria, and the Ibisii, composed of the eukaryotic

nucleocytoplasmic component and all of the archaeal domains. If Archaea prove paraphyletic, then the term names a paraphyletic grade composed of prokaryotic Ibisii with ether-linked lipid membranes.

Clades represent the underlying organismal phylogeny. Federations represent HGT, lineage fusion events, and the flow of genes outside of the underlying organismal backbone of the tree of life. Paraphyletic Grades represent the ecofunctional levels and stages of adaptation that life has progressed through in its journey through time. Here we expand the bipartite system that names clades and paraphyletic grades(Huxley, 1959b) to a tripartite system that also names federations, because the combination of all three fully reflects the complexity of the evolutionary history of life on Earth.

Figure 1: LUCA (Last Universal Cellular Ancestor, LACA (Last Archaeal Common Ancestor, LAECA (Last Archaeal Eukaryal Common Ancestor, LECA (Last Eukaryal Common Ancestor). Panel A: 3-Domain Tree showing the cellular empire and corresponding terminology. In this tree, Archaea are monophyletic. Archaea and Eukarya are the domains of the ibisial realm. Bacteria is the only domain in the realm Bacterii. Panel B: 4-Domain Tree showing the cellular empire and corresponding terminology. In this tree, Archaea are not a proper taxonomic name, but a paraphyletic grade. Crenarchaea, Euryarchaea, and Eukarya are the domains on the Ibisial side of the tree. Bacteria are the only domain on the Bacterii side of the tree. This topology requires one additional taxonomic level between domain and realm and a name for the clade composed of the Crenarchaea and the Eukarya which we will leave to the discovers of this split to assign. Panel C: 5-Domain tree, if proposed deep branching archaeal lineages are verified as proper clades. Here Lokiarchaea are promoted to a domain (Spang et al., 2015), but there are other possible 5-Domain trees, depending on which Archaea is found to be sister-taxon to the Eukarya. These trees require 2 additional taxonomic levels and contain two new clades that require names. Panel D: 6-Domain tree, if the so-called DPANN group (Williams and Embley, 2014) is verified as a proper clade and not just a Long Branch Attraction artifact. In this case, DPANN is also promoted to a domain and thus would require a proper name. This topology requires 3 additional rankings and contains 3 new clades that require naming by the discoverers.


A
3 Domain Tree
Bacterii
Ibeii
Bacteria
Archaea
Eukarya
LACA
LECA
LAECA
LUCA
Domain
Realm
Empire
B
4 Domain Tree
Bacterii
Ibeii
Bacteria
Crenarchaea
Eukarya
Euryarchaea
LECA
LAECA
LACA
LUCA
Domain
Realm
Empire
Grade:
- Archaea
C
5 Domain Tree
Bacterii
Ibeii
Bacteria
Proteoarchaea
Eukarya
Lokiarchaea
Euryarchaea
LECA
LAECA
LACA
LUCA
Domain
Realm
Empire
Grade:
- Archaea
D
6 Domain Tree
Bacterii
Ibeii
Bacteria
"DPANN"
Proteoarchaea
Eukarya
Lokiarchaea
Euryarchaea
LECA
LAECA
LACA
LUCA
Domain
Realm
Empire
Grade:
- Archaea

Chapter 4: The Relationship Between Archaea and Eukarya

4.1. Eukaryal Stem Branch Length

This chapter contains one article (Fournier et al., 2011). The article uses three tubulin paralogs that diverged along the eukaryal stem branch to investigate the length of this stem branch as a measure of eukaryal evolution. The chapter also contains an addendum that repeats the analysis using the four histone paralogs, which were also duplicated on the Eukaryal stem branch.

The research presented here was conceptualized with the assistance of Dr. J. Peter Gogarten and Dr. Greg Fournier. The phyml tree (figure 1) in the article was produced by myself, while the rest of the analyses were carried out by Dr. Greg Fournier and Dr. David Willams. The research in the addendums was carried out by myself. Dr. J. Peter Gogarten supervised the research and assisted with writing the article. Permission for reprinting the article was granted by Elsevier, see appendix.

4.2. Evolution of the Archaea: Emerging Views on Origins and Phylogeny

ELSEVIER

INSTITUT PASTEUR

Research in Microbiology 162 (2011) 92e98

www.elsevier.com/locate/resmic

Evolution of the archaea: emerging views on origins and phylogeny

Gregory P. Fournier [a,*], Amanda A. Dick [b], David Williams [b], J. Peter Gogarten [b]

[a] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[b] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269, United States

Received 13 April 2010; accepted 10 September 2010
Available online 27 October 2010

Abstract

Of the three domains of life, the Archaea are the most recently discovered and, from the perspective of systematics, perhaps the least understood. More than three decades after their discovery, there is still no overwhelming consensus as to their phylogenetic status, with diverse evidence supporting in varying degrees their monophyly, paraphyly, or even polyphyly. As a further complication, their evolutionary history is inextricably linked to the origin of Eukarya, one of the most challenging problems in evolutionary biology. This exclusive relationship between the eukaryal nucleocytoplasm and the Archaea is further supported by a new methodology for rooting the ribosomal Tree of Life based on amino acid composition. Novel approaches such as utilizing horizontal gene transfers as synchronizing events and branch length analysis of deep paralogs will help to clarify temporal relationships between these lineages, and may prove useful in evaluating the numerous conflicting hypotheses related to the evolution of the Archaea and Eukarya.

Keywords: Archaea; Tree of life; Eocyte; Eukaryogenesis

1. Introduction

Before the development of the tools of molecular evolution and the discovery of Archaea (Woese and Fox, 1977a,b; Woese et al., 1978) the central feature of the tree of life was clearly understood to be the dichotomy between prokaryote and eukaryote, polarized by the apparently obvious difference in complexity between the two (Sapp, 2006). However, the discovery of the Archaea gave rise to the "three domain" model for the Tree of Life, with a trichotomy that certainly was conceptually central, but not necessarily basal or aboriginal in an evolutionary sense (Woese and Fox, 1977a,b; Woese and Gupta, 1981; Gogarten and Taiz, 1992). In an unrooted three domain tree of life, one cannot decide which two of the domains are more closely related; however, (Woese et al., 1978) argued strongly against viewing the Archaea as proto-eukaryotes and in favor of a tripartition of life. The evolutionary relationship between these three groups became one of the most important questions in evolutionary biology. This question is best understood in two interrelated parts: (1) are the Archaea truly a monophyletic group defined by shared derived characters, or do they only appear as such from the perspective of the distantly related Bacteria and Eukarya; and (2) given the correct relationships between the three groups, where is the root corresponding to the Most Recent Common Ancestor (MRCA)? (We here use the term monophyletic as originally defined by (Hennig, 1966) to denote a holophyletic group sensu (Ashlock, 1971), and we denote paraphyletic groups explicitly as paraphyletic, i.e., a group of organisms that is derived from a common ancestor that also gave rise to organisms that are not members of this group is labeled as paraphyletic and is not considered a sub-type of monophyly). While the last few decades have seen an explosion of biological data and diverse analytical techniques brought to bear on these questions, an overwhelming consensus has yet to emerge.

* Corresponding author.
E-mail addresses: g4nier@mit.edu (G.P. Fournier), amanda.dick@uconn.edu (A.A. Dick), david.williams@uconn.d-dub.org.uk (D. Williams), gogarten@uconn.edu (J.P. Gogarten).

0923-2508/$ - see front matter Ó 2010 Institut Pasteur. Published by Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.resmic.2010.09.016

Two major competing hypotheses for the topology of the Tree of Life are currently seriously debated, corresponding to a monophyletic Archaea (the "traditional" three-domain model), or a paraphyletic Archaea (the "Eocyte" model). Given these topologies, the location of the root is also in question, with different evidence purported to support a rooting either within the bacterial domain (Lake et al., 2009; Skophammer et al., 2007; Cavalier-Smith, 2010, 2006b, 2002), along the bacterial branch (Gogarten et al., 1989; Iwabe et al., 1989; Brown and Doolittle, 1995; Gribaldo and Cammarano, 1998; Baldauf et al., 1996; Fournier and Gogarten, 2010), along the eukaryal branch (Poole et al., 1999; Glansdorff et al., 2008; Forterre and Philippe, 1999), or within the Archaea (Di Giulio, 2007), even redefining this group as polyphyletic (Lake, 1988; Lake et al., 1985). In contrast, rootings within the Bacteria or on the bacterial branch are compatible with either the monophyly (Yutin et al., 2008) or paraphyly (Cox et al., 2008; Foster et al., 2009) of existing Archaeal groups. While on balance these analyses tend to support the traditional "three domain" topology with a monophyletic Archaea and a root along the bacterial branch, there are enough conflicting data and valid critiques to warrant careful re-consideration, especially given the emerging understanding that horizontal gene transfer and lineage fusion complicates inferring organismal evolutionary histories (Zhaxybayeva and Gogarten, 2004). Additionally, it is increasingly apparent that the origin and evolution of the Eukarya is intimately associated with these questions, and cannot be considered as an independent problem. Not only is the nature of the eukaryal ancestor and its subsequent evolution dependent on both the location of the eukaryal root and the MRCA, but also reasonable inferences concerning the former can actually inform debates concerning the latter.

There are currently two major competing eukaryogenesis hypotheses, with Eukarya either emerging as a derived sister to the Crenarchaeotes (Archaebacterial hypothesis, (Cox et al., 2008; Foster et al., 2009)), or as a deep proto-archaeal or non-archaeal lineage (Archaeozoan hypothesis as described by (Poole and Penny, 2007)). While additional models (bacterial-archaeal fusion events (Searcy, 1992; Zillig et al., 1992), phagocytosis involving primitive RNA-based organisms (Hartman, 1984; Sogin, 1991), or inclusion of an additional viral component are intriguing and not without their advocates), these do not yet have a supporting body of evidence significant enough for serious consideration. Also these models depend upon substantial post-hoc reasoning (Poole and Penny, 2007) and often are not the most parsimonious explanation of the currently available data (see below). The Archaebacterial model is consistent with an Eocyte topology/rooting of the Tree of Life, while the Archaeozoan hypothesis is only compatible with scenarios in which the Archaea are monophyletic, regardless of the location of the root.

Phylogenies based on gene trees have played a substantial role in discriminating between different models for the Tree of Life. Typically, these models attempt to recover either the traditional or Eocyte tree topology, as well as locate the placement of the Eukarya with respect to archaeal sequences. However, these phylogenetic methods (or any reconstruction of ancient organismal evolution) are complicated by artifacts due to long branch attraction (Tourasse and Gouy, 1999), compositional heterogeneity (Cox et al., 2008; Foster et al., 2009), and the exchange of genetic information between divergent organisms (Gogarten et al., 2002). Complications due to phylogenetic analysis also impact methods for rooting the Tree of Life that depend upon phylogeny, specifically those utilizing reciprocal rooting of ancient paralogs (Gogarten et al., 1989; Iwabe et al., 1989; Brown and Doolittle, 1995; Gribaldo and Cammarano, 1998; Baldauf et al., 1996; Fournier and Gogarten, 2010; Zhaxybayeva and Gogarten, 2007; Philippe and Forterre, 1999). It is also important to note that all individual genes that have universal or near universal distribution will have a molecular MRCA that could be interpreted as a proxy for the location of the organismal root. However, given the large amount of gene sharing between lineages, including orthologous replacements, it is almost certain that the different molecular MRCAs did not all coexist in the organismal MRCA (Zhaxybayeva and Gogarten, 2004). It therefore may be problematic to deduce features of the universal tree of life from a character that is associated with a single gene (Skophammer et al., 2007; Lake et al., 2008; Gupta, 2001; Zhaxybayeva and Gogarten, 2007).

The long branch leading to the Eukarya detected in most gene trees is a significant feature directly related to the chronology of events within archaeal evolution, a critical part of both the Archaebacterial and Archaeozoan models. Molecular data suggest that all extant eukaryotes descended from ancestors that had harbored the mitochondrial endosymbiont (Horner et al., 1996; Hampl et al., 2008). This demise of the archezoa as an extant group of Eukaryotes changed how models for the early evolution of Eukaryotes were evaluated (Doolittle, 1996). Many earlier models that postulated a chimeric formation between Bacteria and Archaea giving rise to the eukaryal stem group (Zillig et al., 1992; Hartman, 1984; Sogin, 1991) also became less convincing because many features ascribed to an earlier fusion could also be due to the uptake of the endosymbiont that evolved into the mitochondrium and hydrogenosome (e.g., (Martin and Muller, 1998; Martin and Koonin, 2006)). However, substantial support remains for the proto-eukaryote lineage accumulating numerous derived characters before the endosymbiotic event leading to the mitochondria (Poole and Penny, 2007; Margulis, 2009; Gray, 1993), including the development of the nuclear, cytoskeletal, and spliceosomal machinery. If the pre-mitochondrial derivation of proto-eukaryotes was rapid, then a more recent Crenarchaeal origin is more plausible, as proposed in the Archaebacterial hypothesis. However, if the accumulation of these derived eukaryal characters was gradual, the more ancient branching that is a feature of the Archaeozoan hypothesis would be more chronologically compatible.

Novel approaches in molecular evolution are required to address these problems. Phylogenetic trees of paralogous genes undergoing duplications along the eukaryal stem can be analyzed for evidence of symmetric/asymmetric accumulation of substitutions along internal branches, supporting either gradual or rapid divergence. Compositional analysis of conserved positions along deep branches can independently

identify the organismal root within gene families with potentially different topologies. Finally, ancient horizontal gene transfer (HGT) events can be used not only as characters to define clades, but as a means of synchronizing evolutionary events across the Tree of Life, providing differential support for hypotheses of early life evolution, including archaeal and eukaryal origins.

2. Methods

2.1. Tree Construction

Tubulin and TyrRS amino acid sequences were downloaded from the NCBI using Entrez (http://www.ncbi.nlm.nih.gov/). Tubulin sequences were selected to provide a representative sampling of eukaryal diversity. TyrRS taxa were selected to provide a similar sampling of eukaryal and archaeal diversity, with a smaller subset of bacterial sequences as outgroup. Initial alignments of tubulin homologs, TyrRS homologs, and ribosomal protein families were performed using Muscle (Edgar, 2004). Profile alignments between tubulin homologs were generated using Clustal X 2.0 (Larkin et al., 2007). Trees were constructed using phylobayes (default CAT (Lartillot et al., 2009) and phyml (WAG model, estimated PINVAR + Gamma, four rate categories, 100 bootstrap replicates) (Guindon and Gascuel, 2003). Trees were visualized with FigTree v1.2.2 (Rambaut, 2007). Slow and fast evolving sites were determined using an in house PERL script modified from (Swithers et al., 2009b).

2.2. Absence/presence analysis

Ribosomal protein sequences for all complete archaeal genomes were downloaded from the NCBI using Entrez (http://www.ncbi.nlm.nih.gov/). Apparent absences within archaeal genomes were confirmed using both blastp and tblastn (Altschul et al., 1990).

3. Results and discussion

3.1. The eukaryal stem

The tubulin protein family is present in all Eukarya and is essential for the formation of microtubules that are involved in mitosis, cytokinesis, and vesicle transport. Distantly related to the prokaryotic FtsZ protein, at least three tubulin homologs are universally distributed among Eukarya (α, β, and γ-tubulin), indicating they diverged at a time before the most recent common eukaryal ancestor. In a critique of paralog rooting methods, it has been suggested that an accelerated rate of evolution in the diverging homolog after a duplication event will invariably create long deep branches (and subsequently lead to long branch attraction artifacts) (Cavalier-Smith, 2006a,b). For this reason, one cannot use stem branch lengths as direct evidence for the age of the tubulin gene family (and therefore a lower-bound age for the eukaryal stem group). However, since the eukaryal crown group within each paralog encompasses approximately the same amount of absolute evolutionary time, comparing branch lengths across paralogs within these groups allows for an indirect inference of evolutionary rates, which can then be applied to the stems of each paralog. For a phylogenetic tree of α, β, and γ-tubulins (Fig. 1), paralog stems show branch lengths of 1.081, 0.717, and 1.978 substitutions/site, respectively. Comparing congruent branches within the crown group of each paralog shows close to a 1:1 linear correlation between α and β branch lengths ($R = 0.836$, $p = 0.005$), both of which are approximately 45% the length of the corresponding branches within the γ-tubulin phylogeny ($R = 0.632$, $p = 0.005$). These ratios are similar to those of the branch lengths of the stems (Wilcoxon two-sample test, p(α vs. β) = 0.667, p(α, β vs. γ) = 0.8526). This symmetry between paralog branches would not be expected in the case of rapid divergence following duplication, as in such a case initial rates of divergence would be independent of subsequent rates within each crown group. Assuming the placement of the root at a position deep on the γ stem, this is in agreement with a substantial amount of evolution occurring in a long proto-eukaryal organismal history, with a fairly constant evolutionary rate for each paralog across both stem and crown groups. While rapid evolution following paralog divergence should still occur, its contribution would be substantially less than what is required

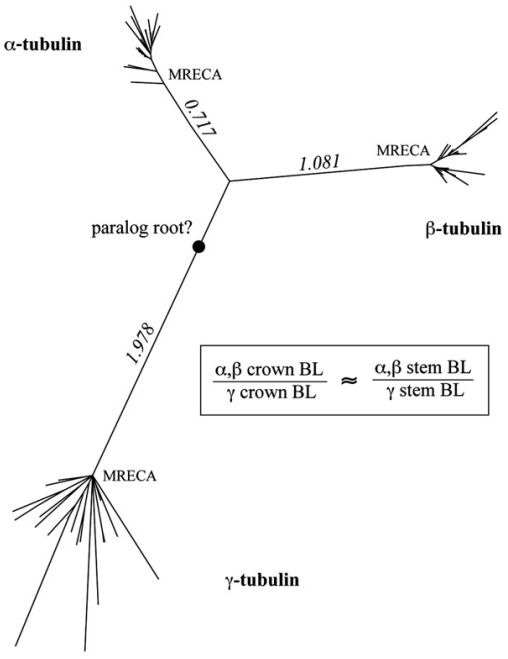


Fig. 1. Maximum-likelihood phylogeny of tubulin proteins. Estimated evolutionary rates within the eukaryal crown group for each tubulin paralog are proportional to branch lengths for each corresponding paralog stem, suggesting a continuous gradual evolution of paralogs within a proto-eukaryal lineage. Numbers on stem branches represent branch lengths (average substitutions/site). Bayesian phylogenetic analysis (not shown) produced a topology with highly similar branch lengths. MRECA refers in each case to "Most Recent Eukaryal Common Ancestor". Putative midpoint rooting agrees with root placement based on paralog outgroups (data not shown).

to fit a more recent eukaryal origin. In that alternative scenario, stem branch lengths would primarily be a factor of the rapid evolution of each paralog after divergence, and should show no correlation to evolutionary rates within each respective crown group, as each would be generated by independent selective and/or neutral processes. Additionally, in such a scenario one would expect asymmetry of stem branch lengths (e.g., between α and β), as the paralog experiencing neofunctionalization and divergence should accumulate more substitutions.

Inferences using branch lengths of deep paralogs are also complicated by site saturation, which will result in an underestimate of the true distance between sequences, and may render such analyses meaningless (Philippe and Forterre, 1999). However, at least in the case of tubulin paralogs, analysis shows that even the long stem branches have undergone only limited site saturation of slow evolving sites, and that many positions remain conserved enough to be phylogenetically informative (Fig. 2). Extending this approach to additional paralogs diverging within the eukaryal stem is necessary to validate this methodology, and to determine if it provides consistent robust support for the Archaeozoan hypothesis.

3.2. HGT, rooting and composition

Many methods for rooting universal trees assume an overall phylogenetic topology that will clearly support at least one of the proposed hypotheses. However, in the case of extensive horizontal transfer, reconstructing the series of events relating the organismal tree to the gene tree can be nontrivial (Gogarten and Townsend, 2005). For example, the phylogeny of tyrosyl-tRNA synthetase (TyrRS) including representatives from all three domains (Fig. 3) contains several groups inconsistent with both three domain and Eocyte trees, which can only be explained by a series of HGT events (see discussion below).

In such cases, identifying the location of the root can greatly assist in determining the most probable set of transfer events. As described in (Fournier and Gogarten, 2010) this can be accomplished by identifying the internal branch with a signal of amino acid composition closest to that determined to be a remnant signal of a more primitive genetic code. Ancestral sequence reconstruction performed on a concatenation of 29 universally conserved ribosomal proteins revealed the bacterial stem to have the greatest non-physiological compositional bias, containing an excess of primordial amino acids (e.g., Gly, Ala, Asn), and an under-representation of more recent additions (e.g., Trp, Tyr, Cys, Phe) (Fig. 4). Since physiology also impacts amino acid usage, this same approach can be used to polarize transfers between groups. For example, in the case of the subset of Haloarchaea which appear as a sister group to the opisthokonts in the TyrRS phylogeny, an excess of negatively charged residues (Glu, Asp) at their ancestral node would identify Haloarchaea as the donor, and the opisthokonts as the recipient, as an over-abundance of these amino acids is strongly correlated with a halophilic physiology. Compositional polarizing of transfer events can also be used for other characters that leave a signal in amino acid composition, such as thermophily or substantial differences in genomic G + C content.

Aside from compositional and phylogenetic analyses, ribosomal proteins are also useful for constructing archaeal phylogeny via presence/absence studies. As these proteins are rarely transferred between organisms (Sorek et al., 2007) and are even less frequently gained or lost along lineages, the few events that can be identified may make excellent phylogenetic markers. A substantial number of ribosomal proteins (~30) are found only within both Eukarya and Archaea. Of these, ribosomal proteins S25e,

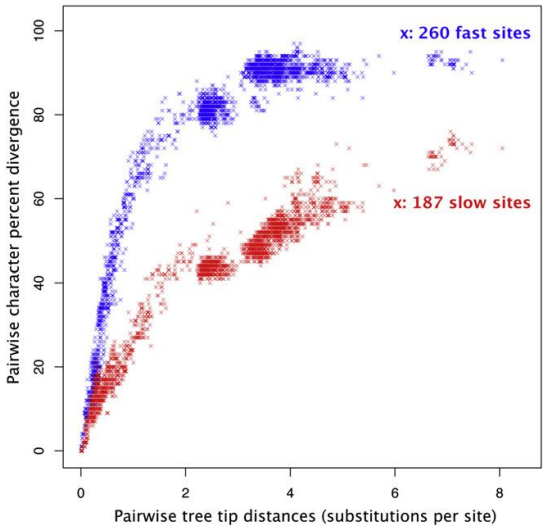


Fig. 2. Site saturation within tubulin protein sequences. The scatter plot compares % sequence divergence in slow and fast sites to inferred phylogenetic distance. The plot reveals that while fast evolving sites are in saturation, slow sites are not. The multiple sequence alignment was divided into fast and slow evolving sites. Slow evolving sites contained 7 or fewer different amino acids in the alignment column (260 sites). The fast sites had 8 or more amino acids per alignment column (187 sites). Alignment columns with more than 50% gaps were not considered in the analysis; gaps were treated as missing data. Phylogenetic distances depicted on the *x*-axes were inferred using phylobayes (Lartillot et al., 2009).

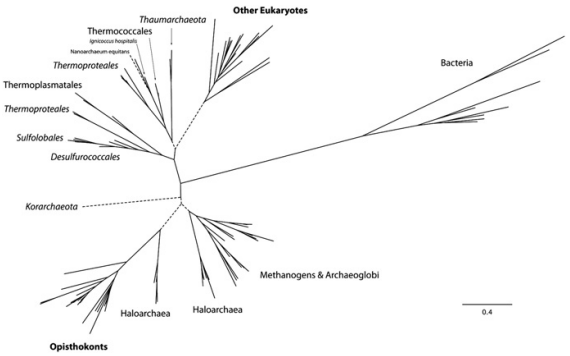


Fig. 3. Bayesian phylogeny of tyrosyl-tRNA synthetases. A complex mixed distribution of eukaryal, crenarchaeal, and euryarchaeal clades inconsistent with either three domain or eocyte tree topologies suggests a history of extensive horizontal gene transfer. Eukaryal, crenarchaeal, and euryarchaeal groups are indicated with bold, italic, and regular fonts, respectively. Dotted lines represent groups supported with less than 70% posterior probability.

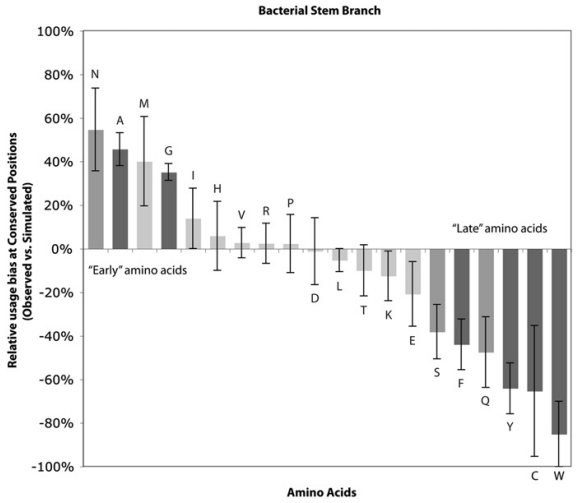


Fig. 4. Compositional signature of a primitive genetic code on the bacterial stem branch. Dark grey indicates amino acids for which significant bias was detected in agreement with a consensus of models of code evolution. Medium grey indicates significant bias in amino acids with ambiguous or conflicting placements in models of code evolution. Light grey indicates amino acids with no significant bias.

S26e, S30e, and L13e are absent within Euryarchaea. Additionally, L14e and L30e are absent within several euryarchaeal orders (Table 1). These proteins show an especially interesting absence pattern among deep archaeal lineages, with Euryarchaea-absent proteins always being present within Korarchaeota and Thaumarchaeota (with the exception of L13e within Thaumarchaeota), and always absent within Nanoarchaeota (with the exception of S16e). Phylogenetic analyses of each of these proteins (data not shown) places these groups on deep unresolved long branches, making it unlikely that the additional presence of L13e or S16e is due to a recent HGT following an earlier loss. The most parsimonious explanation assuming a three-domain rooting of the Tree of Life would therefore place the Nanoarchaeota as a deep euryarchaeal lineage sharing in an ancestral loss of S25e, S30e, and L13e, with the additional loss of S26e occurring after their divergence. Korarchaeota is more likely to be a deep branching crenarchaeal group. Thaumarchaeota appears to have a mixed set of characters, with an absence of L13e resembling Nanoarchaota/Euryarchaeota, and a presence of S25e and S30e resembling Korarchaeota/Crenarchaeota; this suggests that this lineage has undergone independent losses of at least one of these genes, and its placement cannot be inferred.

Table 1
Presence (P)/Absence (A) of Ribosomal Proteins within Archaea.

	Crenarch-aeota	Korarch-aeota	Thaumarch-aeota	Nanoarch-aeota	Euryarch-aeota*
S25e	P	P	P	A	A
S26e	P	P	P	P	A
S30e	P	P	P	A	A
L13e	P	P	A	A	A
L14e	P	P	A	P	Archaeoglobales Thermoplasmatales Halobacteriales Methanosarcinales Methanomicrobiales
L30e	P	P	P	P	Thermoplasmatales Halobacteriales

(* Absences within euryarchaeal orders are listed).

Assuming a tree with an Eocyte topology, it is no longer necessary to presume that these proteins are aboriginal within Archaea; they could have arisen on the branch leading to the Crenarchaeota/Eukarya. In this case, the pattern of presence/absence within Nanoarchaeota and Thaumarchaeota can be explained by either independent losses or gains via HGT, depending on their location in the tree. However, this pattern is not sufficient to distinguish between these scenarios.

Ribosomal proteins L14e and L30e clearly illustrate that unambiguous lineage-specific losses of these protein families do occur within Archaea. Given the current understanding of euryarchaeal phylogeny (Bapteste et al., 2005; Gribaldo and Brochier-Armanet, 2006), the pattern of absences in L14e can be explained by a single loss following the divergence of basal methanogenic lineages (Methanopyrales, Methanobacteriales, Methanococcales), while L30e would require independent losses within the Halobacteriales and Thermoplasmatales.

3.3. Horizontal gene transfer as a tool in reconstructing the Net of life

Horizontally transferred genes that are maintained in the recipient lineage provide characters useful in phylogenetic reconstruction (Huang and Gogarten, 2006). In addition, these transfers also can be used to correlate and date evolutionary events that occurred in different parts of the Tree/Web of Life. The choice of reference phylogeny onto which reticulation events can be mapped in reconstructing the reticulate history of life remains an open question (Swithers et al., 2009a). We do not consider the average phylogenetic signal retained in genomes as useful, because it neither reflects organismal evolutionary history, nor does it necessarily reflect the history of any individual gene (Gogarten and Townsend, 2005). Use of the ribosomal phylogeny as backbone appears more appropriate because ribosomal components are only infrequently transferred between divergent organisms (Sorek et al., 2007). One complication in using the ribosomal phylogeny as backbone is that it cannot be rooted using ancient paralogs; however, the echo from the early expansion of the genetic code that is found in ribosomal proteins can be used to root the ribosomal Tree of Life directly (see discussion above and (Fournier and Gogarten, 2010)).

One transfer that can be used to inform on the relative timing of evolutionary events in the different domains of life is that of tyrosyl-tRNA synthetase (tyrRS) from within the Haloarchaea to the ancestor of the opisthokonts (i.e. the clade formed by animals and fungi) (Huang et al., 2005; Woese et al., 2000). This transfer confirms the fungi and animals as sister kingdoms, further supports microsporidia as a derived rather than basal lineage, and demonstrates that at the time of the opisthokont ancestor, the Haloarchaea had already diverged from other

euryarchaeotes. While more recent phylogenetic analyses suggest that this HGT may actually be secondary to a previous HGT from a deep archaeal lineage to a haloarchaeal ancestor (Fig. 3), the aforementioned inferences remain valid.

In a similar way, the transfer of over 50 genes from within the Chlamydiae to the ancestor of the archaeplastida (plants, green and red algae, and glaucocystophytes) (Huang and Gogarten, 2007, 2008; Becker et al., 2008; Moustafa et al., 2008) confirms the archaeplastida as a monophyletic group, and reveals that at the time the primary plastid was established, Chlamydiae already had split into the groups containing the genera Parachlamydia and Chlamydia, respectively. These transfers reveal that the Archaea and Bacteria were already diversified into different genera before at least two major eukaryal kingdoms diverged, and likely close to the time of radiation for all known eukaryal kingdoms.

In some cases, HGT can also be used to infer an absolute dating of evolutionary events. The transfer of two genes encoding enzymes that allow for the use of acetate as substrate in methanogenesis from within the cellulolytic clostridia to Methanosarcinacaea (Fournier and Gogarten, 2008) suggests that this pathway for acetoclastic methanogeneses evolved relatively recently. As this group likely diversified only after land plants had evolved and produced a cellulose-rich substrate in freshwater environments, these genes were probably transferred no earlier than the Mid-Ordovician (about 475 Mya).

4. Conclusion

With recent phylogenetic analyses utilizing compositional heterogeneity raising new doubts about the monophyly of the Archaea, methods free from potential tree reconstruction artifacts and inconclusive conflicting signals are needed for determining the most accurate scenario for archaeal evolutionary history. To this end, we propose utilizing ancient horizontal gene transfers as synchronizing events between groups of distantly related organisms. Given enough transfer events, this approach will produce a chronological ordering of the emergence of various archaeal and eukaryal clades, supporting the model of archaeal evolution with which the best reconciliation can be made. Such reconciliation must also take into account molecular evolution within the stem leading to the eukaryal domain, as the hypotheses concerning eukaryogenesis are too intimately related to archaeal evolution to be considered independently.

Finally, a novel, recently devised method for rooting the ribosomal Tree of Life utilizing an expected unique bias in amino acid usages on the branch closest to the origin of the genetic code unambiguously locates the root on the branch leading to the bacterial domain. This result is in agreement with previous analyses using reciprocal rooting of ancient gene paralogs, and joins them in refuting the placement of the root within the Archaea under some versions of the Eocyte model (Lake, 1987). While generally associated with three monophyletic domains, this “traditional” placement of the root is also compatible with an archaebacterial origin of Eukarya within a paraphyletic archaeal clade in a more conservative version of the Eocyte model (Cox et al., 2008; Foster et al., 2009; Rivera and Lake, 2004).

Acknowledgements

This work was supported through grants from the NASA Exobiology (NNX07AK15G) and NSF Assembling the Tree of Life (DEB 0830024) Programs to JPG, and an appointment from the NASA Postdoctoral Program to GPF at the Massachusetts Institute of Technology.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Ashlock, P.D., 1971. Monophyly and associated terms. Syst. Zool. 20, 63–69.

Baldauf, S.L., Palmer, J.D., Doolittle, W.F., 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. U S A 93, 7749–7754.

Bapteste, E., Brochier, C., Boucher, Y., 2005. Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. Archaea 1, 353–363.

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

Brown, J., Doolittle, W., 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc. Natl. Acad. Sci. USA 92, 2441–2445.

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

Cavalier-Smith, T., 2006a. Cell evolution and Earth history: stasis and revolution. Philos. Trans. R Soc. Lond B. Biol. Sci. 361, 969–1006.

Cavalier-Smith, T., 2006b. Rooting the tree of life by transition analyses. Biol. Direct 1, 19.

Cavalier-Smith, T., 2010. Deep phylogeny, ancestral groups and the four ages of life. Philos. Trans. R Soc. Lond B. Biol. Sci. 365, 111–132.

Cox, C.J., Foster, P.G., Hirt, R.P., Harris, S.R., Embley, T.M., 2008. The archaebacterial origin of eukaryotes. Proc. Natl. Acad. Sci. U S A 105, 20356–20361.

Di Giulio, M., 2007. The tree of life might be rooted in the branch leading to Nanoarchaeota. Gene 401, 108–113.

Doolittle, W.F., 1996. Some Aspects of the Biology of Cells and Their Possible Significance. In: Roberts, D.M., Sharp, P., Alderson, G., Collins, M.A. (Eds.), Symposium of the Society for General Microbiology; Evolution of Microbial. Cambridge University Press, Cambridge, pp. 1–21.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32, 1792–1797.

Forterre, P., Philippe, H., 1999. Where is the root of the universal tree of life? Bioessays 21, 871–879.

Foster, P.G., Cox, C.J., Embley, T.M., 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos. Trans. R Soc. Lond B. Biol. Sci. 364, 2197–2207.

Fournier, G.P., Gogarten, J.P., 2008. Evolution of acetoclastic methanogenesis in Methanosarcina via horizontal gene transfer from cellulolytic Clostridia. J. Bacteriol. 190, 1124–1127.

Fournier, G.P., Gogarten, J.P., 2010. Rooting the ribosomal tree of life. Mol. Biol. Evol. 27, 1792–1801.

Glansdorff, N., Xu, Y., Labedan, B., 2008. The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. Biol. Direct 3, 29.

Gogarten, J.P., Doolittle, W.F., Lawrence, J.G., 2002. Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. 19, 2226–2238.

Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J., Bowman, B.J., Manolson, M.F., Poole, R.J., et al., 1989. Evolution of the vacuolar H+-

ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. U S A 86, 6661–6665.
Gogarten, J.P., Taiz, L., 1992. Evolution of proton pumping ATPases: rooting the tree of life. Photosynthesis Res. 33, 137–146.
Gogarten, J.P., Townsend, J.P., 2005. Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3, 679–687.
Gray, M.W., 1993. Origin and evolution of organelle genomes. Curr. Opin. Genet. Dev. 3, 884–890.
Gribaldo, S., Brochier-Armanet, C., 2006. The origin and evolution of Archaea: a state of the art. Philos. Trans. R Soc. Lond B. Biol. Sci. 361, 1007–1022.
Gribaldo, S., Cammarano, P., 1998. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. J. Mol. Evol. 47, 508–516.
Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.
Gupta, R.S., 2001. The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. Int. Microbiol. 4, 187–202.
Hampl, V., Silberman, J.D., Stechmann, A., Diaz-Trivino, S., Johnson, P.J., Roger, A.J., 2008. Genetic evidence for a mitochondriate ancestry in the 'amitochondriate' flagellate Trimastix pyriformis. PLoS ONE 3, e1383.
Hartman, H., 1984. The origin of the eukaryotic cell. Speculations Sci. Technol. 7, 77–81.
Hennig, W., 1966. Phylogenetic Systematics. University of Illinois Press, Urbana.
Horner, D.S., Hirt, R.P., Kilvington, S., Lloyd, D., Embley, T.M., 1996. Molecular data suggest an early acquisition of the mitochondrion endosymbiont. Proc. Biol. Sci. 263, 1053–1059.
Huang, J., Gogarten, J.P., 2006. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends Genet. 22, 361–366.
Huang, J., Gogarten, J.P., 2008. Concerted gene recruitment in early plant evolution. Genome Biol. 9, R109.
Huang, J., Gogarten, P., 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biol. 8, R99.
Huang, J., Xu, Y., Gogarten, J.P., 2005. The presence of a haloarchaeal type tyrosyl-tRNA synthetase marks the opisthokonts as monophyletic. Mol. Biol. Evol. 22, 2142–2146.
Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., Miyata, T., 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. U S A 86, 9355–9359.
Lake, J.A., 1987. Prokaryotes and archaebacteria are not monophyletic: rate invariant analysis of rRNA genes indicates that eukaryotes and eocytes form a monophyletic taxon. Cold Spring Harb Symp. Quant Biol. 52, 839–846.
Lake, J.A., 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature 331, 184–186.
Lake, J.A., Clark, M.W., Henderson, E., Fay, S.P., Oakes, M., Scheinman, A., Thornber, J.P., Mah, R.A., 1985. Eubacteria, halobacteria, and the origin of photosynthesis: the photocytes. Proc. Natl. Acad. Sci. U S A 82, 3716–3720.
Lake, J.A., Servin, J.A., Herbold, C.W., Skophammer, R.G., 2008. Evidence for a new root of the tree of life. Syst. Biol. 57, 835–843.
Lake, J.A., Skophammer, R.G., Herbold, C.W., Servin, J.A., 2009. Genome beginnings: rooting the tree of life. Philos. Trans. R Soc. Lond B. Biol. Sci. 364, 2177–2185.
Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., et al., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.
Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.
Margulis, L., 2009. Genome acquisition in horizontal gene transfer: symbiogenesis and macromolecular sequence analysis. Methods Mol. Biol. 532, 181–191.
Martin, W., Koonin, E.V., 2006. Introns and the origin of nucleus-cytosol compartmentalization. Nature 440, 41–45.
Martin, W., Muller, M., 1998. The hydrogen hypothesis for the first eukaryote. Nature 392, 37–41.
Moustafa, A., Reyes-Prieto, A., Bhattacharya, D., 2008. Chlamydiae has contributed at least 55 genes to Plantae with predominantly plastid functions. PLoS ONE 3, e2205.
Philippe, H., Forterre, P., 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. 49, 509–523.
Poole, A., Jeffares, D., Penny, D., 1999. Early evolution: prokaryotes, the new kids on the block. Bioessays 21, 880–889.
Poole, A.M., Penny, D., 2007. Evaluating hypotheses for the origin of eukaryotes. Bioessays 29, 74–84.
Rambaut, A., 2007. Figtree. http://tree.bio.ed.ac.uk/software/figtree/.
Rivera, M.C., Lake, J.A., 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431, 152–155.
Sapp, J., 2006. Two faces of the prokaryote concept. Int. Microbiol. 9, 163–172.
Searcy, D.G., 1992. Origins of mitochondria and chloroplasts from sulfur based symbiosis. In: Hartman, H., Matsuno, K. (Eds.), The Origin and Evolution of the Cell. World Scientific, pp. 47–78.
Skophammer, R.G., Servin, J.A., Herbold, C.W., Lake, J.A., 2007. Evidence for a gram-positive, eubacterial root of the tree of life. Mol. Biol. Evol. 24, 1761–1768.
Sogin, M.L., 1991. Early evolution and the origin of eukaryotes. Curr. Opin. Genet. Dev. 1, 457–463.
Sorek, R., Zhu, Y., Creevey, C.J., Francino, M.P., Bork, P., Rubin, E.M., 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318, 1449–1452.
Swithers, K.S., Gogarten, J.P., Fournier, G.P., 2009a. Trees in the web of life. J. Biol. 8, 54.
Swithers, K.S., Senejani, A.G., Fournier, G.P., Gogarten, J.P., 2009b. Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. BMC Evol. Biol. 9, 303.
Tourasse, N.J., Gouy, M., 1999. Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced from rRNA and protein-coding genes. Mol. Phylogenet. Evol. 13, 159–168.
Woese, C.R., Fox, G.E., 1977a. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U S A 74, 5088–5090.
Woese, C.R., Fox, G.E., 1977b. The concept of cellular evolution. J. Mol. Evol. 10, 1–6.
Woese, C.R., Gupta, R., 1981. Are archaebacteria merely derived 'prokaryotes'? Nature 289, 95–96.
Woese, C.R., Magrum, L.J., Fox, G.E., 1978. Archaebacteria. J. Mol. Evol. 11, 245–251.
Woese, C.R., Olsen, G.J., Ibba, M., Soll, D., 2000. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol. Mol. Biol. Rev. 64, 202–236.
Yutin, N., Makarova, K.S., Mekhedov, S.L., Wolf, Y.I., Koonin, E.V., 2008. The deep archaeal roots of eukaryotes. Mol. Biol. Evol. 25, 1619–1630.
Zhaxybayeva, O., Gogarten, J.P., 2004. Cladogenesis, coalescence and the evolution of the three domains of life. Trends Genet. 20, 182–187.
Zhaxybayeva, O., Gogarten, J.P., 2007. Horizontal gene transfer, gene histories and the root of the tree of life. In: Pudritz, R.E., Higgs, P.G., Stone, J. (Eds.), Planetary Systems and the Origins of Life (Cambridge Astrobiology). Cambridge University Press, Cambridge, UK, pp. 178–192.
Zillig, W., Palm, P., Klenk, H.-P., 1992. A model of the early evolution of organisms: the arisal of the three domains of life from the common ancestor. Origin Evol. Cell, 163–182.

4.3. Addendum: Eukaryal Stem Branch Length in Histones

4.3.1. Introduction

Histones are a good candidate protein family for examination when looking for long eukaryal lines of descent. There are 4 histone paralogs, H2A, H2B, H3, H4, common to all Eukarya, with an archaeal histone homolog outgroup (Bell and White, 2010; Thatcher and Gorovsky, 1994). Each of these 4 paralogs can therefore provide 4 independent measures of the eukaryal SBL, in addition to the 3 provided by the tubulins.

If phylogenetic reconstructions show much more divergence in eukaryal histones compared to archaeal histones, and long branches along the eukaryal lines compared to branch lengths within the crown group, then Eukarya are uniquely eukaryal and are at least as ancient as Archaea or have undergone an increased rate of evolution. This will provide a time constraint on the origin of Eukarya and circumstantial evidence in favor of archaeal monophyly, although a polytomy cannot be ruled out. If long branches are not observed, then the circumstantial evidence will be in favor of archaeal paraphyly and the origin of Eukarya from within the Archaea, because short Eukaryal stem branches are expected under this topology.

4.3.2. Methods

Histone amino acid sequences were downloaded from NCBI and T-COFFEE was used to modify taxa names (Notredame et al., 2000). Deepview was used to create a structural alignment of the four histone subunits (Kaplan and Littlejohn, 2001). Profile alignments in Clustal X 2.0 were used to add histone sequences to the structural alignment (Larkin et al., 2007). Jalview was used to manually adjust the histone alignment (Clamp et al., 2004). Clustal X 2.0 was used to

convert alignments to phylip format. Treeview X was used to convert trees to nexus format (Saldanha, 2004). FigTree1.2.2 was used to visualize trees (Rambaut and Drummond, 2008). Archaeal histone sequences were used to root trees.

PhyML was used to produce Maximum Likelihood trees (Guindon et al., 2005), using the WAG substitution model, 100 bootstrap replicates, estimated proportion of invariable sites, 4 categories of substitution rates, and an estimated gamma distribution parameter. MrBayes was run for at least 1000000 generations to estimate Bayesian trees (Huelsenbeck et al., 2001), with the following priors: unconstrained branch lengths, with a mean of 0.1 substitutions per site, 4 categories for the gamma shape with an exponential distribution and a mean of 1.0, and 2 categories of rate substitution. If MrBayes failed to converge by 1000000 generations, additional 250000 generations were run at a time until convergence.

Extreme divergence and short sequences make histone alignments difficult. To judge the quality of the alignment, Maximum Likelihood trees were generated and examined for eukaryal monophyly. Since no one believes that the Eukarya evolved independently more than once or that a Eukaryon has reverted to a prokaryote (Blackstone, 2013), it is universally agreed upon that the Eukarya, and therefore LECA, are monophyletic. LECA is present four times on the histone tree, one for each of the four subunits, and only when all four instances are resolved as monophyletic with high bootstrap support value, will the alignment be accepted.

4.3.3. Results

The histones were easy to align within subunits, but extremely difficult to align between subunits, which was why the structures were used. In order to obtain a high quality alignment, the sequences were first aligned within subunit and then to the structural alignment. The N-

terminus of the H2A sequences did not correctly align with the homologous region in the H2A sequence from the structural alignment, so Jalview was used to manually correct this error.

Figure 1 shows that branches within a histone subunit are relatively long compared with those of archaeal histones, indicating copious evolution in isolation from Archaea along the eukaryal line, which indirectly supports monophyletic Archaea, although the very long branches joining subunits could be due to increased evolution associated with gene duplications (Goodman et al., 1987). The same Maximum Likelihood tree is shown in Figure 1 and 2, unrooted and rooted on the archaeal histones, respectively. Figure 3 shows the same dataset analysed in Mr. Bayes. The Bayesian tree is in complete agreement with the Maximum Likelihood tree when it comes to long eukaryal stem branches.

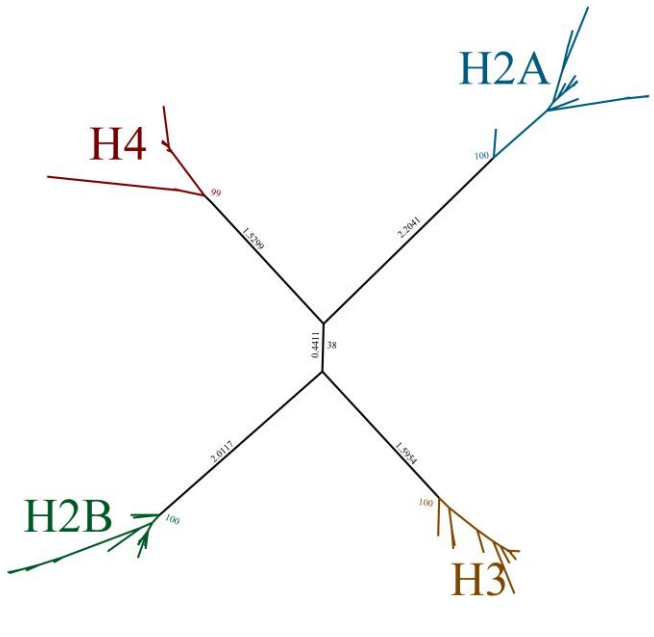


Figure 1: Maximum Likelihood tree after manual adjustment of the structural alignment and histone profile alignment. All four instances of LECA are resolved as monophyletic with high bootstrap support value and all four eukaryal SBLs are long.

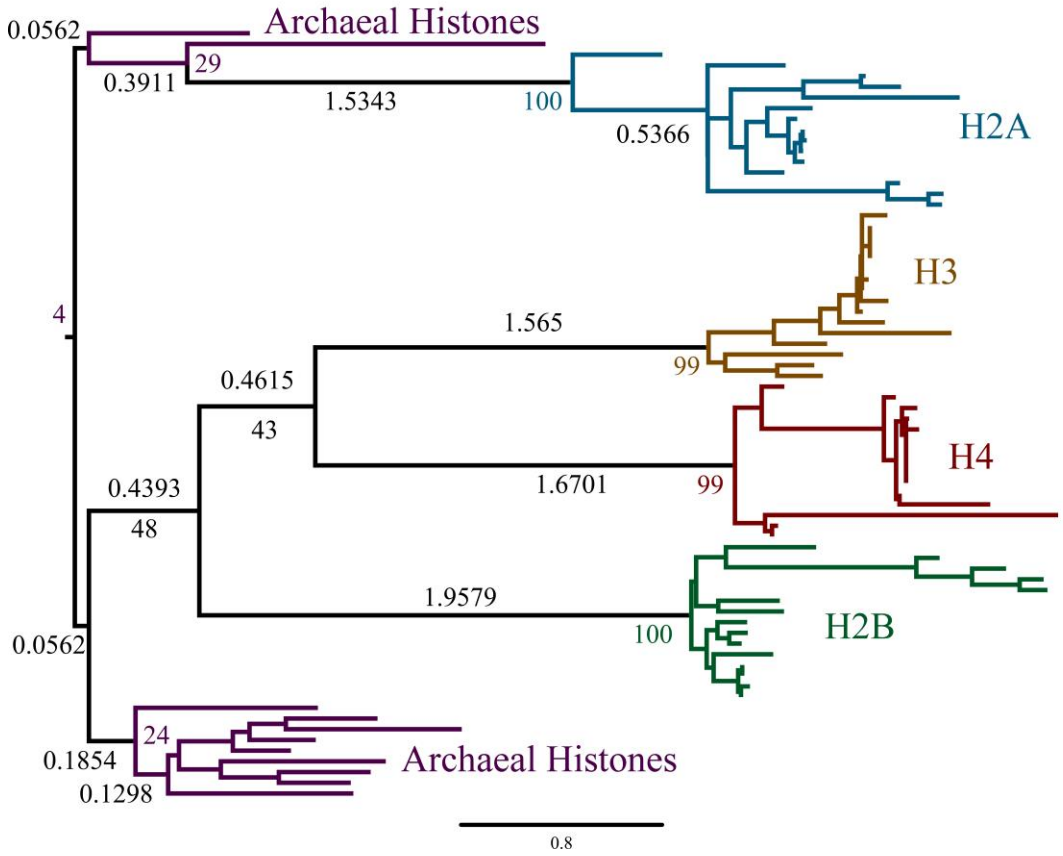


Figure 2: Another view of the Maximum Likelihood tree, with bootstrap values, of the manually adjusted profile structural alignment, this time with the archaeal histones (AH) added. Archaeal histones appear paraphyletic, possibly due to unresolved paralogy. Long branches are exhibited along the eukaryal stem and within the Eukarya.

Figure 3: Bayesian tree of the manually adjusted profile structural alignment with the archaeal histones (AH) added. The posterior probabilities are shown in blue and branch lengths shown in red. The archaeal histone mono or para-phyly is not resolved. As with the Maximum Likelihood tree, long branches are seen along the eukaryal stems and within the eukaryal crowns.

4.3.4. Discussion

All 4 eukaryal histones show longer stem branch lengths than the branch lengths observed within the archaeal histones. All 4 eukaryal histone stem branch lengths are longer than their respective eukaryal crown branch lengths. If the rate of evolution were constant, then this would imply that the duration of the eukaryal stem's existence, from the time the Eukarya split with the Archaea to the time of LECA, when the Eukarya radiated out into its living descendants, is equal to the age of LECA. Since red algae fossils tell us that the time of LECA was prior to 1.2 billion years ago, then the eukaryal stem would be approximately 1.2 billion years old. Summing the two, would indicate that the Eukarya diverged from the Archaea at least 2.4 billion years ago, around the time Archaea first appeared in the fossil record. If this were the case, archaeal paraphyly would be extremely unlikely, because the Archaea would have to be much older than evidence currently supports. This evidence supports the hypothesis of Archaeal monophyly, or possibly a polytomy between Archaea and Eukarya.

Unfortunately branch lengths are the product of both *time* and *evolutionary rate* and there is no way to tease the two apart. The branch lengths in the tree indicate that evolutionary rate is not constant, because the branch lengths within the Archaea are so short. Under any tree topology, except one in which the Archaea evolved from within the Eukarya which there is no evidence to support and no one believes, the sum of the length of the branches in the archaeal crown and the archaeal stem would be expected to have branch lengths at least as long as the corresponding measurement for the Eukarya. But the Archaeal stem is very short, making the total archaeal branch length quite a bit shorter than total eukaryal branch length. Given the three topologies under consideration, I must conclude that the Archaea are evolving more slowly than the Eukarya.

If the Eukarya are evolving faster than the Archaea, then part of the long stem branch lengths seen in Eukarya is likely due to this increase in rate of evolution. Unfortunately, there is no way of knowing how much of it is due to the increase rate of evolution and how much is due to time. The shorter the time, the higher the rate of evolution. But as estimates of divergence time get to the shortest extremes that have been suggested in the literature, the rate of evolution must increase so much as to seem unbelievable. Therefore, at the very least, I argue that scenarios where the Archaea diverged from the Eukarya shortly before the lineages descending from LECA radiated from each other, can be ruled out.

4.3.5. Conclusion

The 4 measurements of long eukaryal stem branch length in the 4 histone subunits are in agreement with the 3 measurements of long eukaryal stem branch length in the 3 tubulin subunits. A long eukaryal stem branch is a common feature in phylogenetic trees (Derelle and Lang, 2012; Fournier et al., 2011). And although variation in rate of evolution prevents placing a time estimate on these long stem branches, they cannot be ignored. This is a piece of the puzzle that needs to be considered when discussing possible scenarios for the origin of the Eukarya.

The extreme divergence within Eukarya indirectly supports the monophyletic Archaea hypothesis or archaeal polytomy. It is possible that the two archaeal kingdoms diverged from each other before the split with Eukarya, but remained a cohesive unit through continual HGT. If that were the case, some genes could show support for monophyletic Archaea, others for paraphyletic Archaea, and they could both be right. It is also possible that these two events occurred so closely in time, and so long ago, that we will never be able to recover with certainty

which split occurred first. But whatever the case, Eukarya have evolved in isolation from Archaea for a long period of time.

Chapter 5: The Watershed of life

5.1. Abstract

Additional data do not seem to be solving the problem of whether the Eukarya evolved from within the Archaea or as sister to the Archaea. In light of this, the problem may not be the data, but how we interpret the data. Here, life is described as a watershed with two types of signals, one from vertical descent and the other from Horizontal Gene Transfer, giving rise to the complex patterns we see in molecular gene trees. This new perspective sheds light on the different signatures of the two signal types, which could lead researchers to tease the two apart sufficiently to answer the question of whether the Archaea are indeed a monophyletic group.

5.2. The Watershed of Life

The problem is whether the Archaea are paraphyletic or monophyletic and the intense debate with equal evidence on both sides. My attempt in Chapter 4 at using stem branch length to resolve the issue is inconclusive, even with positive results, because the nature of branch length is a factor of two variables, time and rate of evolution, that cannot be teased apart (T. H. Jukes and Cantor, 1969). As I have done this work, others have also gathered additional evidence regarding this question, in the form of genes that were formally thought to be specific to Eukarya being found patchily distributed in Archaea. But even though the evidence has increased, there has been no progress towards an answer. We remain just as uncertain as ever and researchers have even asked if we are at a phylogenetic impasse; if we have a question here that cannot ever be answered. This conundrum led me and my advisor, Dr. Gogarten, to consider the nature of cellular life itself. Might the problem not be the data, but how the data is interpreted? Is a simple

bifurcating tree the best representation of cellular life, or is a rooted net of life (Olendzenski and Gogarten, 2009; Williams et al., 2011) the better metaphor?

One of the issues with determining whether the Eukarya group within the Archaea or as sister taxon to the Archaea, is that the splits between the Euryarchaea, Crenarchaea, and Eukarya happened with relatively few molecular changes to support them (Rochette et al., 2014). In addition, these splits likely occurred over two billion years ago, allowing ample time to erode the phylogenetic signal. Add to that the web of biased HGT (Andam et al., 2010b; Lester et al., 2006) and the picture is very muddled. Perhaps reexamining the existing evidence in light of a new model, that of the Watershed of Life, will shed new light on the problem. The two alternative hypotheses are depicted in Figure 1 Panels A and B: a monophyletic and a paraphyletic Archaea. Both result in data that produce similarly muddled messes as scientists are currently observing with regards to the Archaea.

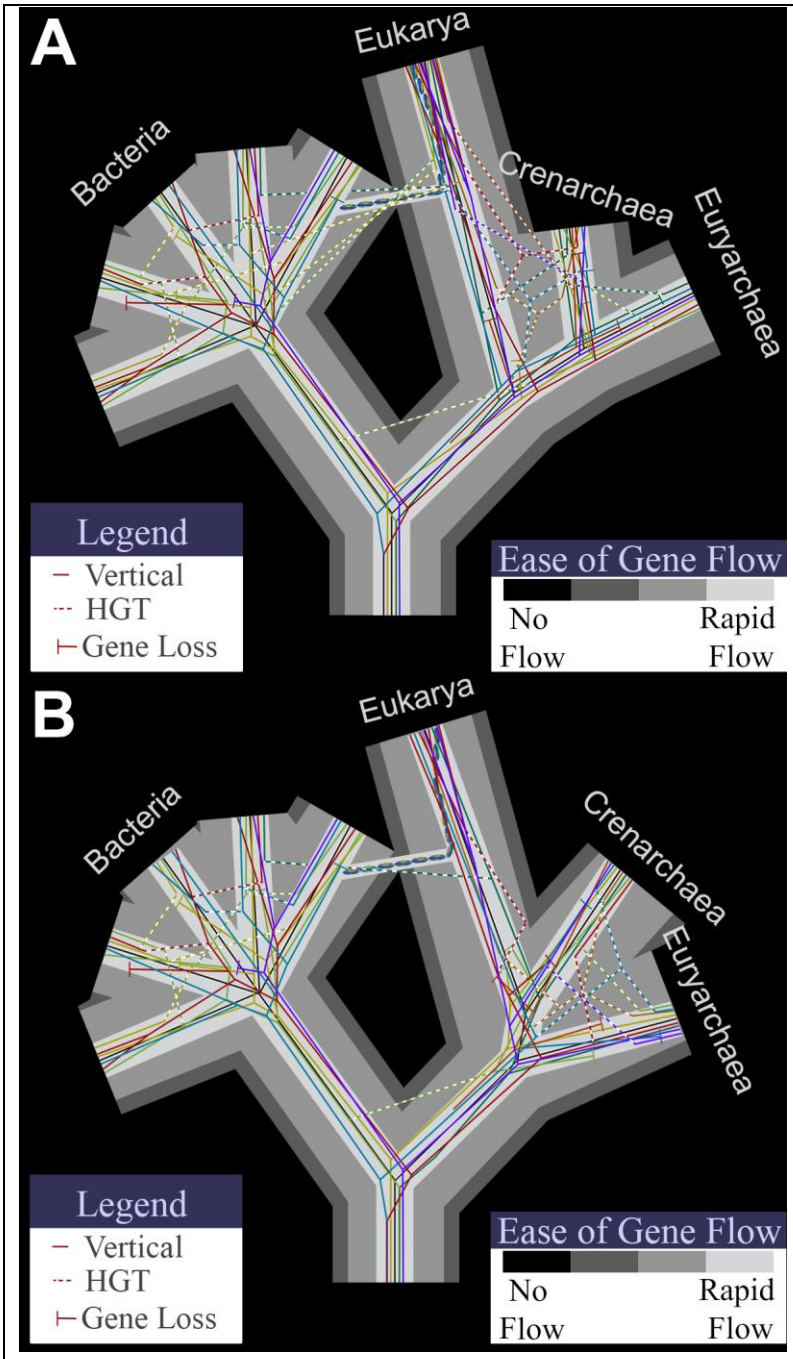


Figure 1: The underlying webs of HGT in Figure 1 Panels A and B are depicted as a watershed that runs underneath the tree of vertical descent. Each set of colored lines represents a gene family. The shades of grey represent the underlying watershed of HGT.

In both Panels A and B of Figure 1, there is an underlying mostly, but not strictly, bifurcating tree that depicts the unbroken line of vertical descent of cellular life. The lightest shade of grey in Panel A and B of Figure 1 represents the taxon boundaries and can be thought of as similar to the species boundaries with multiple genes (represented as colored lines, coexisting within each species) except that the focus of evolution is zoomed out to the various kingdoms or domains. The kingdoms of Eukarya are not individualized, because current research suggests that LECA was a complex modern Eukaryeon (Koumandou et al., 2013), so the core features of the Eukarya evolved on the stem eukaryal branch. The kingdoms of Bacteria are simplified into a polytomy comprising five kingdoms to depict the current consensus that Bacteria underwent a rapid radiation at the base of the domain (Hori and Osawa, 1987); the actual number of bacterial kingdoms and their phylogenetic relationship to each other matters little in this case, because we are interested in the evolution of the Ibisii.

The watersheds of HGT are shown as multi-dimensional surfaces, depicted as varying shades of grey. The lightest shade can be thought of as the deepest region of this surface, where genes flow the most freely, like water flows most swiftly through deeper canals. This represents genes passing from ancestor to descendant in a vertical manner. The second lightest shade of grey represents the second deepest region of the surface, where genes flow quickly due to highways of HGT. These canals link kingdoms that more readily share genes. The next shade of grey represents a shallower region of the canal, where genes can flow, but only rarely. HGT, represented by dashed lines, occasionally reaches across vast expanses, represented in these figures by the fact that only once does a gene cross the darker region of grey. The black represents dry land, where genes cannot flow and are thus not exchanged. The challenge with

this multi-dimensional watershed of HGT is that it can change over time. For example, the acquisition of the mitochondria by the Eukarya, represented by the channel connecting Eukarya and Bacteria, made HGT from Bacteria to Eukarya more conducive, resulting in additional gene transfer events traveling through the channel. The sum of the watershed of HGT through time is what is important in fully characterizing HGT. Unfortunately, this integral is difficult or even impossible to calculate, obscuring our ability to subtract it out from the big evolutionary picture. That is why parsing out the vertical signal in trees is so difficult.

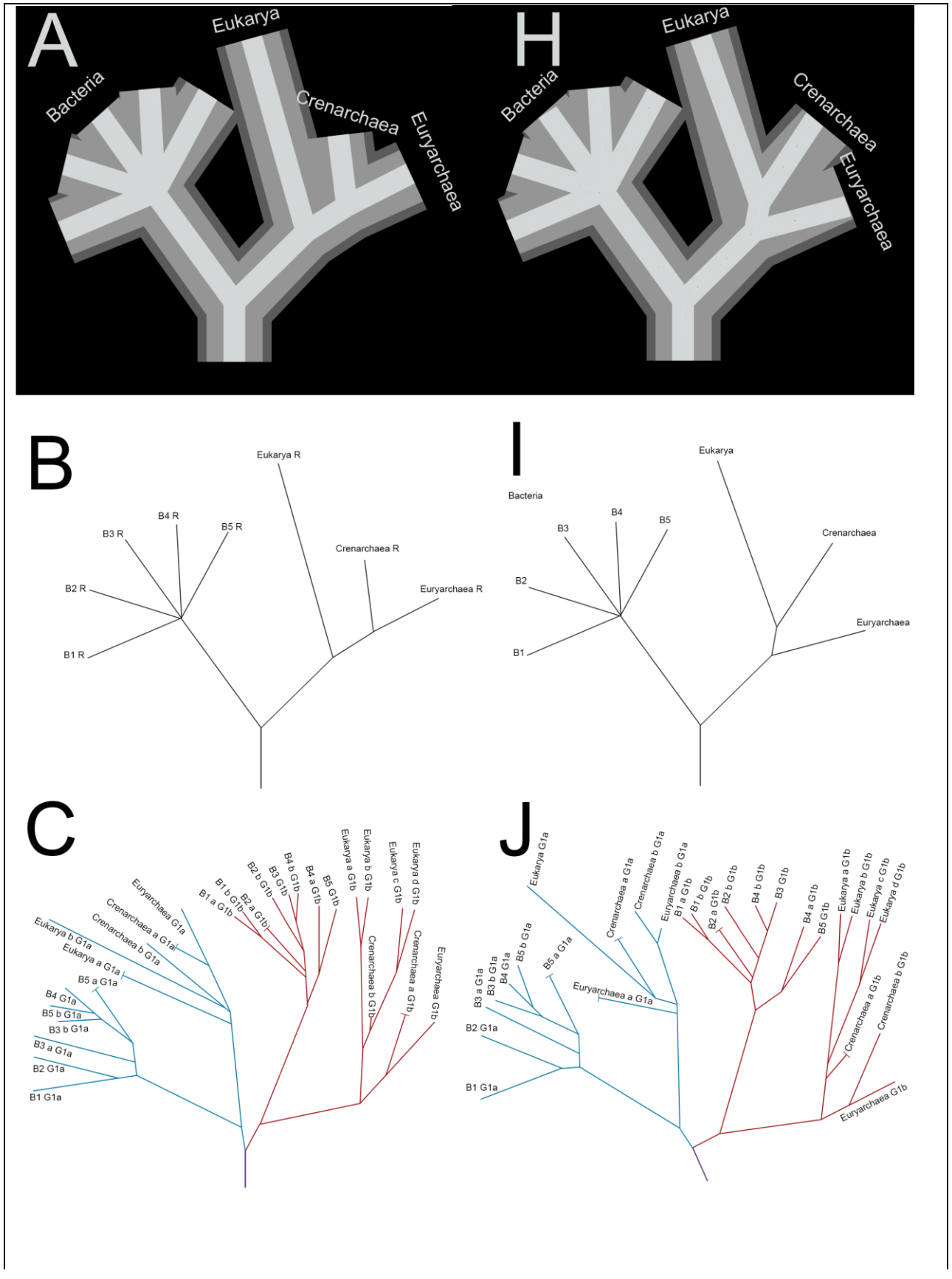
Everything within the lightest shade, or the quickest flowing canal, represents vertical descent and is left out of Panels A and H of Figure 2, which depict only the underlying watershed of HGT. HGT tends to reinforce cellular ancestral relationships, because homologous gene replacement occurs more frequently between close relatives (Andam et al., 2010b). Therefore, the waterways are deepest surrounding the lines of the species tree. In addition to biased HGT, there are also highways of gene sharing, creating links of rapidly flowing genes between parts of the tree that are connected for reasons other than recent shared ancestry (Beiko et al., 2005), where the underlying web of HGT does not closely mimic the cellular tree.

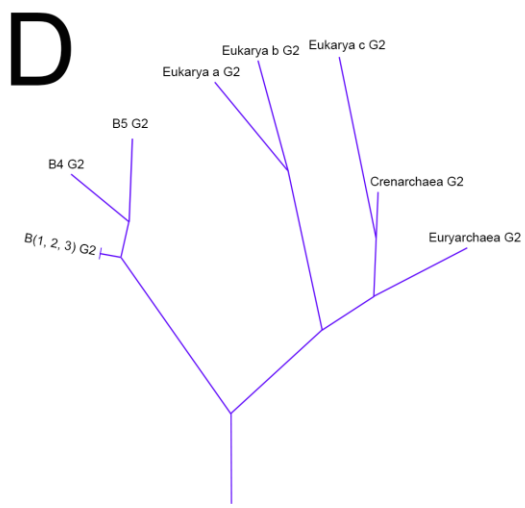
The ribosomal phylogeny is often assumed to represent only the vertical signal in the data. Therefore, a hypothetical ribosomal phylogeny, shown by the black line, is used in Figure 1 Panels A and B to represent the purely vertical signal. This signal is isolated in Figure 2 Panels B and I. Together with the underlying watershed of HGT shown in Figure 2 Panel B, they make up total evolutionary history of all the genes in the tree of life.

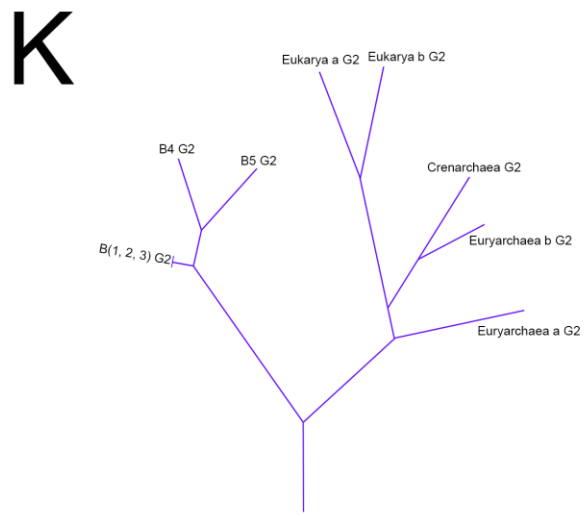
Panels A and B of Figure 1 show the two possibilities for how the tree of vertical descent and the underlying web of HGT interact to produce the molecular patterns that can be observed

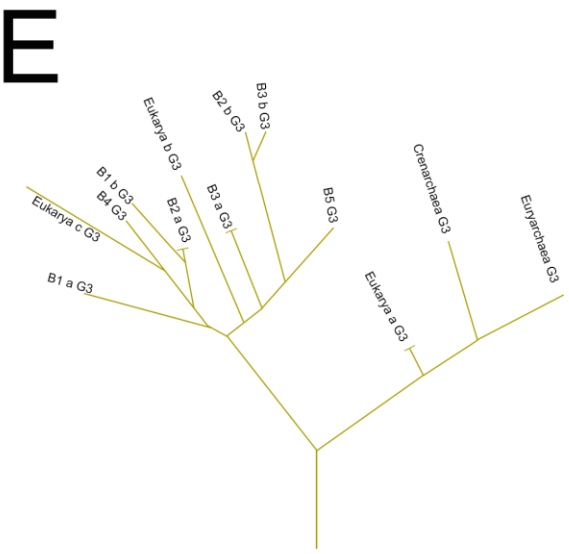
in cellular life today. In both possibilities, vertical descent and HGT play significant roles in the history of cellular evolution. Figure 1 Panel A, represents the monophyletic, or 3-domain, hypothesis, in which all three domains are monophyletic and the paraphyletic signal is the result of HGT. Figure 1 Panel B, represents the paraphyletic Archaea, the so called “2 Domain” (but really 4 or more), or Eocyte (Cox et al., 2008; Poole and Neumann, 2011) hypothesis, in which the Eukarya evolve from within the Archaea and the signal linking Crenarchaea and Euryarchaea is a result of HGT and of sympleisomorphies inherited from their shared ancestor, but lost in Eukarya.

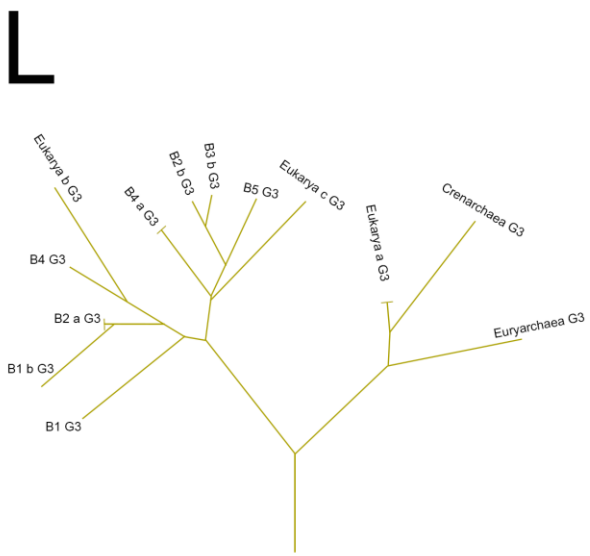
Each of Panels C through N of Figure 2 depict the molecular history of one gene found at the base of the cellular tree, pre-LUCA. These trees represent the types of phenomenon seen in molecular data, such as the ancient duplication of Gene 1 in Panels C and J of Figure 2 that results in all living species having both version a and version b. Recent duplication by itself results in multiple copies of the gene with a similar history being found in living species, such as G1b version a and b found in Eukarya in Figure 2 Panel C, and G2 versions a and b found in Eukarya in Figure 2 Panels D and K. Gene duplication and differential loss or incomplete lineage sorting can be seen in the Bacteria with Gene 3 in Figure 2 Panels E and L, resulting in a gene history that does not directly correspond to the species tree.

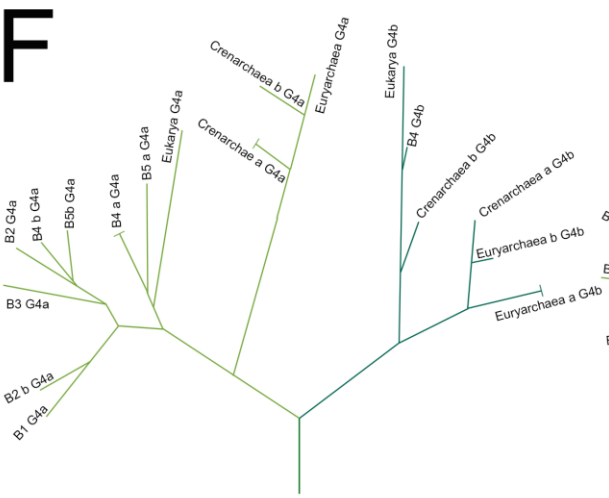

A
Eukarya
Bacteria
Crenarchaea
Euryarchaea
H
Eukarya
Bacteria
Crenarchaea
Euryarchaea
B
Eukarya R
B4 R
B3 R
B5 R
Crenarchaea R
B2 R
Euryarchaea R
B1 R
I
Eukarya
Bacteria
B4
B3
B5
Crenarchaea
B2
Euryarchaea
B1
C
Euryarchaea G1a
Crenarchaea a G1a
Crenarchaea b G1a
Eukarya b G1a
Eukarya a G1a
B5 a G1a
B4 G1a
B5 b G1a
B3 b G1a
B3 a G1a
B2 G1a
B1 G1a
B1 b G1b
B1 a G1b
B2 b G1b
B2 a G1b
B3 G1b
B4 b G1b
B4 a G1b
B5 G1b
Eukarya a G1b
Eukarya b G1b
Eukarya c G1b
Eukarya d G1b
Crenarchaea b G1b
Crenarchaea a G1b
Euryarchaea G1b
J
Eukarya G1a
Crenarchaea a G1a
Crenarchaea b G1a
Euryarchaea b G1a
B1 a G1b
B1 b G1b
B2 a G1b
B2 b G1b
B4 b G1b
B3 G1b
B4 a G1b
B5 G1b
Eukarya a G1b
Eukarya b G1b
Eukarya c G1b
Eukarya d G1b
B3 a G1a
B3 b G1a
B4 G1a
B5 b G1a
B5 a G1a
Euryarchaea a G1a
B2 G1a
B1 G1a
Crenarchaea a G1b
Crenarchaea b G1b
Euryarchaea G1b

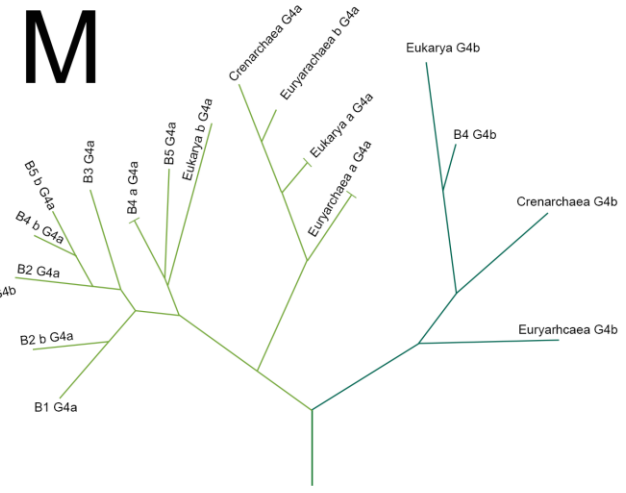

D
Eukarya b G2
Eukarya c G2
Eukarya a G2
B5 G2
B4 G2
Crenarchaea G2
B(1, 2, 3) G2
Euryarchaea G2


K
Eukarya a G2
Eukarya b G2
B4 G2
B5 G2
Crenarchaea G2
B(1, 2, 3) G2
Euryarchaea b G2
Euryarchaea a G2


E
Eukarya b G3
B2 b G3
B3 b G3
B1 b G3
B4 G3
B3 a G3
B2 a G3
Eukarya c G3
B5 G3
Crenarchaea G3
Euryarchaea G3
B1 a G3
Eukarya a G3


L
Eukarya b G3
B2 b G3
B3 b G3
B4 a G3
B5 G3
Eukarya c G3
Crenarchaea G3
Eukarya a G3
B4 G3
B2 a G3
Euryarchaea G3
B1 b G3
B1 G3


F
Crenarchaea b G4a
Euryarchaea G4a
Eukarya G4b
B2 G4a
B4 b G4a
B5b G4a
B4 a G4a
B5 a G4a
Eukarya G4a
Crenarchae a G4a
B4 G4b
Crenarchaea b G4b
Crenarchaea a G4b
Euryarchaea b G4b
B3 G4a
Euryarchaea a G4b
B2 b G4a
B1 G4a


M
Crenarchaea G4a
Euryarchaea b G4a
Eukarya G4b
B5 b G4a
B3 G4a
B5 G4a
Eukarya b G4a
B4 a G4a
Eukarya a G4a
B4 G4b
B4 b G4a
Euryarchaea a G4a
Crenarchaea G4b
B2 G4a
B2 b G4a
Euryarhcaea G4b
B1 G4a

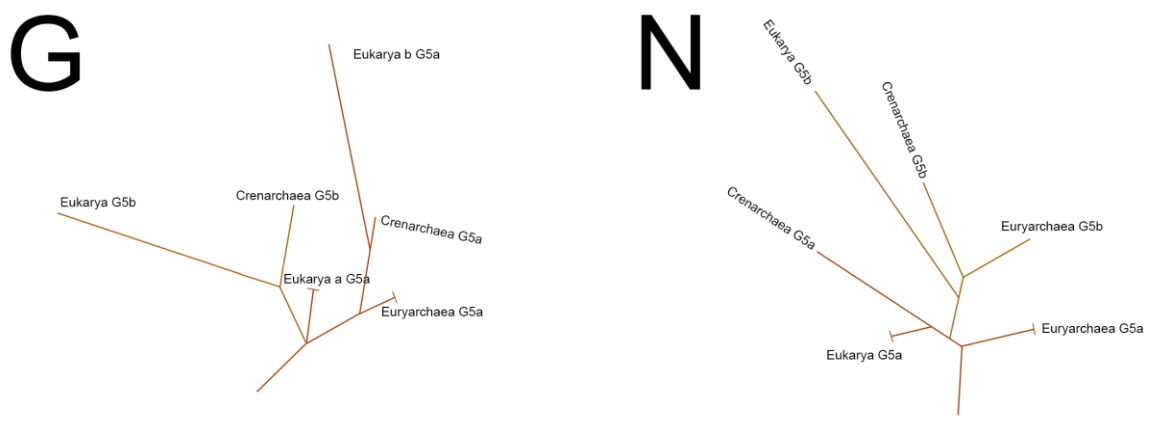


Figure 2: The individual gene trees that make up the watershed of life. Panels A through G correspond to the 3-Domain watershed. Panels H through N correspond to the 4-Domain watershed. Panels A and H show the underlying watersheds of life. Panels B and I show the ribosomal trees. Panels C and J show Gene 1. Panels D and K show Gene 2. Panels E and L show Gene 3. Panels F and M show Gene 4. Panels G and N show Gene 5.

HGT is pervasive in the watersheds shown in Panels A and B of Figure 1, represented by the dashed lines crossing grey regions. This results in multiple copies with different histories being found in living species, such as G1a version a and b found in B3, G2 versions a and b and version c in Eukarya in Figure 2 Panel D, G2 version a and b in Euryarchaea in Figure 2 Panel K, G3 versions a and b found in B1 in Figure 2 Panels E and L, and G4a versions a and b found in B2 and B5 and G4b and G4a version b found in B4 in Figure 2 Panels F and M.

Occasionally genes can be horizontally transferred across the dark grey region of the watershed, i.e. G4a being found in the Ibisial branch in Figure 2 Panels F and M. In addition, highways of HGT can form bridges between distantly related kingdoms, allowing for easy flow

of genes. This can be seen with G3 version b and c in Eukarya in Figure 2 Panel E, with G4a in Eukarya in Figure 2 Panel F, and with G4a version b in Eukarya in Figure 2 Panel M. These genes link Eukarya with Bacteria and may have come in with the mitochondria or have been picked up during phagocytosis (Doolittle, 1998).

HGT alone leads to difficulty in interpreting phylogenetic trees, but it is even worse when transfer is followed by loss of the original copy, as shown in Figure 2 Panel C with G1a versions a and b in B5, in Panels E and F with G3 versions a and b found in B2 and B3 and versions a, b, and c in Eukarya, and in Panels F and G with versions a and b of G4a found in B4. Even more difficulty is caused when these transferred genes are first transferred to the Mobilome or now extinct lineages, diverging from the standard vertical copy, and then transferred back to surviving cellular lineages, shown by dashed lines that drift out into the grey regions of the web of gene sharing, before drifting back. An example where a gene from the Mobilome or extinct lineage is transferred to only one surviving cellular lineage is G1b version c and d found in Eukarya in Figure 2 Panel J. Examples where two surviving cellular lineage both pick up the same gene from the Mobilome or from extinct lineages are G1b versions c and d found in Eukarya and G1a version b in Eukarya and G1a version b in Crenarchaea of Panel C of Figure 2, version c in Eukarya and version b in Crenarchaea of G5 in Panel G of Figure 2, and version b in Eukarya, Crenarchaea, and Euryarchaea in G5 in Panel N of Figure 2.

Examination of the gene trees shown in Figure 2 Panels C through G and J through N reveals that there are differences between vertically and horizontally acquired genes in terms of the appearance of their respective phylogenies. HGT leads to patchy, inconsistent data, where one gene unites a clade with one specific lineage, while another gene unites the clade with a different, although equally specific, lineage (Lane, 2011). Then when a supermatrix of all of

these genes is constructed, strong support is found for either the average of these signals or the tree with the least conflict with both topologies, because this is a feature of the matrix concatenation approach (Lapierre et al., 2014). On the other hand, genes mostly undergoing vertical descent show broad presence in the clade that is the proposed sister group, especially when the gene is postulated to have been present in the ancestor. These genes then consistently show the same picture, grouping in the same position, with the same sister taxa, and similar divergence times (although not necessarily identical).

5.3. Applying the Watershed of Life

Given that the true evolutionary history must either be that shown in figure 1 Panel A, or that shown in Panel B, we can examine existing evidence in a new light. The one place where the sort of patchy gene distribution characteristic of the HGT network can readily be found is in the work of the proponents of the Eocyte hypothesis (Cox et al., 2008; Ettema et al., 2011; Makarova et al., 2010; Nunoura et al., 2011; Spang et al., 2015; Williams et al., 2013; Wolf et al., 2012; Yutin et al., 2009; Yutin and Koonin, 2012). Oddly enough these researchers have compiled rather strong evidence in favor of the 3-Domain tree and a monophyletic Archaea, by showing that shared Crenarchaea-Eukarya Signature Proteins (ESPs sensu (Hartman and Fedorov, 2002)) all show the pattern associated with HGT. The evidence collected by the 3-Domain supporters is consistent with the vertical signal expected with the 3-domain tree. (Baldauf et al., 1996a; Brown et al., 1997; Brown and Doolittle, 1995; Gribaldo and Cammarano, 1998). No evidence has been found that is consistent with either the vertical signal expected with the Eocyte tree nor the underlying web of HGT expected with the Eocyte tree. Although researchers have searched for

the vertical signal, there has been no concerted effort to look for the horizontal signal. Thus, the question of how many patchy genes result from frequent HGT between Euryarchaea and Crenarchaea, has been left largely unanswered.

Owing to the dearth of research on HGT between the Euryarchaea and the Crenarchaea, this quantity cannot be compared to the quantity of HGT between Crenarchaea and Eukarya. Only when this comparison can be made, can the question of which scenario has led to the unique relationship between Eukarya and Archaea be answered. Support for the Eocyte tree should be looked for, by examining Crenarchaea and Euryarchaea for shared, but patchily represented genes. Although this endeavor is likely obscured by the fact that HGT networks are not stationary, nor are they set in stone for all time, there should still be evidence in the form of patchily distributed genes that do not closely recapitulate the known phylogeny (i.e. the type the Eocyte group has already gathered supporting the 3-domain tree). Likewise, if the Eocyte tree correctly represents the true linear history of cellular life, then there should be some genes left in the genome to show this. Such genes would be broadly represented in all Eukarya and Crenarchaea, having come from their hypothetical common ancestor. These genes would produce phylogenies that support known clades. No known genes fit these criteria.

It has been suggested that although Eukarya have exchanged HGT for Meiosis as their primary method for gene sharing, they went through a phase in their early evolution when they readily acquired horizontally transferred genes. The fact that there are more bacterial-related genes in Eukarya (Rochette et al., 2014) implies that Eukarya at one point had a very high rate of HGT. Phagocytosis of a wide range of prey may have provided the mechanism by which this happened (Doolittle, 1998). These genes were then incorporated into the eukaryal cell and put to

new, innovative usage. This might explain why Eukarya have acquired so many crenarchaeal genes from various sources, even though they no longer appear to be continuing acquisition of horizontally transferred genes on the same scale.

There are either synapomorphies or horizontally acquired genes common to the Archaea that give this group its key signature way of life. These features were acquired after the split with Eukarya, or shared via HGT. These traits include the unique ether linked lipid membrane, derived tRNA and rRNAs, the ability to colonize various extreme habitats, (Woese et al., 1978), and a slow evolutionary rate.

The reason why the branch separating Eukarya from Archaea is so short, could be that Archaea evolved a slower rate of evolution immediately after branching from Eukarya. Tighter control of mutation rate might have been important in adapting to extreme environments or simply the result of a more accurate polymerase. Or a slower rate of evolution could have evolved in the stem Ibisii and then changed back in Eukarya, speeding evolution back up. Once evolution was stuck at a slower rate, Archaea were less able to compete with the more quickly evolving Bacteria. HGT from Bacteria and Eukarya would then have become an important work-around to acquire new functional genes in order to remain competitive.

Note that the direction of transfer is not known. Stem Eukarya may have picked up these shared genes from a variety of Crenarchaea, or a variety of Crenarchaea may have gained genes from stem Eukarya, or both. If the genes are crenarchaeal gains, then this could be a valuable way for slowly evolving lineages to acquire more genes. Per (Brown, 2003), “If gene occur in most species of one domain but only a few species of another, then it is probable that a species from the more populous group was the donor.” If this is the case, genes shared between Eukarya and Archaea most likely originated in Eukarya.

On the other hand, if patchy genes are eukaryal gains, then they could be the result of phagocytotic life-style (Doolittle, 1998). In the latter case, this is also a valid explanation for how and why the genes of non-mitochondrial bacterial origin entered the eukaryal line (assuming they are not in fact vertically inherited genes present in LUCA, but lost in Archaea). The eukaryal stem would thus have had a tendency to pick up and retain genes. Such tendency may be a pre-requisite to the type of morphological complexity exhibited by various eukaryal cells (Koonin, 2010).

5.4. Implications for Symbiosis in the Origin of Eukarya

Symbiotic hypotheses for the origin of Eukarya in which the Eukarya arose from within the Archaea (Martin and Müller, 1998; Moreira and López-García, 1998) are generally thought to be more parsimonious, because they do not require the existence of a proto-eukaryal lineage, but they are not simpler. The latest symbiotic fusion papers claim that the Last Archaea Eukarya Common Ancestor (LAECA) was the proto-eukaryal lineage, and as such was a fully-fledged Archaeon that was more complex than any modern Archaeon (Wolf et al., 2012; Yutin et al., 2009). An important point is that the Last Eukarya Common Ancestor (LECA) was a complex modern Eukaryeon (Archibald et al., 2000; Collins and Penny, 2005; DeGrasse et al., 2009; Field and Dacks, 2009; Koumandou et al., 2013; Neumann et al., 2010), so the list of eukaryal features that need to either be present in LAECA or acquired shortly after is close to the list of eukaryal signature proteins *sensu* (Hartman and Fedorov, 2002). If all the genes found in desperate archaeal species were present in the LAECA, then gene loss was rampant. This is in fact a very convoluted scenario, because the proto-eukaryal lineage is still proposed to have existed. It has only been moved further back in time and given less time to evolve, in addition to requiring a

large number of independent gene losses and convergent genome simplification events to explain the genomes of the present-day Archaea. The most parsimonious choice is that Archaea are monophyletic and shared genes are the product of HGT. But, evolution is not necessarily parsimonious and the Eukarya could very well have evolved from within the Archaea. However, Archaea/Bacteria fusion scenarios for the origin of Eukarya require various implausible intermediates with unlikely membrane loses, that are not supported on a cell-biology level (Cavalier-Smith, 2014) (Jekely, 2006; Poole and Penny, 2006). Fusion hypotheses, relying on extensive HGT from various branches of a tree to create a complex, undocumented hypothetical host ancestor that was then involved in a symbiosis with implausible intermediates, should not be considered a simpler explanation to autogenesis of eukaryal traits (Cavalier-Smith, 2014; Jékely, 2007; Keeling, 2014; Poole and Penny, 2007; Rochette et al., 2014) with a highway of gene sharing. And, there is significant scientific evidence supporting a link between Eukarya and Euryarchaea, the archaeal Kingdom proposed to be most distantly related to Eukarya per the currently favored fusion hypotheses, including Supertree whole genome analyses (Pisani et al., 2007). If the Eukarya are more closely related to the Crenarchaea, then Supertrees, which have the advantage of not suffering from the inherent concatenation artifact of combing the signal of HGT with that of vertical descent to produce a phylogeny that represents neither, should not group Eukarya with Euryarchaea to the exclusion of Crenarchaea. The fact that Eukarya go with Euryarchaea in some analyses and Crenarchaea in others mildly supports the hypothesis that

Eukarya go with both, equally, but the branch length separating them is too short to produce a reliable signal.

It has been said that the Eukarya did not exist before the formation of the federation with the Alphaproteobacteria that became the mitochondria (Williams et al., 2013). In this way of thinking, the Eukarya are nothing more than a fusion between an Archaeon and a Bacterium, but this is not the case. Even if the Eukarya evolved within the Archaea and not as sister to them, they still evolved from a very derived ancestor that possessed many or possibly all the eukaryal innovations. These innovations are not found in Bacteria and did not come with the endosymbiont; complex molecular machineries and membrane bound compartments do not simply spring up when an endosymbiont is gained. Even if these traits were gained after the mitochondria was acquired, even with simpler homologs performing similar, but fundamentally simpler functions were acquired and built upon, the full interwoven cellular and molecular complexity of it all was still invented by cells that gave rise to LECA. Eukaryal features are fundamentally eukaryal innovations and the eukaryal clade is much more than just the fusion of the parts of the federation. Even if the Eukarya evolved from within the Archaea, they are not Archaea. They lack the level of cellular simplicity and specific features shared by the Archaea. They are the only known living members of a more cellularly complex grade, called the Eukaryotes.

5.5. Conclusion:

There is hope that one day a definitive answer can be obtained with regards to the topology of the split between Eukarya and Archaea. But the type of data that is missing is not the type of data that researchers are currently looking for. An effort needs to be made to examine and

quantify the network of Horizontal Gene Transfer that joins Crenarchaea and Euryarchaea. When this last piece of the puzzle is obtained, then a monophyletic archaeal scenario can be teased apart from the paraphyletic archaeal scenarios.

The present data, with an ever-increasing number of Eukaryal Signature Proteins found patchily distributed in distantly related Archaea, supports only the 3-Domain topology, because patchy genes are a sign of Horizontal Gene Transfer. Furthermore, symbiotic fusion scenarios for the origin of the Eukarya are less parsimonious than autogenesis scenarios for the origin of Eukarya, explaining less while requiring unnecessary and convoluted steps. Unless new evidence is found that supports symbiosis or archaeal paraphyly, then the 3 Domain tree and autogenesis as the mechanism for the origin of Eukarya must remain the null hypotheses.

Chapter 6: The Root of the Eukarya

6.1. Introduction

Currently no consensus exists as to where the root within the Eukarya lies (Cavalier-Smith and Chao, 2010; Derelle and Lang, 2012; He et al., 2014; Pace et al., 1986; Rogozin et al., 2009; Stechmann and Cavalier-Smith, 2002; Stiller et al., 1998; Wideman et al., 2013; Woese et al., 1990). The fossil record is no help in this regard, because the first definitive eukaryal fossil belongs to the red algae (Butterfield et al., 1985), which belong to a larger group, the Archaeaplastida, and are not even the deepest branch within this group (Huang and Gogarten, 2007), making them an unlikely candidate for the location of the root (unless one believes figure 7). With no fossil evidence to inform on the problem, one is left only with molecular sequence information and phylogenetic reconstruction.

The main problem with rooting the Eukarya using phylogenetic reconstruction has to do with the long eukaryal stem branch discussed in Chapter 4. Because of this long branch, the closest outgroup, the Archaea, are incredibly far away. Using Archaea to root the Eukarya is analogous to throwing a dart across a football field and trying to hit a target. But some phylogenetic reconstruction programs, such as phyml (Guindon et al., 2010), must return a bifurcating tree, even if there is no evidence in the dataset to support any bifurcation pattern. So the result of such an exercise is that the long branch to the archaeal outgroup attracts to some random branch of the eukaryal tree, creating a Long Branch Attraction artifact.

In order to root the Eukarya, a closer outgroup is needed. It is impossible to find a lineage that is a closer outgroup than the Archaea, because all such lineages went extinct. But although whole cell lineages did not speciate during this time period or have not yet been sequenced, many individual genes did proliferate through the processes of gene duplication. Paralogs that

diverge from each other on the Eukaryal stem branch are closer outgroups to the Last Eukaryal Common Ancestor (LECA) than Archaea and should therefore be more likely to accurately locate the root within Eukarya.

An additional problem with rooting the Eukarya is that individual protein trees lack resolution when it comes to branching order. A multi-protein concatenation may multiply phylogenetic signal and allow for deeper resolution, so a concatenation of paralogs could provide a closer root and increase the phylogenetic signal. On the other hand, concatenation will also multiply the signal from artifacts and could produce a consensus tree that does not correspond to the actual evolutionary history of any of the component proteins (Bapteste et al., 2005). In this chapter, I attempt to counter these problems by producing protein concatenations that are limited to actual evolutionary units, such as the histone complex, the tubulins, and the proteasome.

The four histone subunits come together to form the histone complex and have likely evolved together throughout the history of the Eukarya, as have the three tubulins, which interact to produce the microtubule system, and the 14 subunits of the proteasome, which form a complex macromolecule responsible for protein degradation. Using histone, tubulin, and proteasome protein concatenations will increase resolution and allow for independent determinations of branching order at the deepest level in the eukaryotic tree.

There are 4 histone paralogs, H2A, H2B, H3, H4, common to all Eukarya (Bell and White, 2010; Thatcher and Gorovsky, 1994). Alpha, Beta, and Gamma Tubulin are present in most Eukarya and believed to have been present in LECA (Fournier et al., 2011). It was unknown whether the fourteen subunits of the proteasome were all present in LECA, or if gene duplication continued after LECA, but my preliminary analysis indicated that all fourteen

subunits each are monophyletic. If gene duplication continued after LECA, one or more subunits would group inside another, creating a paraphyletic subunit. Thus, because all fourteen subunits are monophyletic, all fourteen must have been present in LECA and are suitable for placing the root within Eukarya.

6.2. Methods

6.2.1. Histones

Although the histone data had previously been used for analyzing the eukaryal stem branch length in chapter 4, the data were reassembled from scratch in 2013 for this project. Histone amino acid sequences were downloaded from NCBI's list of fully sequenced genomes. If more than one genome of a closely related group was available, one was chosen at random. Histone sequences were first identified from the list of annotated proteins and then verified using reciprocal best BLAST hits. An in house Perl script was used to modify taxa names. Clustal X 2.0 was used to create initial alignments within a subunit and also profile alignments of the subunits to the structural alignment (Larkin et al., 2007). Deepview was used to create a structural alignment of the four histone subunits (Kaplan and Littlejohn, 2001). Jalview was used to manually adjust the histone alignment (Clamp et al., 2004). Clustal X 2.0 was used to convert alignments to phylip format (Larkin et al., 2007). FigTree1.2.2 was used to visualize trees (Rambaut and Drummond, 2008). An in house python script was used to concatenate the subunits. Permutations of concatenation order were used to root the concatenations, as depicted in figure 1 (modified from (Hilario and Gogarten, 1993)). With 4 histone subunits, only 4 permutations are needed so that each individual protein in the concatenation is rooted with all of its paralogs, to break up the long branch to the outgroup as much as possible.

PhyML 3.0 was used to produce Maximum Likelihood trees (Guindon et al., 2005), using the WAG substitution model, 100 bootstrap replicates, estimated proportion of invariable sites, 4 categories of substitution rates, and an estimated gamma distribution parameter.

Extreme divergence and short sequences make histone alignments difficult. To judge the quality of the alignment, Maximum Likelihood trees were generated and examined for eukaryal monophyly. Since no one believes that the Eukarya evolved independently more than once or that a Eukaryeon has reverted to a prokaryote, it is universally agreed upon that the Eukarya, and therefore LECA, are monophyletic. LECA is present four times on the histone tree, one for each of the four subunits, and only when all four instances are resolved as monophyletic with high bootstrap support value, will the alignment be accepted.

6.2.2. Tubulins

The tubulin dataset was curated in the same manner as that of the histone dataset. The data were reassembled from scratch in 2013. Tubulin amino acid sequences were downloaded from the same genomes used in the histone analysis. Tubulin sequences were first identified from the list of annotated proteins and then verified using reciprocal best BLAST hits. An in house Perl script was was used to modify taxa names.

When taxon sampling of NCBI proved insufficient to root the Eukarya, transcriptome datasets and whole genome sequences freely available were searched for additional sequences using TBLASTN (Gertz et al., 2006). Because the initial sequence used as query affects the results of TBLASTN searches, multiple searches were performed on the virdiplantae using *Angomonas deanei* sequences (an euglenozoa-type excavate) as seed and *Arabidopsis thaliana* sequences as seed. *A. deanei* sequences were used under the hypothesis that the excavates are the

root of the Eukarya and therefore would be an outgroup to all the newly sequenced genomes under consideration, making them an equally good seed for all. *A. thaliana* sequences were used under the hypothesis that as a far closer relative to the virdiplantae sequences, being a member of the virdiplantae themselves, they would make for a better seed and be more likely to return the full sequence of the desired ortholog. Upon comparison of the two sets of TBLASTN searches, the excavate seed performed equally well as the viridiplantae seed, when analyzing multiple virdiplantae genomes. The excavate, *A. deanei* was used as the only seed for all further TBLASTN searches. Alpha and Beta subunits were identified, but gamma TBLASTN searches returned only partial matches. Preliminary trees showed all partial gamma sequences grouping together, to the exclusion of the full gamma sequences, even though both sets were interspersed throughout the eukaryal tree. This was ruled to be an artifact caused by the truncated sequences and all partial gamma sequences removed from further analysis.

An in house PERL script was written to modify taxa names. Protein alignments were performed using Muscle (Edgar, 2004). Clustal X 2.0 was used to create profile alignments (Larkin et al., 2007). Clustal X 2.0 was used to convert alignments to phylip format. FigTree1.2.2 was used to visualize trees (Rambaut and Drummond, 2008). An in house python script was used to concatenate the subunits. Permutations of concatenation order were used to root the concatenations, as depicted in figure 2. With 3 tubulin subunits, only 3 permutations are needed so that each individual protein in the concatenation is rooted with all of its paralogs, to break up the long branch to the outgroup as much as possible.

Protest 2.4 was used to determine that the LG + Gamma +I model maximized the likelihood of the tree (Abascal et al., 2005). PhyML was used to produce Maximum Likelihood trees (Guindon et al., 2005), using the LG substitution model, 100 bootstrap replicates, estimated

proportion of invariable sites, 4 categories of substitution rates, and an estimated gamma distribution parameter. Although aligning tubulins was much simpler than histone and proteasome alignment, Maximum Likelihood trees were still generated and examined to judge alignment quality; mostly sequences causing Long Branch Attraction artifacts were removed.

6.2.3. Proteasome subunits

Proteasome amino acid sequences were downloaded from NCBI, as described for the histone and tubulin data sets. When taxon sampling of NCBI proved insufficient to root the Eukarya, transcriptome datasets and whole genome sequences freely available were searched for additional sequences using TBLASTN (Gertz et al., 2006). As described for the tubulin data set, *Angomonas deanei* sequences (an euglenozoa-type excavate) and *Arabidopsis thaliana* sequences were both used as seed for multiple virdiplantae genomes. When they performed equally well, *A.* deanei was chosen for all further TBLASTN searches. An in house perl script was used to modify taxa names. Each subunit was first aligned individually using Muscle (Edgar, 2004). Then Promals3D was used to create a structural alignment for the 14 subunits (Pei et al., 2008). The number of subunits stymied the profile alignment capabilities of popular alignment programs, so Seaview was used to manually preform a profile alignment.

Clustal X 2.0 was used to convert alignments to phylip format (Larkin et al., 2007). FigTree1.2.2 was used to visualize trees (Rambaut and Drummond, 2008). An in house python script was used to concatenate the subunits. Permutations of concatenation order were used to root the concatenations. With 14 proteasome subunits, 14 permutations would be needed so that each individual protein in the concatenation is rooted with all of its paralogs. But because it is unknown whether the alpha subunits diverged from the beta subunits on the eukaryal stem or

before the Last Archaeal/Eukaryal Common ancestor, and because the branch separating the alphas from the betas is long, the dataset was split in two: The alpha proteasomes and the beta proteasomes, so that alpha subunits are concatenated and rooted with only alpha subunits and the beta subunits are concatenated and rooted with only beta subunits.

Protest 2.4 was used to determine that the LG + Gamma +I model maximized the likelihood of the tree (Abascal et al., 2005). PhyML 3.0 was used to produce Maximum Likelihood trees (Guindon et al., 2005), using the LG substitution model, 100 bootstrap replicates, estimated proportion of invariable sites, 4 categories of substitution rates, and an estimated gamma distribution parameter. When PhyML failed to produced trees in time, FastTree (Price et al., 2010) was used with default settings (JTT, the "CAT" approximation, and "SH-like local supports") to create quick and dirty place holder trees.

Proteasome alignment was complicated by the misnaming of the subunits for which crystal structure was available. Additional subunits were found to be misannotated, with top BLAST hits also misannotated, so placement in Maximum Likelihood trees was used to confirm subunit identity. Additionally, taxa and or subunits causing long branch attraction artifacts in Maximum Likelihood trees were removed.

Figure 1: Using Histones concatenations to root LECA. Each of the 4 histones subunits is rooted using concatenations composed of the other 3 subunits in every other possible order, as well as a concatenation made from repeating the archaeal histone sequence 4 times.

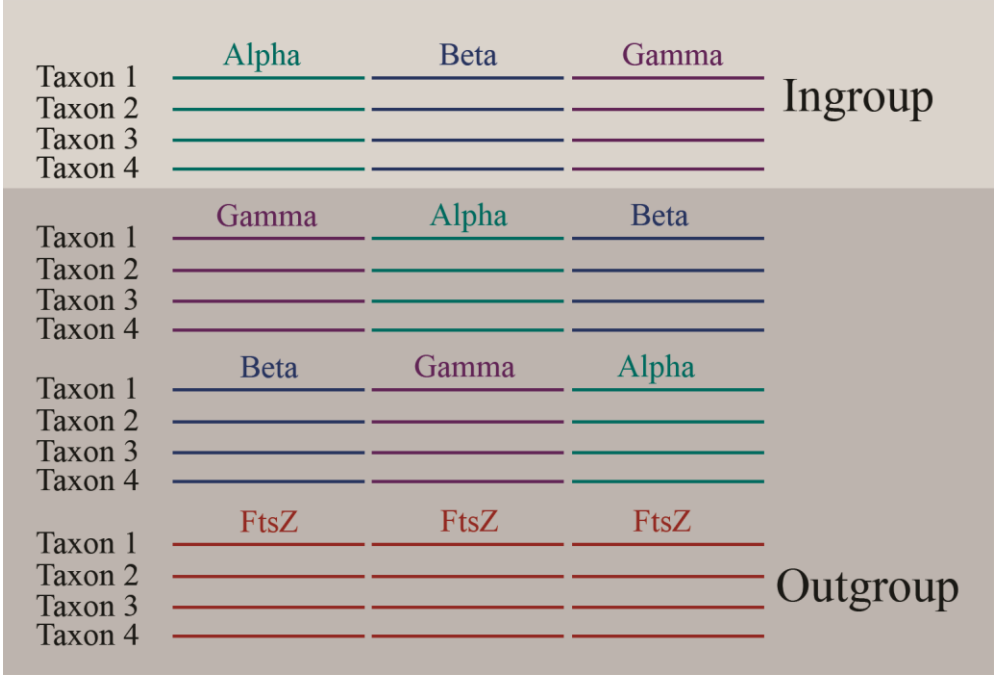


Figure 2: Using tubulins to root LECA. Each of the 3 eukaryal tubulin subunits, alpha, beta, and gamma, is rooted using concatenations composed of the 2 subunits in every other possible order, as well as a concatenation made from repeating the FtsZ sequence 3 times.

6.3. Results

6.3.1. Histones

The histones were so conserved within a subunit that they could be aligned by hand. However, they were so divergent between subunits that a structural alignment had to be performed. Unfortunately, histone sequences are extremely short ~200 bp per sequence, so not even concatenating all four together provided a reliable branching order within the Eukarya, shown in Figure 3. I therefore conclude that the histones do not contain enough phylogenetic signal at the level of LECA to root the Eukarya.

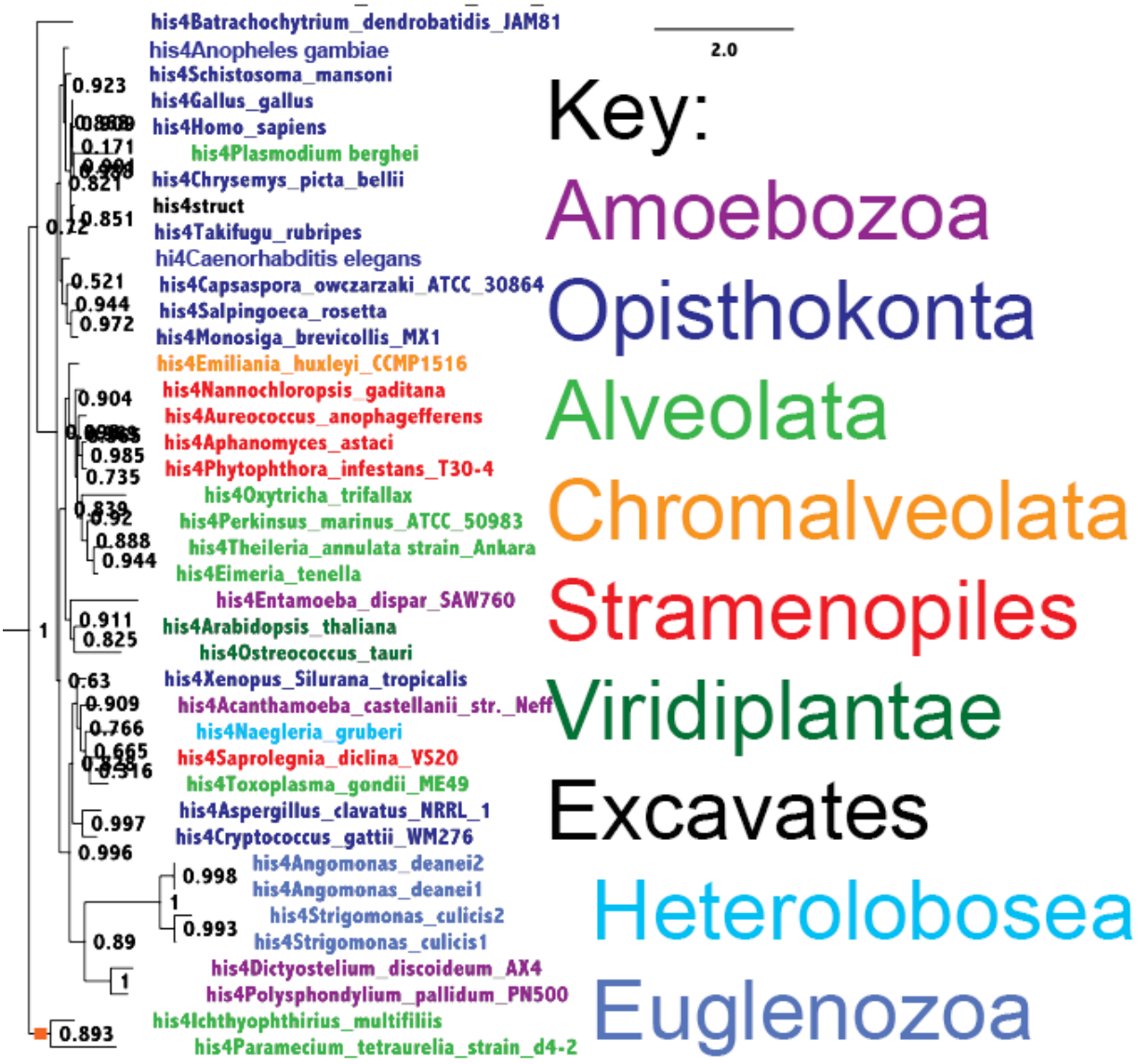


Figure 3: One of the 4 histone concatenations, rooted with the other 3 histone concatenations.

The other 3 are not shown, because they are suffering from the same problems as this one.

6.3.2. Tubulins

Obtaining alpha and beta tubulin sequences from new genomes was straightforward, owing to the high levels of sequence conservation in these subunits. The same cannot be said for the gamma subunit, whose sequences were highly divergence across the Eukarya. When TBLASTN yielded a sequence, it was almost always a partial sequence. A preliminary tree of these sequences showed the new partial sequences grouping together to the exclusion of the full gamma sequences. Therefore, the partial gamma sequences were thrown out of the dataset and only alpha and beta tubulin were used to represent the newly sequenced genomes and transcriptomes.

Figures 4, 5 and 6 show the three tubulin concatenations. They were calculated as one Maximum Likelihood tree, but broken into three parts, in order to display them. All 3 exhibit a paraphyletic excavate root, but all three show signs of artifacts. The deepest branch is always Euglena, but this looks like Long Branch Attraction. The Archeaplastida and SAR supergroups are not present as monophyletic groups. One or two taxa grouping in unexpected places could be due to errors in the phylogenetic assignment to supergroup or horizontal gene transfer, especially in taxa known for horizontal gene transfer or known to harbor eukaryal endosymbionts, such as the chromalveolates. But there are just so many taxa out of place that there is little hope for this dataset in resolving the split at the deepest levels in the eukaryal phylogeny.

To verify that this tree is suffering from some irreconcilable artifact, all of the deepest branches were removed and the tree rerun, checking to see if the interior of the tree was stable. It wasn't, as seen in Figure 7. The topology of the tree completely changed, likely because there is no support for any of the deeper nodes, and the root shifted to within the red algae. The data is

not shown, but this process of removing deeply branching taxa was continued until there were so few taxa left that it was clear that the instability of the tree was not a factor of a few long branches.

It is possible that a combination of identifying long branches to remove, adding the full sequences of the gamma subunit from newly sequences genomes, and adding data from taxa yet to be sequenced, might provide new information that would allow the tubulin family to resolve the branching order at the deepest levels within the Eukarya. But as it is now, with only the alpha and beta subunits available for half of the dataset, there just is not enough phylogenetic signal to produce a reliable branching order.

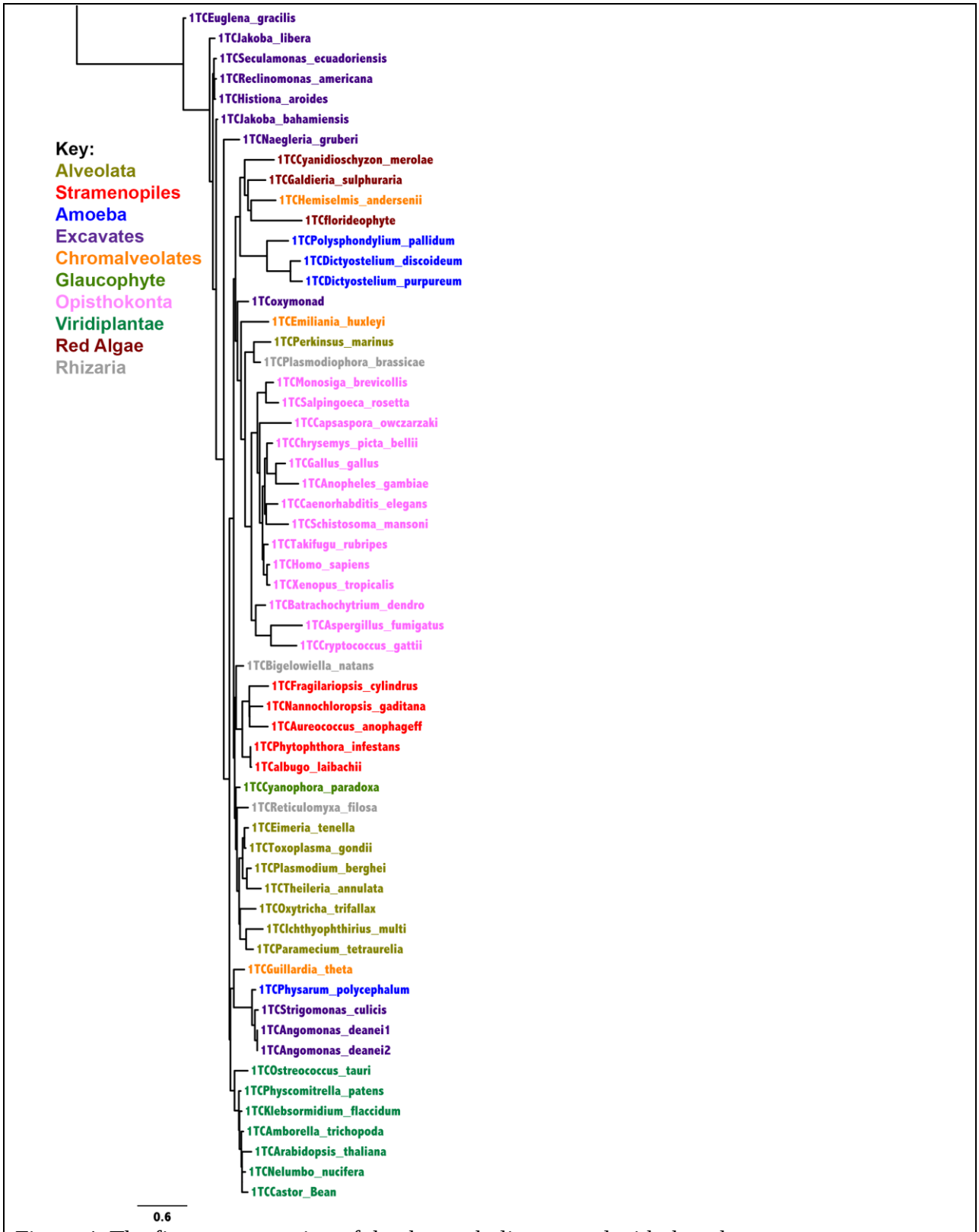


Figure 4: The first concatenation of the three tubulins, rooted with the other two.

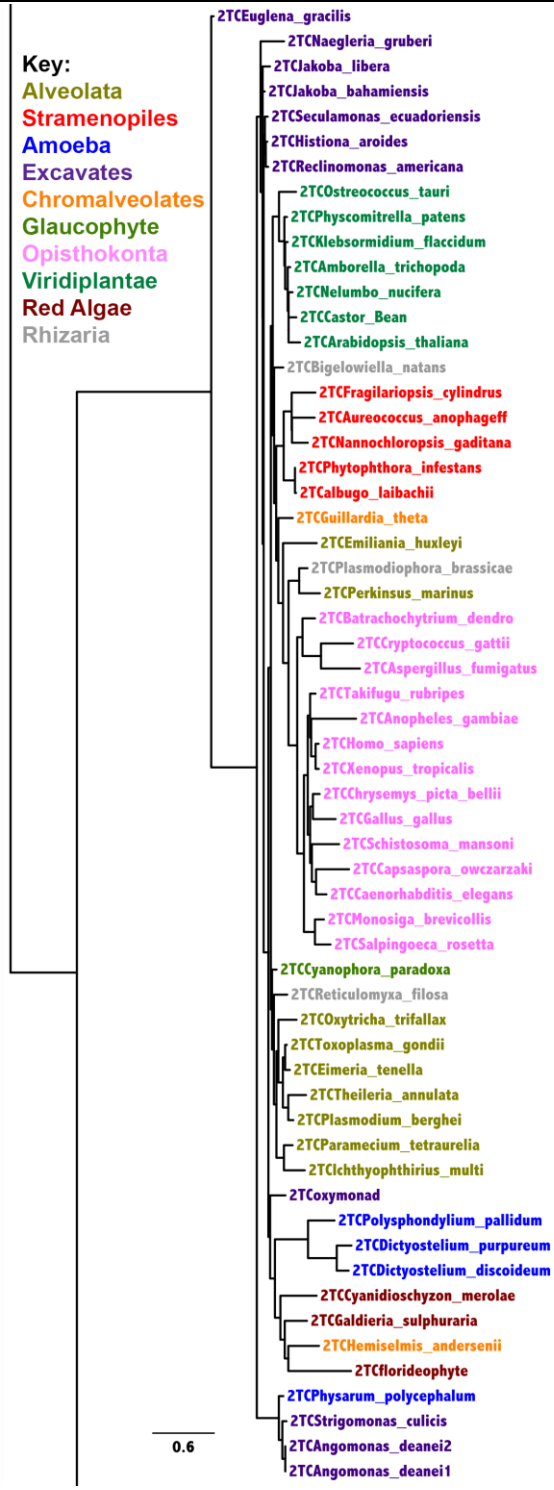


Figure 5: The second tubulin concatenation, rooted with the other 2. Colors are the same as in Figure 4.

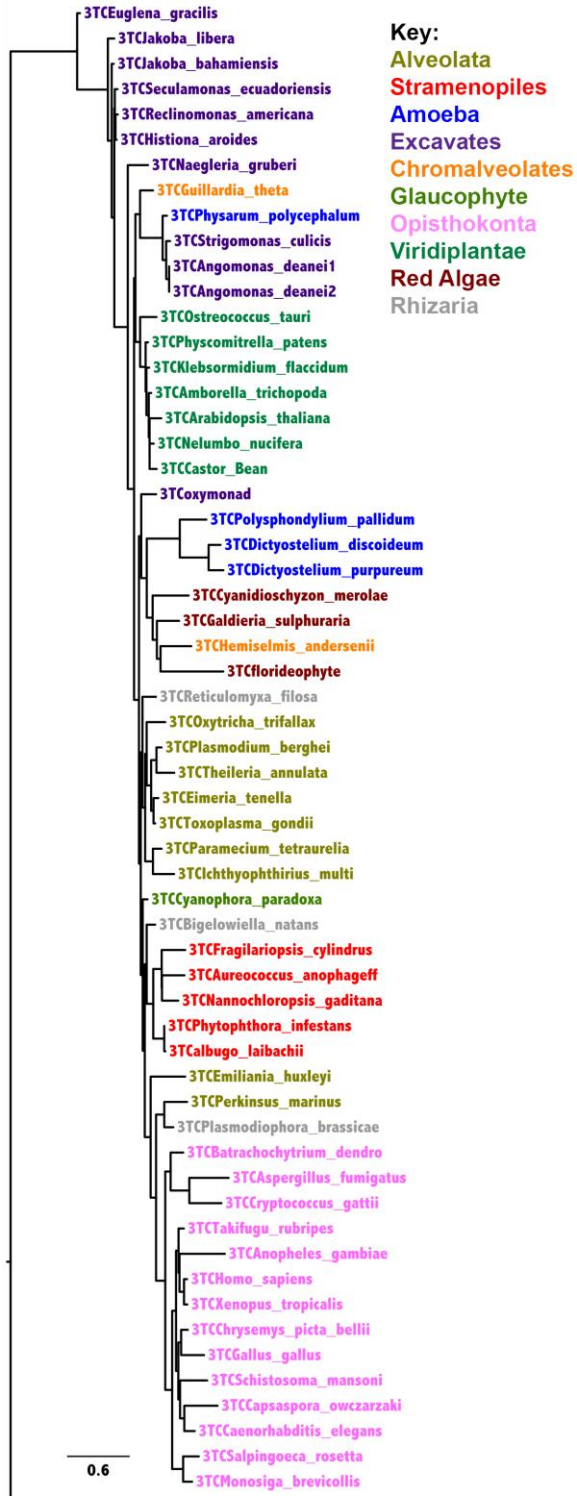


Figure 6: The third tubulin concatenation, rooted with the other 2. Colors are the same as in Figure 4.

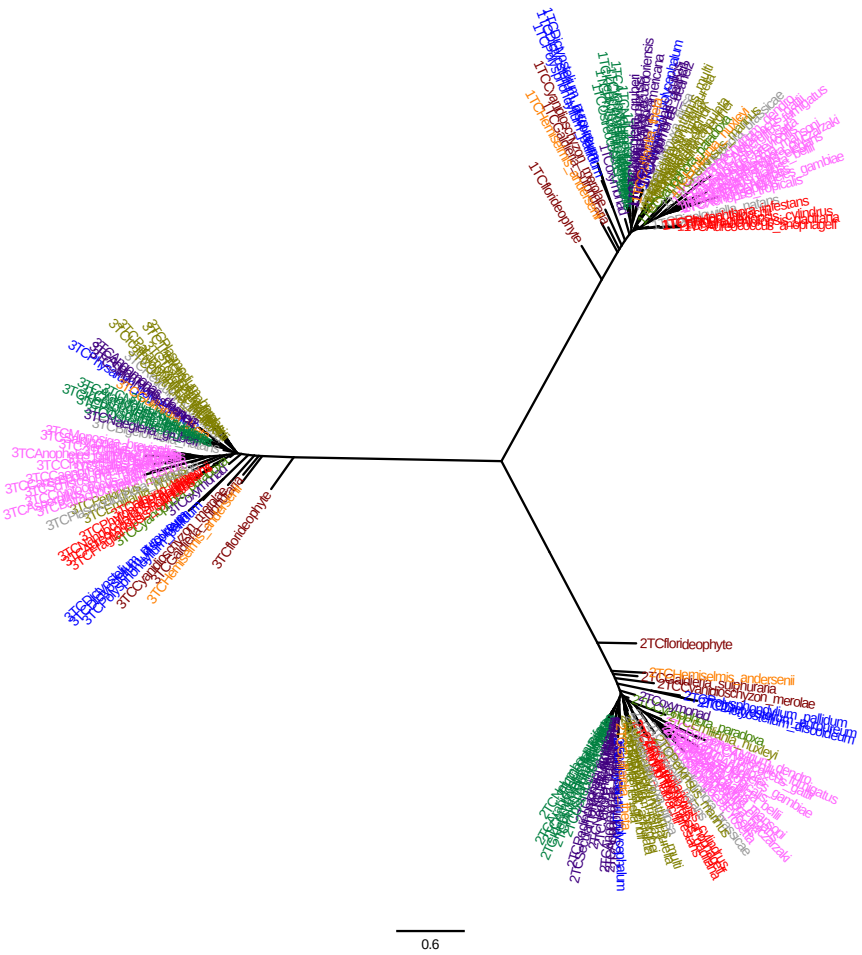


Figure 7: The deepest branches from figures 4, 5, and 6, the Euglena, were removed and the tree re-estimated. Now the root is within the red algae and known supergroups are still not recovered. Colors are the same as in Figure 4.

6.3.3. Proteasome

Figure 8 shows a phylogenetic reconstruction of all 14 of the eukaryal proteasome subunits and the 2 archaeal proteasome subunits. From this tree, it appears that either the ancestor of the Archaea and Eukarya had a 2 subunit proteasome, or, the Archaea acquired a 2 subunit proteasome from a stem group Eukaryon, prior to the divergence of the 7 alphas and the 7 betas. Although the topology within the eukaryal subunits is not resolved, all 14 subunits are clearly resolved as monophyletic and therefore had to have been present in the Last Eukaryal Common Ancestor (LECA).

Figure 9 shows only the 14 eukaryal proteasome subunits, but with the addition of a number of new sequences obtained from newly sequenced genomes and from transcriptome projects. All 14 subunits are still monophyletic. The topology within each subunit is still unresolved.

Figure 10 shows the 7 individual beta subunits that are concatenated in Figure 11. Neither shows resolution at the deepest nodes in the eukaryal tree, but occasionally known super groups are recovered. One of the subunits in Figure 11 is scaled up in size in Figure 12, to increase visibility. Support values, representing "SH-like local supports," are shown in black. All support values are so low as to be unusable, represent known clades whose topology is not in question, or represent obvious cases of Long Branch Attraction.

Figure 13 shows the 7 individual alpha subunits that are concatenated in Figure 14. One of the subunits in Figure 14 is scaled up in size in Figure 15, to increase visibility and support values, representing "SH-like local supports," are shown in black. All support values are so low as to be unusable, represent known clades whose topology is not in question, or represent obvious cases of Long Branch Attraction.

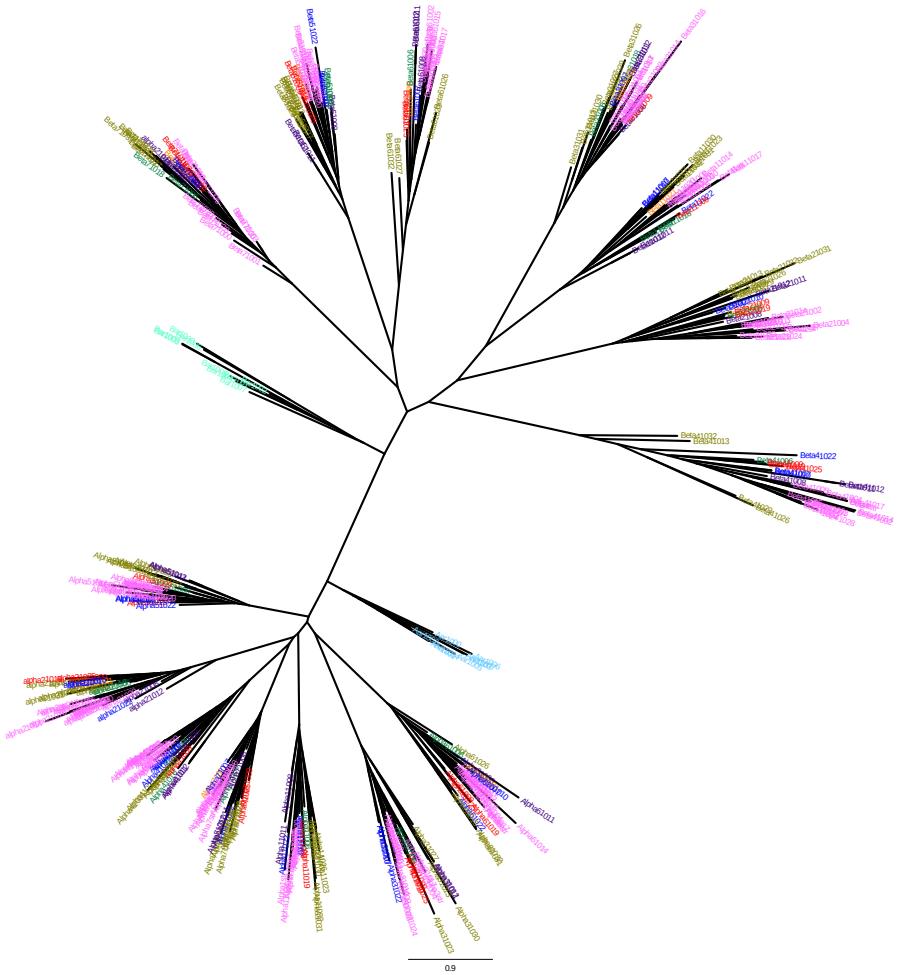


Figure 8: Phyml Maximum Likelihood tree of all 14 of the proteasome subunits and the 2 Archaeal proteasome subunits. Colors are the same as in Figure 4, with the addition of pastel blue for the archaeal alpha subunits and seafoam green for the archaeal beta subunits.

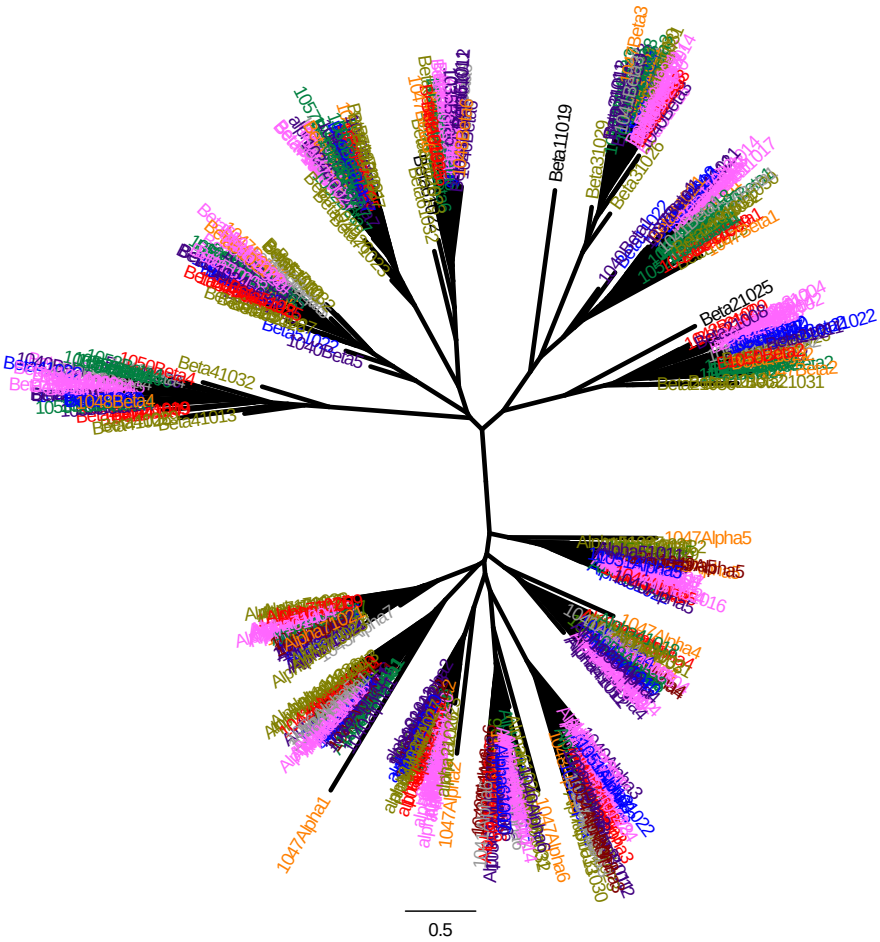


Figure 9: FastTree with default settings of the 14 eukaryal proteasome subunits with a broader sampling. Colors are the same as in Figure 4.

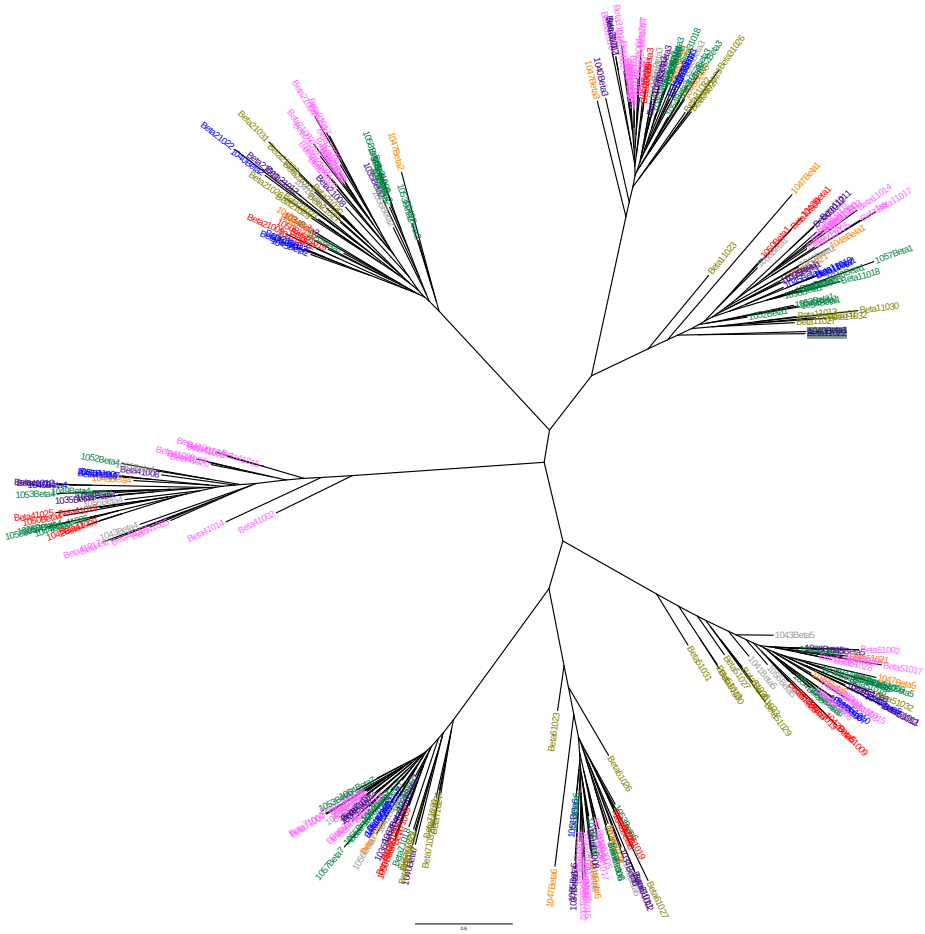


Figure 10: FastTree with default settings of only the 7 eukaryal beta proteasome subunits, used for the beta concatenations. Colors are the same as in Figure 4.

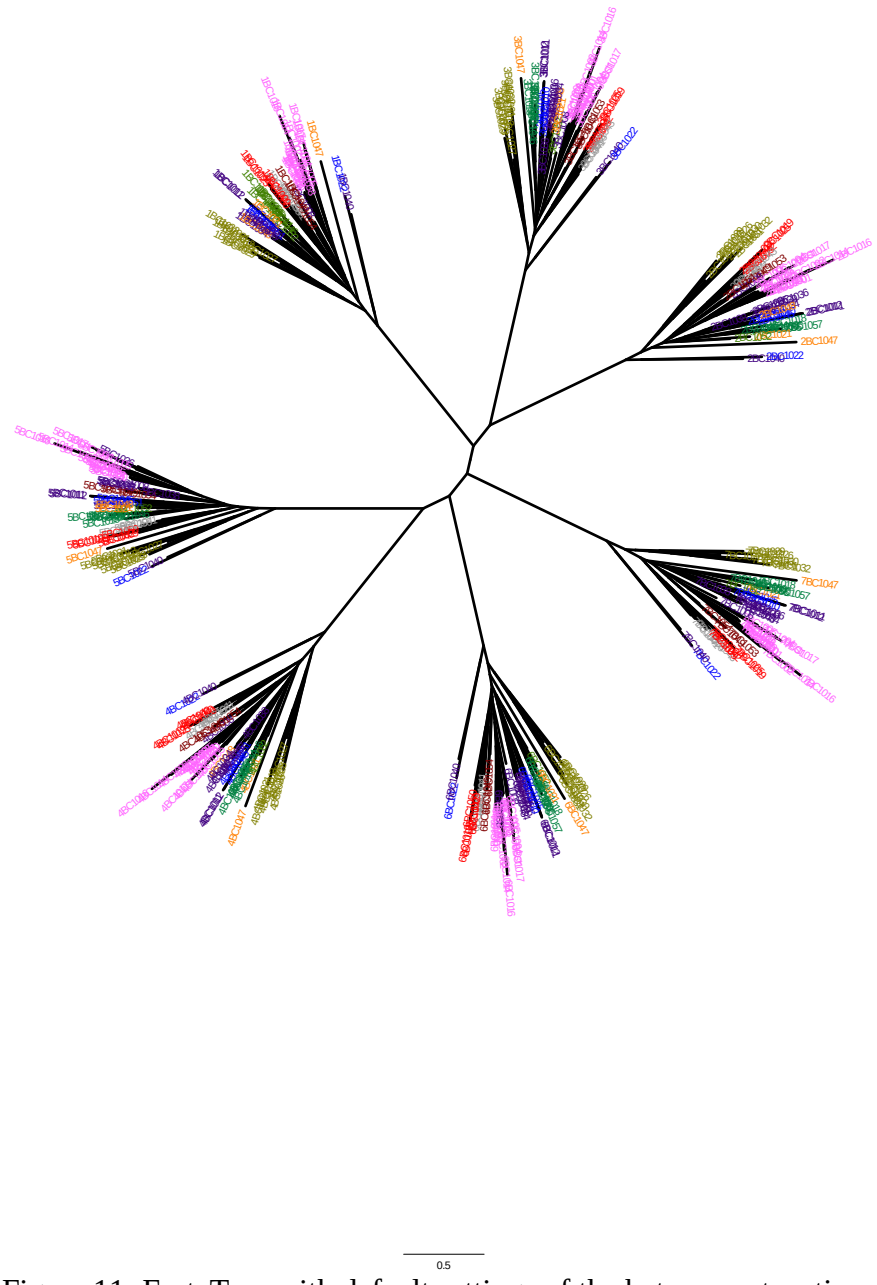


Figure 11: FastaTree with default settings of the beta concatenation. Colors are the same as in Figure 4.

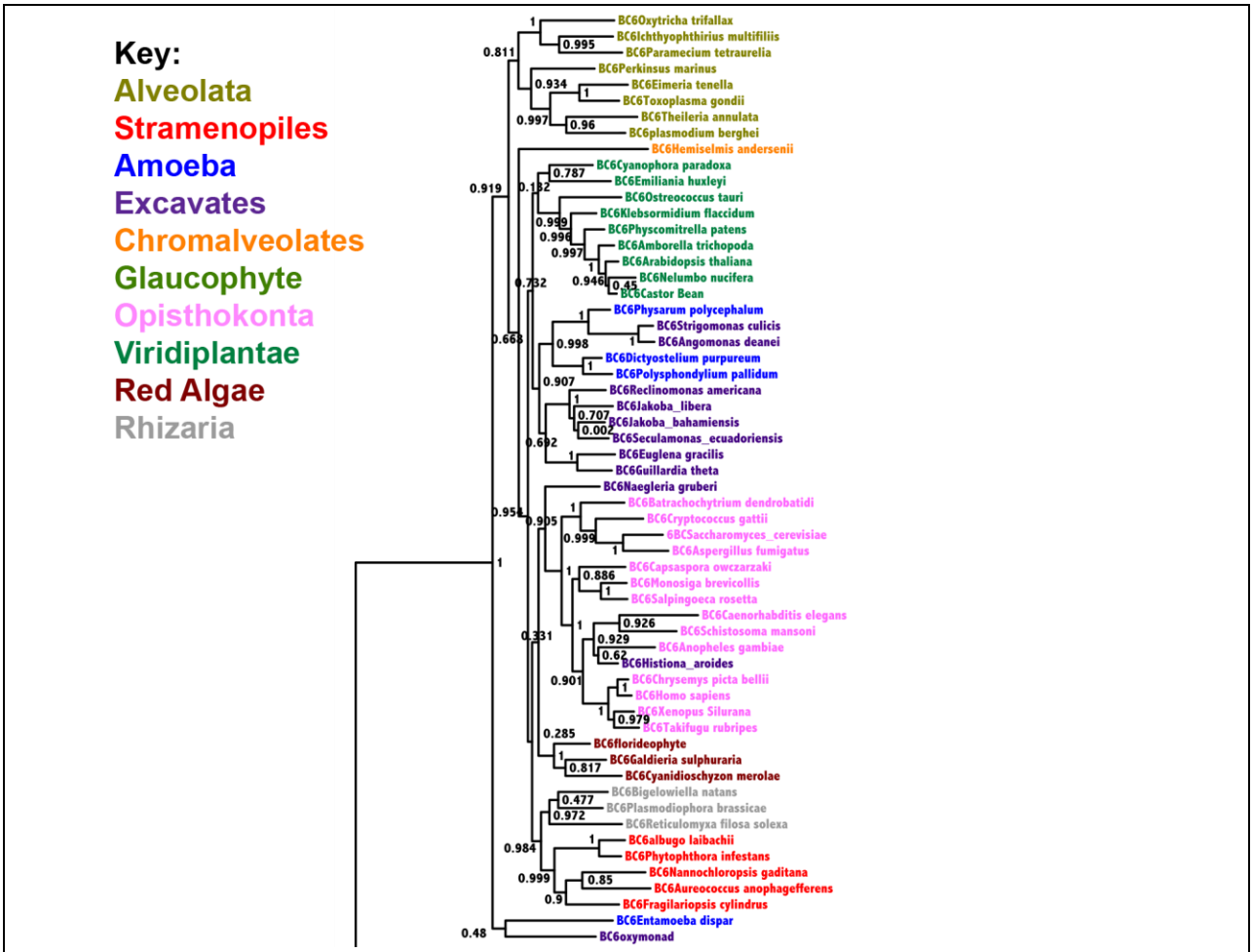


Figure 12: Same as Figure 11, but zoomed in to show only one of the 7 concatenations. Support values represent "SH-like local supports."

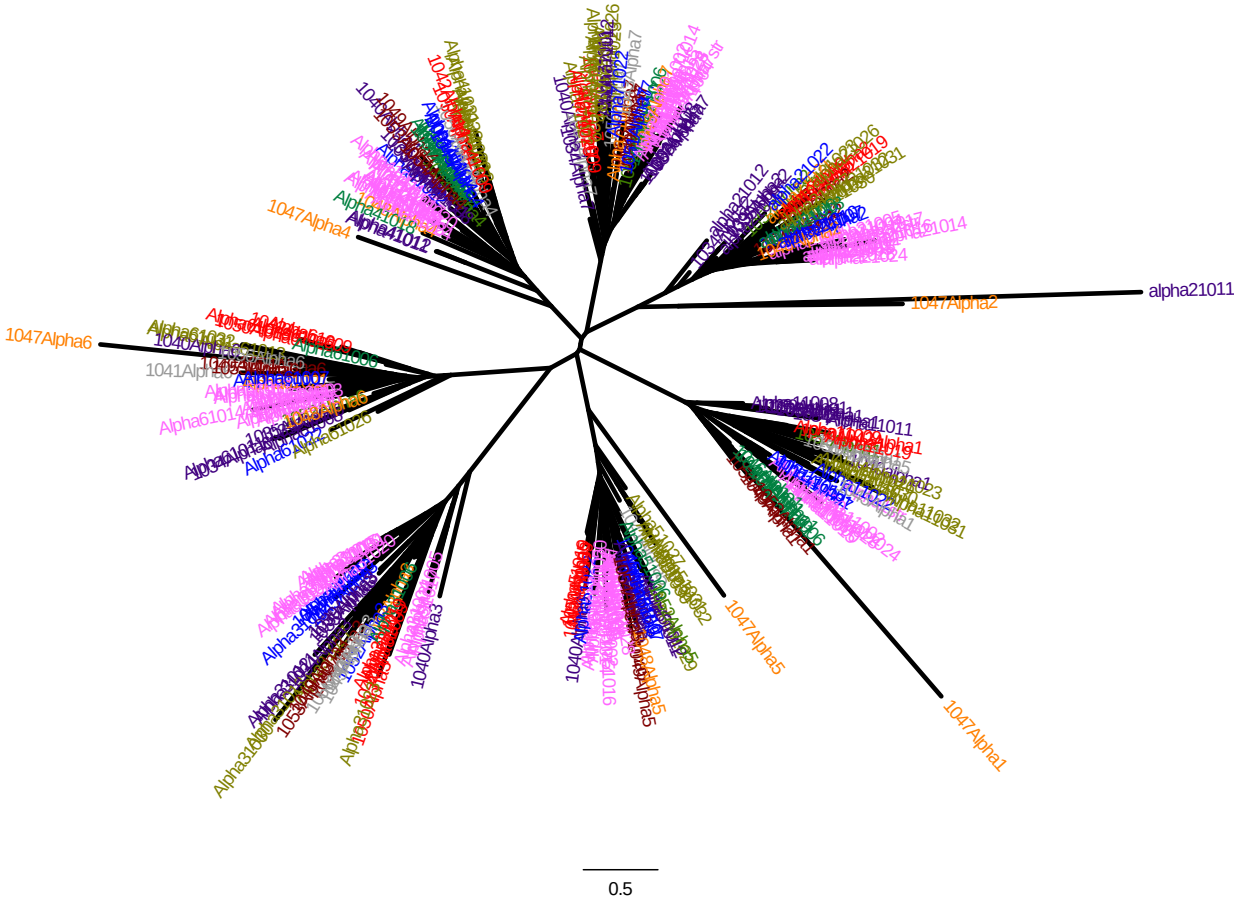


Figure 13: Phyml Maximum Likelihood tree with only the 7 eukaryal alpha proteasome subunits.

Colors are the same as in Figure 4.

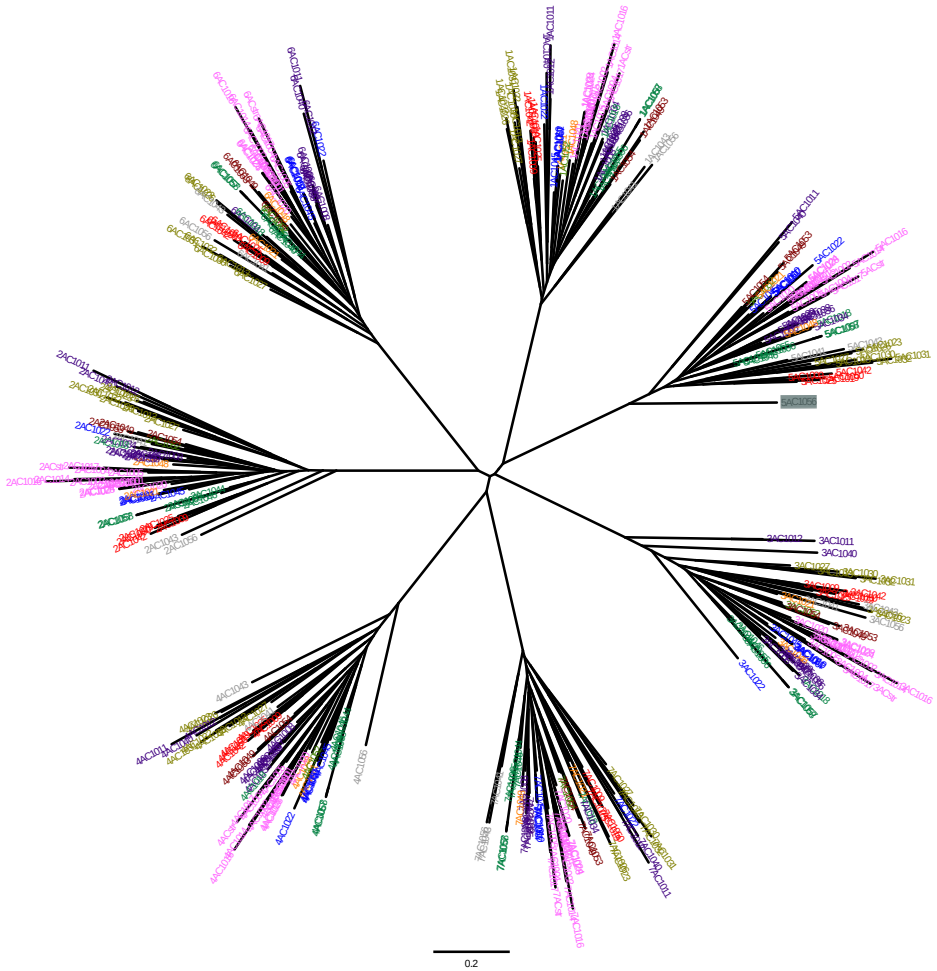


Figure 14: FastTree with default settings of the 7 alpha concatenations. Colors are the same as in Figure 4.

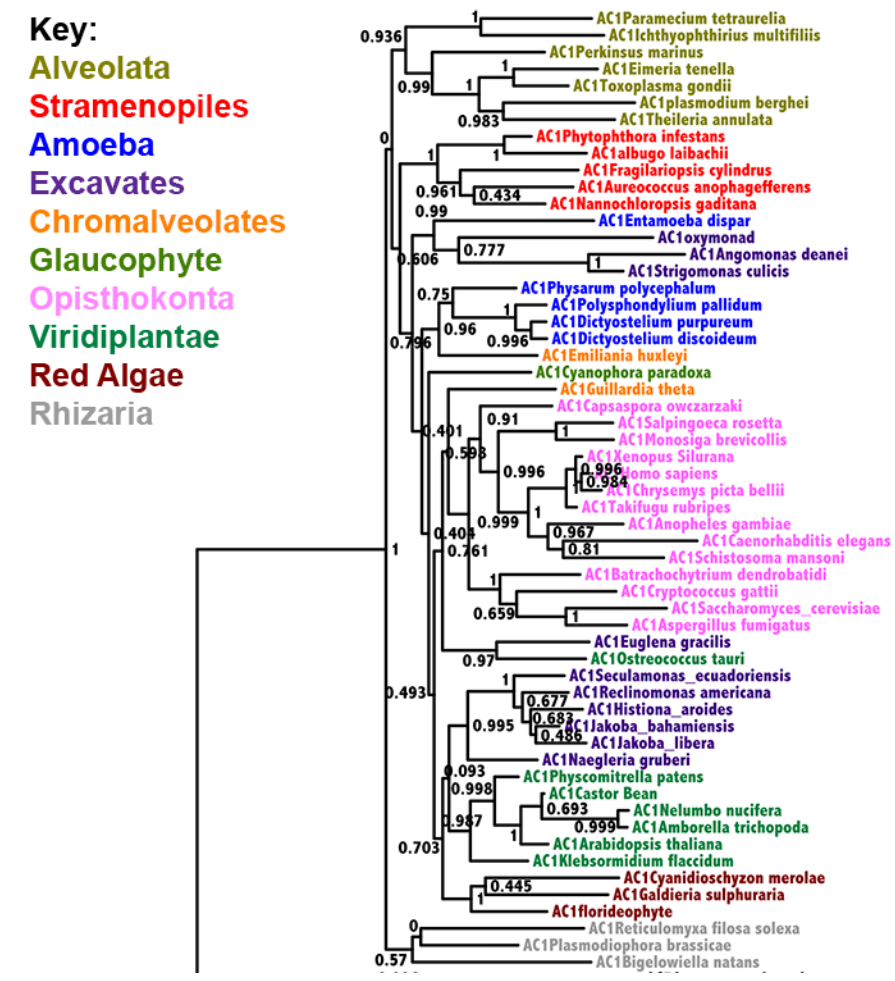


Figure 15: Same as Figure 14, but zoomed in to show only one of the 7 concatenations. Support values represent "SH-like local supports."

6.4. Discussion

It was thought that concatenation might increase resolution and that adding additional taxa would break up long branches, allowing for the determination of the location of the deepest split, and therefore the root within Eukarya. Rooting the eukaryal tree will provide polarization to the tree, which will help future scientists choose between various evolutionary scenarios. Rooting the tree will also help identify synapomorphies, which are needed to properly divide the Eukarya into Kingdoms.

Unfortunately, there is a tradeoff between breaking up long branches and decreasing support values, because the more taxa that are added to an alignment, the lower the support values become. After exploring a myriad of options for using histone, tubulin, and proteasome concatenations to root LECA, there is not enough phylogenetic signal in any of these datasets to produce an answer with confidence. The analogy of throwing a dart across a football field seems to be holding true here and these datasets are not very good at throwing darts.

Given the lack of resolution in these phylogenetic trees and the Long Branch Attraction artifacts, no difference is observed between these trees and those previously published rooting the Eukarya. There do seem to be a number of supergroups that are being seen over and over again; it is just that where the root is within these groups changes. Except, that while I acknowledge a complete lack of trust in root placement, my competitors have thrown fewer darts and published on whichever rooting they happened to find.

The proteasome dataset is not fully exhausted. Taxa suffering from Long Branch Attraction and those that have experienced Horizontal Gene Transfer could still be identified and thrown out. Or alternatively, a supermatrix approach could be used instead of gene

concatenation. But ultimately, that long Eukaryal stem branch leading to the outgroup is the biggest remaining problem. Identifying a gene duplication that occurred just prior to the diversification of LECA would be the ideal solution to this problem. Dr. Gogarten has suggested looking into the Chaperone family with this aim in mind; a task I will leave to incoming graduate students.

7 References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.

Adler, D., Murdoch, D., 2012. RGL: 3D visualization device system (OpenGL).

Allwood, A.C., Grotzinger, J.P., Knoll, A.H., Burch, I.W., Anderson, M.S., Coleman, M.L., Kanik, I., 2009. Controls on development and diversity of Early Archean stromatolites. Proc. Natl. Acad. Sci. 106, 9548–9555.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. doi:10.1016/S0022-2836(05)80360-2

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

Andam, C.P., Williams, D., Gogarten, J.P., 2010a. Natural taxonomy in light of horizontal gene transfer. Biol. Philos. 25, 589–602.

Andam, C.P., Williams, D., Gogarten, J.P., 2010b. Biased gene transfer mimics patterns created through shared ancestry. Proc. Natl. Acad. Sci. 107, 10679–10684.

Archibald, J.M., Logsdon Jr, J.M., Doolittle, W.F., 2000. Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in CCT genes. Mol. Biol. Evol. 17, 1456–1466.

Arisue, N., Hasegawa, M., Hashimoto, T., 2005. Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. Mol. Biol. Evol. 22, 409–420.

Baldauf, S.L., 2003. The deep roots of eukaryotes. Science 300, 1703–1706.

Baldauf, S.L., Palmer, J.D., Doolittle, W.F., 1996a. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. 93, 7749–7754.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., Doolittle, W.F., 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290, 972–977.

Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R.L., Doolittle, W.F., 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evol. Biol. 5, 33.

Basaran, P., Basaran, N., Cakir, I., 2001. Molecular differentiation of Lactococcus lactis subspecies lactis and cremoris strains by ribotyping and site specific-PCR. Curr. Microbiol. 42, 45–48.

Battistuzzi, F.U., Hedges, S.B., 2009. A major clade of prokaryotes with ancient adaptations to life on land. Mol. Biol. Evol. 26, 335–343.

Beiko, R.G., Harlow, T.J., Ragan, M.A., 2005. Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. U. S. A. 102, 14332–14337.

Bell, S.D., White, M.F., 2010. Archaeal chromatin organization, in: Bacterial Chromatin. Springer, pp. 205–217.

Berney, C., Pawlowski, J., 2006. A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. Proc. R. Soc. Lond. B Biol. Sci. 273, 1867–1872.

Bhattacharya, D., Medlin, L., 1995. The phylogeny of plastids: a review based on comparisons of small-subunit ribosomal RNA coding regions. J. Phycol. 31, 489–498.

Biller, S.J., Berube, P.M., Lindell, D., Chisholm, S.W., 2015. Prochlorococcus: the structure and function of collective diversity. Nat Rev Micro 13, 13–27.

Blackstone, N.W., 2013. Why did eukaryotes evolve only once? Genetic and energetic aspects of conflict and conflict mediation. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 368, 20120266. doi:10.1098/rstb.2012.0266

Boussau, B., Gouy, M., 2012. What genomes have to say about the evolution of the Earth. Gondwana Res. 21, 483–494.

Boussau, B., Gueguen, L., Gouy, M., 2008. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. BMC Evol Biol 8, 272.

Brown, J.R., 2003. Ancient horizontal gene transfer. Nat. Rev. Genet. 4, 121–132.

Brown, J.R., Doolittle, W.F., 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc. Natl. Acad. Sci. 92, 2441–2445.

Brown, J.R., Robb, F.T., Weiss, R., Doolittle, W.F., 1997. Evidence for the early divergence of tryptophanyl-and tyrosyl-tRNA synthetases. J. Mol. Evol. 45, 9–16.

Butterfield, N.J., Knoll, A.H., Swett, K., 1985. A Bangiophyte Red Alga from the Proterozoic of arctic Canada. J Appl Phys 58, 1456.

Caro-Quintero, A., Ritalahti, K.M., Cusick, K.D., Löffler, F.E., Konstantinidis, K.T., 2012. The chimeric genome of Sphaerochaeta: nonspiral spirochetes that break with the prevalent dogma in spirochete biology. mBio 3. doi:10.1128/mBio.00025-12

Cavalier-Smith, T., 2014. The neomuran revolution and phagotrophic origin of eukaryotes and cilia in the light of intracellular coevolution and a revised tree of life. Cold Spring Harb. Perspect. Biol. 6, a016006.

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

Cavalier-Smith, T., 1992. The number of symbiotic origins of organelles. Biosystems 28, 91–106.

Cavalier-Smith, T., 1987. The origin of eukaryotic and archaebacterial cells. Ann. N. Y. Acad. Sci. 503, 17–54.

Cavalier-Smith, T., 1982. The origins of plastids. Biol. J. Linn. Soc. 17, 289–306. doi:10.1111/j.1095-8312.1982.tb02023.x

Cavalier-Smith, T., Chao, E.E., 2010. Phylogeny and evolution of apusomonadida (protozoa: apusozoa): new genera and species. Protist 161, 549–576.

Ciccarelli, F.D., Doerks, T., Von Mering, C., Creevey, C.J., Snel, B., Bork, P., 2006. Toward automatic reconstruction of a highly resolved tree of life. science 311, 1283–1287.

Clamp, M., Cuff, J., Searle, S.M., Barton, G.J., 2004. The jalview java alignment editor. Bioinformatics 20, 426–427.

Collins, L., Penny, D., 2005. Complex spliceosomal organization ancestral to extant eukaryotes. Mol. Biol. Evol. 22, 1053–1066.

Cox, C.J., Foster, P.G., Hirt, R.P., Harris, S.R., Embley, T.M., 2008. The archaebacterial origin of eukaryotes. Proc. Natl. Acad. Sci. 105, 20356–20361.

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

de Bary, A., 1879. Die Erscheinung der Symbiose: Vortrag gehalten auf der Versammlung Deutscher Naturforscher und Aerzte zu Cassel. Trübner.

DeGrasse, J.A., DuBois, K.N., Devos, D., Siegel, T.N., Sali, A., Field, M.C., Rout, M.P., Chait, B.T., 2009. Evidence for a shared nuclear pore complex architecture that is conserved from the last common eukaryotic ancestor. Mol. Cell. Proteomics 8, 2119–2130.

Derelle, R., Lang, B.F., 2012. Rooting the Eukaryotic Tree with Mitochondrial and Bacterial Proteins. Mol. Biol. Evol. 29, 1277–1289. doi:10.1093/molbev/msr295

Des Marais, D.J., 1997. Isotopic evolution of the biogeochemical carbon cycle during the Proterozoic Eon. Org. Geochem. 27, 185–193.

Doolittle, W.F., 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet. 14, 307–311.

Doolittle, W.F., Brunet, T.D., 2016. What Is the Tree of Life? PLOS Genet 12, e1005912.

Douzery, E.J., Snell, E.A., Bapteste, E., Delsuc, F., Philippe, H., 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? Proc. Natl. Acad. Sci. U. S. A. 101, 15386–15391.

Dykhuizen, D.E., Green, L., 1991. Recombination in Escherichia coli and the definition of biological species. J. Bacteriol. 173, 7257–7268.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Eisen, J.A., 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr. Opin. Genet. Dev. 10, 606–611.

Embley, T.M., Williams, T.A., 2015. Evolution: Steps on the road to eukaryotes. Nature 521, 169–170. doi:10.1038/nature14522

Eme, L., Sharpe, S.C., Brown, M.W., Roger, A.J., 2014. On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. Cold Spring Harb Perspect Biol 6, 165–180.

Ettema, T.J., Lind\a as, A.-C., Bernander, R., 2011. An actin-based cytoskeleton in archaea. Mol. Microbiol. 80, 1052–1061.

Evolution of Life - Fossils, Molecules and Culture | Syozo Osawa | Springer, n.d.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.

Field, M.C., Dacks, J.B., 2009. First and last ancestors: reconstructing evolution of the endomembrane system with ESCRTs, vesicle coat proteins, and nuclear pore complexes. Curr. Opin. Cell Biol. 21, 4–13.

Foster, P.G., Cox, C.J., Embley, T.M., 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos. Trans. R. Soc. Lond. B Biol. Sci. 364, 2197–2207.

Fournier, G.P., Dick, A.A., Williams, D., Gogarten, J.P., 2011. Evolution of the archaea: emerging views on origins and phylogeny. Res. Microbiol. 162, 92–98.

Fournier, G.P., Gogarten, J.P., 2010. Rooting the ribosomal tree of life. Mol. Biol. Evol. 27, 1792–1801.

Gertz, E.M., Yu, Y.-K., Agarwala, R., Schäffer, A.A., Altschul, S.F., 2006. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 4, 1.

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., 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. U. S. A. 86, 6661–6665.

Gogarten, J.P., Townsend, J.P., 2005. Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3, 679–687. doi:10.1038/nrmicro1204

Goodman, M., Czelusniak, J., Koop, B.F., Tagle, D.A., Slightom, J.L., 1987. Globins: a case study in molecular phylogeny, in: Cold Spring Harbor Symposia on Quantitative Biology. Cold Spring Harbor Laboratory Press, pp. 875–890.

Gribaldo, S., Brochier-Armanet, C., 2006. The Origin and Evolution of Archaea: A State of the Art. Philos. Trans. Biol. Sci. 361, 1007–1022.

Gribaldo, S., Cammarano, P., 1998. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. J. Mol. Evol. 47, 508–516.

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

Group, T.A.P., 1998. An Ordinal Classification for the Families of Flowering Plants. Ann. Mo. Bot. Gard. 85, 531–553. doi:10.2307/2992015

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321. doi:10.1093/sysbio/syq010

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

Guindon, S., Lethiec, F., Duroux, P., Gascuel, O., 2005. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 33, W557–W559.

Hartman, H., Fedorov, A., 2002. The origin of the eukaryotic cell: a genomic investigation. Proc. Natl. Acad. Sci. 99, 1420–1425.

He, D., Fiz-Palacios, O., Fu, C.-J., Fehling, J., Tsai, C.-C., Baldauf, S.L., 2014. An Alternative Root for the Eukaryote Tree of Life. Curr. Biol. 24, 465–470. doi:10.1016/j.cub.2014.01.036

Hennig, W., 1975a. "Cladistic analysis of cladistic classification?": A reply to Ernst Mayr. Syst. Biol. 24, 244–256.

Hennig, W., 1966. Phylogenetic systematics. University of Illinois Press, Urbana.

Hilario, E., Gogarten, J.P., 1993. Horizontal transfer of ATPase genes—the tree of life becomes a net of life. Biosystems 31, 111–119.

Hori, H., Itoh, T., Osawa, S., 1982. The Phylogenic Structure of the Metabacteria. Zentralblatt Für Bakteriol. Mikrobiol. Hyg. Abt Orig. C Allg. Angew. Ökol. Mikrobiol. 3, 18–30. doi:10.1016/S0721-9571(82)80050-1

Hori, H., Osawa, S., 1987. Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences. Mol. Biol. Evol. 4, 445–472.

Hori, H., Osawa, S., 1979. Evolutionary change in 5S RNA secondary structure and a phylogenic tree of 54 5S RNA species. Proc. Natl. Acad. Sci. U. S. A. 76, 381–385.

Huang, J., Gogarten, J.P., 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biol. 8, 1.

Huang, J., Xu, Y., Gogarten, J.P., 2005. The presence of a haloarchaeal type tyrosyl-tRNA synthetase marks the opisthokonts as monophyletic. Mol. Biol. Evol. 22, 2142–2146.

Huelsenbeck, J., Hillis, D., 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42, 247–264.

Huelsenbeck, J.P., Ronquist, F., others, 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Huet, J., Schnabel, R., Sentenac, A., Zillig, W., 1983. Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type. EMBO J. 2, 1291.

Huxley, J.S., 1959a. Clades and grades. Funct. Taxon. Importance Ed AJ Cain 21–22.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., Miyata, T., 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. 86, 9355–9359.

Javaux, E.J., Knoll, A.H., Walter, M.R., 2001. Morphological and ecological complexity in early eukaryotic ecosystems. Nature 412, 66–69.

Javaux, E.J., Marshall, C.P., Bekker, A., 2010. Organic-walled microfossils in 3.2-billion-year-old shallow-marine siliciclastic deposits. Nature 463, 934–938.

Jékely, G., 2007. Origin of eukaryotic endomembranes: a critical evaluation of different model scenarios, in: Eukaryotic Membranes and Cytoskeleton. Springer, pp. 38–51.

Jukes, T., Cantor, C., 1969. Evolution of protein molecules. Munro HN Ed Mamm. Protein Metab. Acad. Press N. Y.

Kandler, O., 1995. Cell wall biochemistry in Archaea and its phylogenetic implications. J. Biol. Phys. 20, 165–169.

Kaplan, W., Littlejohn, T.G., 2001. Swiss-PDB viewer (deep view). Brief. Bioinform. 2, 195–197.

Keeling, P.J., 2014. The impact of history on our perception of evolutionary events: Endosymbiosis and the origin of eukaryotic complexity. Cold Spring Harb Perspect Biol 6, a016196.

Koonin, E.V., 2015. Archaeal ancestors of eukaryotes: not so elusive any more. BMC Biol. 13, 84.

Koonin, E.V., 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. Genome Biol 11, 209.

Koski, L.B., Golding, G.B., 2001. The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52, 540–542. doi:10.1007/s002390010184

Koumandou, V.L., Wickstead, B., Ginger, M.L., van der Giezen, M., Dacks, J.B., Field, M.C., 2013. Molecular paleontology and complexity in the last eukaryotic common ancestor. Crit. Rev. Biochem. Mol. Biol. 48, 373–396.

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., others, 2007. EMBL nucleotide sequence database in 2006. Nucleic Acids Res. 35, D16–D20.

Lake, J.A., Larsen, J., Sarna, B., Rafael, R., Pu, Y., Koo, H., Zhao, J., Sinsheimer, J.S., 2015. Rings Reconcile Genotypic and Phenotypic Evolution within the Proteobacteria. Genome Biol. Evol. 7, 3434–3442.

Lane, N., 2011. Energetics and genetics across the prokaryote-eukaryote divide. Biol Direct 6, 35.

Lapierre, P., Lasek-Nesselquist, E., Gogarten, J.P., 2014. The impact of HGT on phylogenomic reconstruction methods. Brief. Bioinform. 15, 79–90. doi:10.1093/bib/bbs050

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., others, 2007. Clustal W and Clustal X version 2.0. bioinformatics 23, 2947–2948.

Lasek-Nesselquist, E., Gogarten, J.P., 2013b. The effects of model choice and mitigating bias on the ribosomal tree of life. Mol. Phylogenet. Evol. 69, 17–38. doi:10.1016/j.ympev.2013.05.006

Lester, L., Meade, A., Pagel, M., 2006. The slow road to the eukaryotic genome. BioEssays 28, 57–64.

Linnaeus, C., 1758. Systema naturae 1. Editio Decima, Reformata (Holmiae, fasc reprint Bri Mus Nat Hist, London 1939.

Lovejoy, A., n.d. 0. 1936. The great chain of being: a study of the history of an idea. Cambridge, Mass.: Harvard University Press.

Makarova, K.S., Yutin, N., Bell, S.D., Koonin, E.V., 2010. Evolution of diverse cell division and vesicle formation systems in Archaea. Nat. Rev. Microbiol. 8, 731–741.

Maloy, A., Culloty, S., Slater, J., 2009. Use of PCR-DGGE to investigate the trophic ecology of marine suspension feeding bivalves. Mar Ecol Prog Ser 381, 109–118.

Margulis, L., 1995. Symbiosis in Cell Evolution: Microbial Communities in the Archean and Proterozoic Eons. W H Freeman & Co.

Marsh, T.L., Reich, C.I., Whitelock, R.B., Olsen, G.J., 1994. Transcription factor IID in the Archaea: sequences in the Thermococcus celer genome would encode a product closely related to the TATA-binding protein of eukaryotes. Proc. Natl. Acad. Sci. 91, 4180–4184.

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

Mayr, E., 1974. Cladistic analysis or cladistic classification? J. Zool. Syst. Evol. Res. 12, 94–128.

McNulty, S.N., Abubucker, S., Simon, G.M., Mitreva, M., McNulty, N.P., Fischer, K., Curtis, K.C., Brattig, N.W., Weil, G.J., Fischer, P.U., 2012. Transcriptomic and proteomic analyses of a Wolbachia-free filarial parasite provide evidence of trans-kingdom horizontal gene transfer. PloS One 7, e45777. doi:10.1371/journal.pone.0045777

Mindell, D.P., 1992. Phylogenetic consequences of symbioses: Eukarya and Eubacteria are not monophyletic taxa. Biosystems 27, 53–62.

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

Nasir, A., Kim, K.M., Caetano-Anollés, G., 2015. Lokiarchaeota: eukaryote-like missing links from microbial dark matter? Trends Microbiol. 23, 448–450.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garrett, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A., White, O., Salzberg, S.L., Smith, H.O., Venter, J.C., Fraser, C.M., 1999a. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399, 323–329. doi:10.1038/20601

Neumann, N., Lundin, D., Poole, A.M., 2010. Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor. PloS One 5, e13241.

Notredame, C., Higgins, D.G., Heringa, J., 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205–217.

Nunoura, T., Takaki, Y., Kakuta, J., Nishi, S., Sugahara, J., Kazama, H., Chee, G.-J., Hattori, M., Kanai, A., Atomi, H., others, 2011. Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. Nucleic Acids Res. 39, 3204–3223.

Olendzenski, L., Gogarten, J.P., 2009. Evolution of Genes and Organisms. Ann. N. Y. Acad. Sci. 1178, 137–145.

Olsen, G.J., Woese, C.R., 1997. Archaeal genomics: an overview. Cell 89, 991–994.

Pace, N.R., Olsen, G.J., Woese, C.R., 1986. Ribosomal RNA phylogeny and the primary lines of evolutionary descent. Cell 45, 325–326.

Pace, N.R., Sapp, J., Goldenfeld, N., 2012. Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. Proc. Natl. Acad. Sci. 109, 1011–1018.

Parfrey, L.W., Lahr, D.J., Knoll, A.H., Katz, L.A., 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc. Natl. Acad. Sci. 108, 13624–13629.

Peabody, C.R., Chung, Y.J., Yen, M.-R., Vidal-Ingigliardi, D., Pugsley, A.P., Saier, M.H., 2003. Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. Microbiology 149, 3051–3072. doi:10.1099/mic.0.26364-0

Pei, J., Kim, B.-H., Grishin, N.V., 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 36, 2295–2300.

Pisani, D., Cotton, J.A., McInerney, J.O., 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol. Biol. Evol. 24, 1752–1760.

Pittis, A.A., Gabaldón, T., 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature 531, 101–104. doi:10.1038/nature16941

Podell, S., Gaasterland, T., 2007. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biol. 8, R16. doi:10.1186/gb-2007-8-2-r16

Poole, A.M., Neumann, N., 2011. Reconciling an archaeal origin of eukaryotes with engulfment: a biologically plausible update of the Eocyte hypothesis. Res. Microbiol. 162, 71–76.

Poole, A.M., Penny, D., 2007. Evaluating hypotheses for the origin of eukaryotes. Bioessays 29, 74–84.

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS One 5, e9490.

R Developmental Core Team, 2008. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A., Drummond, A., 2008. FigTree: Tree figure drawing tool, version 1.2. 2. Inst. Evol. Biol. Univ. Edinb.

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. CABIOS 13, 235–238.

Richards, T.A., Cavalier-Smith, T., 2005. Myosin domain evolution and the primary divergence of eukaryotes. Nature 436, 1113–1118.

Rivera, M.C., Lake, J.A., 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431, 152–155. doi:10.1038/nature02848

Roberts, D.M., 1996. Evolution of Microbial Life. Cambridge University Press.

Rochette, N.C., Brochier-Armanet, C., Gouy, M., 2014. Phylogenomic test of the hypotheses for the evolutionary origin of eukaryotes. Mol. Biol. Evol. 31, 832–845.

Roger, A.J., 1999. Reconstructing early events in eukaryotic evolution. Am. Nat. 154, S146–S163.

Rogozin, I.B., Basu, M.K., Csürös, M., Koonin, E.V., 2009. Analysis of Rare Genomic Changes Does Not Support the Unikont–Bikont Phylogeny and Suggests Cyanobacterial Symbiosis as the Point of Primary Radiation of Eukaryotes. Genome Biol. Evol. 1, 99–113. doi:10.1093/gbe/evp011

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinforma. Oxf. Engl. 19, 1572–1574.

Ruepp, A., Graml, W., Santos-Martinez, M.L., Koretke, K.K., Volker, C., Mewes, H.W., Frishman, D., Stocker, S., Lupas, A.N., Baumeister, W., 2000. The genome sequence of the thermoacidophilic scavenger Thermoplasma acidophilum. Nature 407, 508–513. doi:10.1038/35035069

Sagan, L., 1967. On the origin of mitosing cells. J. Theor. Biol. 14, 225-IN6. doi:10.1016/0022-5193(67)90079-3

Saldanha, A.J., 2004. Java Treeview—extensible visualization of microarray data. Bioinformatics 20, 3246–3248.

Sheppard, S.K., McCarthy, N.D., Falush, D., Maiden, M.C., 2008. Convergence of Campylobacter species: implications for bacterial evolution. Science 320, 237–239.

Simpson, A.G., 2003. Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). Int. J. Syst. Evol. Microbiol. 53, 1759–1777.

Sogin, M.L., 1991. Early evolution and the origin of eukaryotes. Curr. Opin. Genet. Dev. 1, 457–463.

Sogin, M.L., Elwood, H.J., Gunderson, J.H., 1986. Evolutionary diversity of eukaryotic small-subunit rRNA genes. Proc. Natl. Acad. Sci. 83, 1383–1387.

Sogin, M.L., Gunderson, J.H., Elwood, H.J., Alonso, R.A., Peattie, D.A., 1989. Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from Giardia lamblia. Science 243, 75–77.

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

Stechmann, A., Cavalier-Smith, T., 2002. Rooting the eukaryote tree by using a derived gene fusion. Science 297, 89–91.

Stiller, J.W., Duffield, E.C., Hall, B.D., 1998. Amitochondriate amoebae and the evolution of DNA-dependent RNA polymerase II. Proc. Natl. Acad. Sci. 95, 11769–11774.

Stiller, J.W., Hall, B.D., 1997. The origin of red algae: Implications for plastid evolution. Proc. Natl. Acad. Sci. 94, 4520–4525.

Swithers, K.S., 2012. Genome Structure, Gene transfer and Innovation. UNIVERSITY OF CONNECTICUT.

Swofford, D.L., Waddell, P.J., Huelsenbeck, J.P., Foster, P.G., Lewis, P.O., Rogers, J.S., 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50, 525–539.

Thatcher, T.H., Gorovsky, M.A., 1994. Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. Nucleic Acids Res. 22, 174–179.

The Proterozoic Biosphere [WWW Document], n.d. . Camb. Univ. Press. URL http://admin.cambridge.org/pt/academic/subjects/earth-and-environmental-

science/palaeontology-and-life-history/proterozoic-biosphere-multidisciplinary-study?format=PB (accessed 11.18.16).

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., Martin, W., 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. 5, 123–135. doi:10.1038/nrg1271

Villarreal, L.P., 2016. Viruses and the placenta: the essential virus first view. Apmis 124, 20–30.

Wacey, D., Kilburn, M.R., Saunders, M., Cliff, J., Brasier, M.D., 2011. Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. Nat. Geosci. 4, 698–702.

Whittaker, R.H., 1969. New concepts of kingdoms of organisms. Science 163, 150–160.

Wideman, J.G., Gawryluk, R.M.R., Gray, M.W., Dacks, J.B., 2013. The ancient and widespread nature of the ER-Mitochondria Encounter Structure. Mol. Biol. Evol. mst120. doi:10.1093/molbev/mst120

Williams, D., Fournier, G.P., Lapierre, P., Swithers, K.S., Green, A.G., Andam, C.P., Gogarten, J.P., 2011. A rooted net of life. Biol Direct 6, 45.

Williams, T.A., Embley, T.M., 2014. Archaeal "dark matter" and the origin of eukaryotes. Genome Biol. Evol. 6, 474–481.

Williams, T.A., Foster, P.G., Cox, C.J., Embley, T.M., 2013. An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504, 231–236.

Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. 74, 5088–5090.

Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. 87, 4576–4579.

Woese, C.R., Magrum, L.J., Fox, G.E., 1978. Archaebacteria. J. Mol. Evol. 11, 245–252.

Wolf, Y.I., Makarova, K.S., Yutin, N., Koonin, E.V., 2012. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol. Direct 7, 1–15.

Worning, P., Jensen, L.J., Nelson, K.E., Brunak, S., Ussery, D.W., 2000. Structural analysis of DNA sequence: evidence for lateral gene transfer in Thermotoga maritima. Nucleic Acids Res. 28, 706–709.

Yutin, N., Koonin, E.V., 2012. Archaeal origin of tubulin. Biol Direct 7.

Yutin, N., Wolf, M.Y., Wolf, Y.I., Koonin, E.V., 2009. The origins of phagocytosis and eukaryogenesis. Biol Direct 4, 6150–4.

Zhaxybayeva, O., Doolittle, W.F., Papke, R.T., Gogarten, J.P., 2009a. Intertwined evolutionary histories of marine Synechococcus and Prochlorococcus marinus. Genome Biol. Evol. 1, 325–339.

Zhaxybayeva, O., Gogarten, J.P., 2007. Horizontal gene transfer, gene histories and the root of the tree of life, in: Pudritz, R.E., Higgs, P.G., Stone, J. (Eds.), Planetary Systems and the Origins of Life (Cambridge Astrobiology). Cambridge University Press, Cambridge, UK, pp. 178–192.

Zhaxybayeva, O., Gogarten, J.P., Charlebois, R.L., Doolittle, W.F., Papke, R.T., 2006. Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. Genome Res. 16, 1099–1108. doi:10.1101/gr.5322306

Zhaxybayeva, O., Swithers, K.S., Lapierre, P., Fournier, G.P., Bickhart, D.M., DeBoy, R.T., Nelson, K.E., Nesbø, C.L., Doolittle, W.F., Gogarten, J.P., Noll, K.M., 2009b. On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proc. Natl. Acad. Sci. U. S. A. 106, 5865–5870. doi:10.1073/pnas.0901260106

Personal use

Authors can use their articles, in full or in part, for a wide range of scholarly, non-commercial purposes as outlined below:

- Use by an author in the author's classroom teaching (including distribution of copies, paper or electronic)
- Distribution of copies (including through e-mail) to known research colleagues for their personal use (but not for Commercial Use)
- Inclusion in a thesis or dissertation (provided that this is not to be published commercially)
- Use in a subsequent compilation of the author's works
- Extending the Article to book-length form
- Preparation of other derivative works (but not for Commercial Use)
- Otherwise using or re-using portions or excerpts in other works

These rights apply for all Elsevier authors who publish their article as either a subscription article or an open access article. In all cases we require that all Elsevier authors always include a full acknowledgement and, if appropriate, a link to the final published version hosted on Science Direct.