Utilization of Phylogenetic Analysis Methods to Understand Deep Phylogeny

Jeffrey M. O'Brien
University of Connecticut - Storrs, obrien.jeffrey@gmail.com

Recommended Citation
http://digitalcommons.uconn.edu/gs_theses/1079
Utilization of Phylogenetic Analysis Methods to Understand Deep Phylogeny

Jeffrey M. O’Brien

B.S., University of Connecticut, 2015

A Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science
At the
University of Connecticut
2017
Master of Science Thesis
Utilization of Phylogenetic Analysis Methods to Understand Deep Phylogeny

Presented by
Jeffrey M. O’Brien, B.S.

Major Advisor_______________________________________________
J. Peter Gogarten

Associate Advisor____________________________________________
R. Thane Papke

Associate Advisor____________________________________________
Jonathan Klassen

University of Connecticut
2017
Acknowledgements:

I would like to thank my advisor, Peter Gogarten. He has been my teacher and guide in learning bioinformatics and molecular evolution over the course of three years in lab as an undergraduate and another year as a graduate student. This work would not have been possible without his enduring support and guidance. I would like to acknowledge Tim Harlow for providing the framework for one of the scripts central to this analysis, as well as all of the Gogarten lab members. I would also like to acknowledge the UConn Bioservices Center for providing access to the computational resources necessary to do this work.

I would like to thank my parents, Donna and Shawn O’Brien. Their support and love over the years has encouraged me to accomplish my goals and pursue my dream of going on to medical school. Thank you for always being there for me and helping me become the person I am today.
Table of Contents:

Chapter 1 ......................................................................................................................... 1
  Introduction .................................................................................................................. 1
  Methods ....................................................................................................................... 4
  Results ......................................................................................................................... 7

Chapter 2 ......................................................................................................................... 19
  Introduction ................................................................................................................. 19
  Methods ....................................................................................................................... 23
  Results ......................................................................................................................... 28

Appendix 1 ..................................................................................................................... 40

Works Cited .................................................................................................................... 43

Scripts mentioned in text can be found on GitHub:

<https://github.com/jmo11001/masters_thesis>
Chapter 1: Applying Tree Metrics and Midpoint Rooting to Phylogenies to Infer Ancient Evolutionary Origins

Introduction:

The last universal common ancestor (LUCA), or the cenancestor of all cellular life, was a prokaryotic organism that shares certain functional similarities with extant organisms. LUCA can be better understood as a population of genetically similar cells that gave rise to all cellular life currently inhabiting the planet. This organism had a genome composed of DNA, encoded for the production of 20 amino acids, used ribosomes for protein synthesis based on an RNA transcript template, and utilized membranes capable of chemiosmotic coupling in energy production (Gogarten and Taiz 1992; Goldman et al. 2013). This definition is distinct from that of the progenote, which was originally described by Carl Woese and George Fox in the 80’s as an organism in the process of linking genotype and phenotype (Woese and Fox 1977).

Studies of highly conserved proteins have offered a view of the lifestyle and traits possessed by LUCA. In addition to those already mentioned, LUCA had tRNAs and a variety of elongation factors (Gogarten and Taiz 1992). A more recent phylogenetic analysis was undertaken to determine protein families that were present in LUCA (Weiss et al. 2016). Several of their stated criteria for including a protein family as present in LUCA are questionable. For instance, in order for a protein sequence cluster to be labeled as originating in LUCA, it only needed to have representation for a minimum of 2 major groups in each domain corresponding primarily to order level (and sometimes phylum level) classification. Furthermore, adequate sequence representation for a group was a minimum of only 2 sequences. This means that in
any given dataset there may only be 4 sequences corresponding to either the bacteria or archaea. As can be seen in supplementary table 2 in their publication, many of the families they identified have very sparse coverage in one domain or another.

In order to trace a protein family with sparse coverage back to LUCA, many independent gene loss events would have to occur. This is parsimoniously unlikely. While the authors claim that their methods prevent identification of protein families in which horizontal gene transfer (HGT) has occurred, this explanation can much more easily explain their results particularly in the case of sparse coverage for a domain (Mirkin et al. 2003; Becerra et al. 2007). Since there methods only require representation in two order level groups within a domain, one can easily imagine a scenario where gene family would attributed as a LUCA trait incorrectly because of a post-LUCA HGT from one domain to another followed by a relatively easier transfer from one order to another within the same domain (Beiko et al. 2005; Andam and Gogarten 2011). The shortcomings of taking analytical shortcuts that prevent the identification of the full scope of a protein family and neglect the importance of HGT in post-LUCA evolution have been explored in depth elsewhere by Gogarten and Deamer (Gogarten and Deamer 2016).

In the current work, phylogenetic metric analyses were run to learn whether the protein families identified by the Weiss et al. do indeed represent ancient evolutionary origins, or if they are more likely distributed as such due to recent transfer events. Branch length metrics and rooting analyses are effective measures that can be used to analyze a batch of phylogenies quantitatively. These methods enable understanding of the dynamic mechanisms underpinning the molecular evolution of protein families (Wolf et al. 1999; Pittis and Gabaldón 2016).
Several proteins that were almost certainly present in LUCA may have been included in their dataset, but due to their filtering methods may not represent the full range of sequence diversity for that family. For that reason, datasets were manually curated for ATP synthase catalytic subunits for both bacteria and archaea, as well as a collection consisting of the RNA Polymerase B subunit from crenarchaeotes, B and B’ subunits from euryarchaeotes, and the β-subunit from bacteria. These proteins were certainly present in LUCA (Johann Peter Gogarten et al. 1989; Puhler et al. 1989), and by ensuring that a range of sequence diversity is present, we could use these phylogenies for comparison to the 355 phylogenies curated by the Weiss et al. research cohort.
Methods:

Sequence Data Acquisition

Protein sequence alignments were provided by Bill Martin (University of Hanover, Germany), the corresponding author on the publication describing 355 protein families that were purportedly present in LUCA based on their phylogenetic analysis criterion (Madeline C Weiss et al. 2016). The files were in phylip format, and each sequence was named by a GI number. All 355 alignment files were converted to fasta format using the script phylip2fasta.pl, copied from stackoverflow.com.

Renaming

To facilitate downstream analysis, sequences were renamed according to group abbreviations (Supplementary figure 1, Appendix 1). This was done by modifying a PERL script provided by Tim Harlow for the current purposes. In short, the script takes as input a set of fasta formatted alignment files and outputs them with the only modification being that the header for each sequence is modified (get_taxonomy.pl, GitHub repository). This is done by querying the NCBI taxonomy database using e-utils. A string is returned representing the hierarchical taxonomy for the organism the sequence came from. This string was used to name each sequence based on the domain, group abbreviation, and accession number (ie “Bact_GP_accession”). The abbreviations were taken from supplemental table two in the original publication (Madeline C Weiss et al. 2016). Supplementary table 2 of this publication lists all abbreviations used. A few group abbreviations had to be added to the list taken from supplementary table 2 of the Weiss et al. paper in order to make the re-naming comprehensive.
**Phylogenetic Reconstruction**

Each alignment was passed to RAxML for phylogenetic reconstruction using the same methods as previously described (Stamatakis 2014; Madeline C Weiss et al. 2016). Briefly, the PROTCATWAG model was used for phylogenetic inference of each tree. Special amino acid characters U, O, and J were converted to C, K, and X. Deviations from these methods included using RAxML v.8.1.17 rather than v.7.8.6, which is what was used in the original publication. No significant changes were seen in the update documentation that would impact the analysis.

In the current analysis, branch ratio analysis was conducted on the 355 RAxML phylogenies. Then, a further step was taken to calculate SH-support values for the branches of the 355 unrooted trees using the “-f J” option and inputting each tree with the “-t” option. The SH-support value trees were used for midpoint rooting analysis.

**Branch Length Analysis**

Each of the 355 maximum likelihood phylogenies was next analyzed using a branch length metric analysis. The ‘distance()’ and ‘common_ancestor()’ commands from the BioPhylo module in Python were used to extract the maximum within group tip to tip traversal branch length and the interdomain branch length respectively (see supplementary figure 1, appendix 1) (Talevich et al. 2012). These commands were incorporated into a script that used a loop structure to analyze all 355 phylogenies in batch (branch_ratio.py, GitHub repository). The tip to tip traversal distance was then divided by the interdomain branch length in order to obtain a ratio that could be used to describe the overall relation of these two values. A tree that has a
relatively large interdomain branch length in comparison to the maximum branch length distance within a group would have a ratio less than 1.

**Midpoint Rooting Analysis**

All 355 trees were opened in the graphical user interface program FigTree. When one types “command + m” with the tree window open, FigTree roots the tree at the midpoint of the longest tip to tip traversal distance within the tree. This operation was performed on each of the maximum likelihood trees. Subsequently, each tree was checked manually to determine if midpoint rooting separated all archaeal sequences on one side of the root and all bacterial sequences on the other. If this topology was not achieved upon midpoint rooting, then the tree file was counted as ‘does not support bacterial and archaeal monophyly upon midpoint rooting’.

**‘Gold Standard’ LUCA Phylogenies**

Sequence datasets were compiled manually for the B, B+B’, and β subunits of RNA polymerase, as well as the catalytic subunit of ATP synthase using NCBI and the web BLASTp algorithm (Camacho et al. 2009). These datasets were aligned using MUSCLE v.3.8.31 and then phylogeny was inferred using RAxML v.8.1.17 using the PROTCATWAG model and 100 bootstrap replicates (Edgar 2004; Stamatakis 2014). The resulting phylogenies were then examined for branch length ratio and midpoint rooting as described above.
Results:

After running maximum likelihood inference on the alignments obtained from Bill Martin (University of Hanover, Germany), custom tests were conducted on the topology of the trees. This was done to determine if there is sufficient phylogenetic evidence to support the authors claim that the 355 protein families in question were present in the last universal common ancestor (LUCA) (Madeline C. Weiss et al. 2016). There was reason to suspect that the gene datasets produced by the bioinformatic analysis conducted by Weiss et al. 2016 in some, many, or most, cases may not have been present in LUCA. Sparse representation in one either archaea or bacteria as seen in many of their alignments would have to be explained by hundreds of independent gene loss events under their interpretations of the methods, but horizontal transfer could more easily explain the observed topologies. Additionally, some gene families that were present in LUCA were likely identified by the methods of Weiss et al., but with many missing orthologous sequences due to a combination of HGT and a stringent clustering cutoff.

Branch Length Ratio Tests

The separation between Archaea and Bacteria represents a fundamental split in biology that usually is reflected in sequence diversity of orthologs that were vertically inherited (Zillig et al. 1989; Gogarten et al. 1989; Iwabe et al. 1989; Brown and Doolittle 1995; Puigbò et al. 2009). A reasonable expectation is that the length of the branch connecting archaeal and bacterial vertically inherited orthologs should be longer than the branches within each of the groups (orders and classes) used by Weiss et al (2016) to assemble their datasets. Therefore, in the first
test, a script was written that calculates the ratio between maximum branch length distance within a taxonomic group and the length of the branch separating the bacteria and archaea on the tree (equation 1). The calculated ratios for the 355 trees identified by Weiss et al. 2016 could then be compared to values obtained from gold standard phylogenies almost certainly present in LUCA. These gold standard phylogenies include ubiquitous and indispensable genes such as the catalytic subunit of ATP synthase and RNA polymerase. Previous analyses on these proteins supports that they were present at the time of the last universal common ancestor (Johann Peter Gogarten et al. 1989; Puhler et al. 1989), even though the latter has been transferred between the bacterial and archaeal domains repeatedly (Hilario and Gogarten 1993).

\[
\text{Ratio} = \frac{\text{Maximum within group distance}}{\text{Distance between Archaea and Bacteria}} \tag{1}
\]

If the protein families that passed the criteria in the Weiss et al. publication were indeed present in LUCA, and each dataset adequately captured the sequence diversity for this protein family, then we would expect this ratio to be less than 1. This is due to the fact that the length of the branch separating the archaea from the bacteria ought to be much longer than any traversal distance tip to tip for two sequences in the same order level taxonomic group, meaning there should be a greater number of substitutions separating the two domains and fewer substitutions separating sequences that come from the same taxonomic group. Essentially, we would expect the bulk of sequence change to lie between the two domains.
The measurements were taken by using commands in various python packages, primarily ‘Phylo’ from Bio Python as well as ‘dendropy’ (Talevich et al. 2012). A visual example of one of the trees from the 355 protein families is shown in figure 1.1. One of the trees from the dataset with relatively few taxa was chosen to display the process of determining the ratio. In 1A only the maximum within group difference is shown, however the in-house python script exhaustively calculates all possible tip to tip traversal routes within a single group. For the Thermoproteales (TP) group there are 7 sequences in this tree, which means that there are actually 21 possible tip to tip traversal routes according to equation 2, which describes how to calculate the number of possible traversal routes in a tree for a given number of tips ‘n’.

$$\text{# of Routes} = \sum_{i=1}^{n-1} i$$  \[2\]

The ‘distance()’ command within the BioPhylo module in python takes two phylogenetic tree tip names as arguments and returns the sum of the branch lengths connecting them. This was the core function used in my in-house python script, but in order to isolate the maximum distance I had to organize the steps taken into an algorithm. First, all sequences in a single group would be added to a list. This was straightforward since I had previously named sequences to have a two-letter group abbreviation between the first and second underscore characters in the sequence name. Next, I used two nested for loops to compare each leaf in the list with each other leaf individually. Essentially, I passed every possible pair of arguments to the ‘distance()’ function and save the value into a new list (albeit with the harmless redundancy of each pair occurring twice). These steps would then be repeated for all of the other groups in the tree; in the one shown above these would be Desulfurococcales (DE), Chloroflexi (CF), and Gammaproteobacteria (GP). The traversal distances between tips within
Figure 1.1: Visual aid for understanding within group distance to interdomain branch length ratio calculation. Above are two representations of a single tree calculated from protein family 3621 in the Weiss et al. 2016 dataset. Phylogenetic inference was done in RAxML and the trees were visualized and colored in FigTree. (A) In the upper rendering of the tree the maximum within group distance has been highlighted in red. This distance represents the tip to tip traversal distance between two Thermoproteales (TP) sequences in the phylogeny. The value of this distance is 4.07, and this corresponds to the numerator in equation 1. (B) In the lower rendering of the tree, the branch connecting the Archaeal and Bacterial sequences on the tree has been colored in gold. The interdomain branch length is $2 \times 0.7043 = 1.41$, and this value corresponds to the denominator in equation 1. The calculated ratio for this tree is then 2.88 when with manual calculation, which corresponds to the value determined by the python script (see supplemental table 1).
these groups would then be appended to the running list of values mentioned previously. From here it is straightforward to isolate the largest distance using the ‘max()’ function in python. You will notice that the groups with only one leaf, Sulfolobes (SU) and Betaproteobacteria (BP), were excluded due to the fact that they would not have a within group traversal distance. This combination of steps reproducibly isolates the numerator that is used in the ratio calculation shown in equation 1.

In order to isolate the denominator for this equation, the ‘common_ancestor()’ function in the BioPhylo module was called. This function takes a list of tips in a phylogenetic tree and returns information about the branch connecting this group to the rest of the tree. Again, I took advantage of the way I named my sequences to isolate all archaeal sequences into a list called “arch_list” and all the bacterial sequences into a list called “bact_list”. Next, each of these lists were passed to the common ancestor function as an argument in separate calls. Python interpreted these trees as rooted, so only one of the two domain tip lists would actually return a non-zero branch length when the ‘common_ancestor()’ function was called. The accuracy of this command was verified manually for several trees using FigTree. Then, using conditional statements, I was able to return the interdomain branch length for use as the denominator in equation 1. When considered in conjunction with the maximum within group distance value a branch length ratio could now be calculated for each of the 355 trees. A list of these values can be found in supplementary table 2.

In order to visualize the data graphically, they were read into the RStudio platform, which was running v3.3.2 (R Core Team, 2016). First, a histogram was made using ‘hist()’ from
Figure 1.2: A limited range histogram displaying maximum within group distance to interdomain branch length ratio. In the above figure, it can be seen that very few of the trees have a ratio between 0 and 1, which is the ratio expected for protein families present in LUCA. Keeping in mind that the full dataset includes 355 trees in total, the fact that only 2 trees have ratios between 0 and 1 indicates that the majority of these sequence clusters may not represent a bona fide family present in LUCA. The reduced range of 50 excludes few outliers from the plot. The histograms in supplementary figures 1 and two are more comprehensive and include the outliers at the expense of compressing the majority of the data to the left side of the graph.

The first representation included the entire range of values obtained from the 355 trees and can be found in supplementary figure 1.1 (also see supplementary figure 1.2). Due to the presence of an outlier data point, it was difficult to see the range of the data that was critical to the central question. To see how many of the branch length ratios were less than one, the ratio of the histogram was limited to 50 which eliminated a minority of outlier data points (figure 1.2).
Table 1.1: Summary statistics for the branch length ratio values of 355 trees.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1st Quartile</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quartile</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.000</td>
<td>6.228</td>
<td>12.190</td>
<td>18.100</td>
<td>20.840</td>
<td>724.200</td>
</tr>
</tbody>
</table>

Table 1.2: The same dataset of 355 branch length ratios represented in terms of population percentiles. Based on this analysis it can be stated that 99.5% of the ratio values are greater than 1.

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>2.51</td>
</tr>
<tr>
<td>1%</td>
<td>1.47</td>
</tr>
<tr>
<td>0.50%</td>
<td>1.04</td>
</tr>
</tbody>
</table>

In addition to plotting the data as a histogram, the numeric vector was passed as an argument to the ‘summary()’ function in R. This function collects some basic descriptive statistics about the data including minimum, maximum, mean, median, 1st quartile, and 3rd quartile (Table 1.1). This data considered, in conjunction with the observation garnered from the histogram, tells one that in the majority of trees the interdomain branch is relatively short compared to the maximum within group traversal distance.

To more precisely articulate how many data points fell outside of the 0-1 range, the ‘quantile()’ function in R was called. This function reports the upper bound of the subset of ordered branch length ratios for a given percentage of the population (passed as an argument). It can be seen from these analyses that the problem with this data is more than just a few of their trees falling outside an acceptable branch length ratio range. In fact 99.5%, nearly their entire dataset, does not fall within this expected range.

In summary, only 2 out of 355 trees have an interdomain branch length that is larger than the maximum traversal distance within a taxonomic group. This indicates that for the vast
majority of gene families identified as present in LUCA, the archaeal and bacterial homologs are more related to one another than the most distantly related members of a single taxonomic group within the archaea or bacteria. This is a pattern that one should not expect for protein families present in LUCA.

‘Gold Standard’ LUCA Phylogenies

The summary statistics and plots for the dataset of 355 branch length ratios revealed that the majority of the phylogenies were calculated to have a ratio greater than 1. In fact, this difference was actually rather pronounced with the mean being 18.1 and median of 12.19 for the population (Table 1.1). As mentioned previously, this means that these protein sequence datasets are either not representative of the entire diversity within the protein family or that they were simply not present in LUCA. To further validate this claim, two additional phylogenetic analyses were run in RAxML on genes that were almost certainly present in LUCA, RNA Polymerase and ATP synthase.

The RNA Polymerase dataset was compiled simply by searching the NCBI database. Crenarchaeota sequences were gathered first by searching for the B subunit of RNA polymerase for one taxon and then using this sequence as a query in BLASTp searches filtered based on taxonomic groups (Camacho et al. 2009). A similar course of action was taken for the Euryarchaeota, but by gathering B’ and B’’ subunits, which when pasted end to end are homologous to the Crenarchaeotal B subunit. Next, Bacterial sequences were added by searching for the homologous β subunit. Fig. 1 in Puhler (1989) was used a guide to understand
which RNA polymerase subunits were homologous to each other when constructing the dataset.

For the ATP synthase dataset, the archaeal A-ATPase sequences were collected from an existing dataset for which I had already made a phylogeny in RAxML (See figure 2.2). I chose sequences for each group that had a large traversal distance on this tree. Next, I searched the NCBI database to obtain a sequence for the F-type ATPase Beta catalytic subunit from bacteria. This was then used in subsequent group filtered BLASTp searches to collect a handful of sequences representing a broad spread of genus/family diversity within the group.

Figure 1.3: ATP Synthase catalytic subunit phylogeny of Bacterial and Archaeal sequences. Above is a phylogeny of the catalytic A-ATPase subunits from Archaea and F-type catalytic Beta-ATPase subunits from Bacteria. This tree was inferred by the maximum likelihood program RAxML v.8.1.17 using the PROTCATWAG substitution model (consistent with Weiss et al. methods) and 100 bootstrap replicates (supports not shown). Highlighted in red is the maximum within group traversal distance within this phylogeny, which is found within the Thermoproteales. This distance was measured to be 1.2065. As in the RNA polymerase phylogeny above the inter domain branch length (gold) is larger than any of the within group distances. The length of this branch is 3.1540. By putting these two values into equation 1, this phylogeny is given a branch length ratio of 0.38. Again, the branch length ratio being less than 1 lends support to the fact that protein families present in LUCA ought to fall within this range.
Within both bacteria and archaea several taxonomic groups were selected to search for sequences, and within each of these groups the representative sequences were chosen based on being from either divergent genus or families. In doing this one can ensure that the maximum within group distance is not an underestimate. For an accurate comparison to the ratios from the data compiled by Weiss et al., these sequence files were crafted to cover a broad phylogenetic range within each of the groups included.

As can be seen in figures 1.3 and 1.4, the ‘gold standard’ LUCA phylogenies return branch length ratio values that are far less than one. The interdomain branch length for these trees is relatively large compared to the maximum within group tip to tip traversal distance. The calculated ratio for the RNA Polymerase phylogeny was 0.25 and the ratio for the ATP Synthase phylogeny was 0.38. This long branch separating the two domains indicates that there is more intradomain similarity amongst sequences than there is similarity between the two domains. This is exactly what we would expect for the vast majority of phylogenies of protein families present in LUCA. This conclusion follows naturally from the logic that more substitutions should separate the last archaeal common ancestor and last bacterial common ancestor than for any two species coming from the same group. However, violations to the rule would occur when sequences fall in a different domain on the tree due to a post-LUCA HGT event.

**Midpoint Rooting Analysis**

In addition to calculating branch length metrics on the 355 trees in the dataset, they were also rooted at the midpoint and inspected. This is straightforward to do in the graphical user interface of FigTree by pressing ‘control + m’. This effectively isolates the largest possible
Figure 1.4: Two example trees from the midpoint rooting analysis. (A) Example of a tree that returns domain monophyly when midpoint rooting is done in FigTree. The root can be seen on the left side of the tree, with all of the archaeal sequences above the root and all of the bacteria sequences below. (B) This example does not return domain monophyly upon midpoint rooting in FigTree. One divergent archaeal sequence falls below the root rather than above with the rest of the archaeal sequences. Archaeal sequences and the branches leading to them have been highlighted in red.
traversal distance between two tips in the unrooted tree and then places the root at the midpoint of this branch. Sequences were named with the first four characters corresponding to the domain. This made it straightforward to search tip names using FigTree and visually interpret the topology.

Using the logic described above all 355 trees were rooted and checked for domain monophyly. It was determined for each phylogeny if all bacterial sequences were placed on one side of the root and all archaeal sequences placed on the other. Upon inspection, it was determined that 288 out of 355 trees did support monophyly of both Archaea and Bacteria upon midpoint rooting. An example of a midpoint rooted tree that resolves domain monophyly can be seen in figure 1.3A, and 1.4B represents a tree that does not return domain monophyly upon midpoint rooting.

This analysis was done to lend further support to the previous claim stating that a protein family present in LUCA ought to have a relatively large interdomain branch length when compared to the traversal distances within groups. In midpoint rooting however, we consider the furthest reaches of the tree. If this were a protein family present in LUCA we would expect that the root would be placed between the furthest reaching branch in the Archaea and the furthest reaching branch in the Bacteria. If the interdomain distance were large, then this would likely place the root between the two domains. Since this pattern was not observed for 81.1% of the 355 trees, we can lend further support to the observation that the phylogenies in this dataset are topologically divergent from what we would expect in phylogenies for protein families from LUCA. Both of the ‘gold standard’ phylogenies return domain monophyly when midpoint rooted in FigTree.
Chapter 2: Using ATP Synthase Phylogeny to Infer Deep Evolutionary Relationships Focusing on Newly Discovered Lineages

Introduction

In the past decade or so the number of sequenced organisms in the archaeal domain has skyrocketed, particularly for organisms that were previously uncultivable. This trend has been increasing for some time due to technological advances in single-cell genome sequencing and metagenomic sequencing (Albertsen et al. 2013; Blainey 2013; Gawad et al. 2016). This has led to an expansion in the number of taxa that can be used for phylogenetic reconstruction of the evolutionary history in archaea (Hug et al. 2016).

The novel lineages that have been identified include members of the DPANN superphylum, Hadesarchaeota, Altaiarchaeales, and more recently Theionarchaeota and members of the ASGARD group among others (Rinke et al. 2013; Probst and Moissl-Eichinger 2015; Spang et al. 2015; Baker et al. 2016; Lazar et al. 2017; Zaremba-Niedzwiedzka et al. 2017a). These novel lineages share the common characteristic of being uncultivable to date. Some of these lineages, particularly the representatives of DPANN have very streamlined genomes and relatively high rates of substitution in their protein coding sequences compared to other extant archaea (Castelle et al. 2015).

Particularly for the DPANN group, alternative phylogenetic analyses have reconstructed the evolutionary history of these lineages differently. The research group that originally sequenced and described these organisms found that they group together as a cohesive superphylum basal to both the TACK superphylum lineages and the Euryarchaeota. This
topology was observed in separate maximum likelihood analyses, one based on a concatenation of 38-marker genes and the other based on a concatenation of 15 ribosomal proteins (Rinke et al. 2013; Castelle et al. 2015). Both of these phylogenies revealed that several branches leading to DPANN groups falling basal to extant archaea had high bootstrap support values. However, there have been several analyses that have not reconstructed this same relationship. A Bayesian analysis of a concatenated alignment consisting of 32 ribosomal proteins and 38 new conserved proteins split up the Parvarchaeota, Nanoarchaeota, and Nanohaloarchaeota (Petitjean et al. 2015). Other analyses have also raised suspicion that DPANN may not be a cohesive group, such as one that showed the Nanoarchaeota to group between the Euryarchaeota and the Crenarchaeota under some evolutionary models (Lasek-Nesselquist and Gogarten 2013), and another that places the Nanohaloarchaeota as sister group to the Haloarchaea in a 16S maximum likelihood phylogeny (Narasingarao et al. 2011; Zhaxybayeva et al. 2013).

More recently the ASGARD group has entered the scene as a novel taxonomic classification for four archaeal lineages, Lokiarchaeota, Thorarchaeota, Odinarchaeota, and Heimdallarchaeota (Zaremba-Niedzwiedzka et al. 2017a). One phylogenetic analyses has placed the Eukaryota within the Lokiarchaeota clade (Spang et al. 2015). After more members of the ASGARD group were sequenced a later analysis placed the Eukaryota within this group rather than specifically within the Lokiarchaeota (Zaremba-Niedzwiedzka et al. 2017b). There is reason to suspect that these relationships could be a phylogenetic reconstruction artifact as both the Lokiarchaeota clade and the Eukaryotes are on relatively long branches in comparison to extant archaea. Long-branch attraction tends to group sequences that have a high number of
substitutions per site together even if this does not reflect the true evolutionary history of the sequences (Bergsten 2005).

In the current analysis, a dataset consisting of the A-ATPase catalytic subunit as well as the Eukaryotic V-ATPase was compiled and rendered into a phylogenetic tree. This protein was chosen for the analysis because it is essential for life, ubiquitous, and has one of the lowest substitution rates for known protein sequences. While having a comparatively low substitution rate compared to other proteins due to its importance, when they do occur they are in accordance with rates outlined by common substitution matrices. In addition to these reasons, ATP synthase has a broad and illustrious history of use in phylogenetic analyses, and is especially useful for deciphering deep phylogenetic relationships (Gogarten et al. 1989; Gogarten and Taiz 1992; Lapierre et al. 2006). This gene is not infallible to events that have the potential to distort phylogenetic interpretation, it has in fact been shown to have undergone horizontal gene transfer (HGT) events over the course of evolutionary history (Hilario and Gogarten 1993). Nevertheless, understanding the phylogeny of this gene is useful in our attempt to understand the placement of novel archaeal lineages.

Compositional bias is known to artifactually group sequences together that share similar amino acid or nucleotide preferences at given positions due to environmental pressures (Herbeck et al. 2005; Cox et al. 2008). A discussion of the effects of compositional bias on phylogenetic reconstruction has been covered elsewhere (Lasek-Nesselquist and Gogarten 2013; Rota-Stabelli et al. 2013). One way to deal with this in protein sequence datasets would be to use binning strategies, the most popular of which have been the 6-class and 4-class Dayhoff recoding schemes (Martin et al. 2003; Embley et al. 2002; Hrdy et al. 2004; Susko and
Roger 2007). The current analysis was passed to phylogenetic reconstruction both as a full amino acid dataset and after Dayhoff recoding into 6 and 4 classes.

The goal of the current analysis is to further the understanding about the phylogenetic placement of novel lineages of Archaea. Additionally, the following phylogenetic analyses aim to shed some light on the potential archaeal origins of modern eukaryotes, and how that may relate to the novel ASGARD lineages. Finally, this analysis should add some additional phylogenetic context for understanding the evolutionary relationships of the DPANN lineages.
Methods:

Gathering sequence data:

The National Center for Biotechnology Information (NCBI) genome browser was accessed and filtered for the group “All Archaea” from the Prokaryotes tab on January 5th, 2017. These 891 records were downloaded as an excel spreadsheet. An in-house python script (archaea_NCBI_download_named_by_group.py, GitHub repository) calling the urllib module was used to download the amino acid coding sequence for each genome. This script was used to read the url column in the spreadsheet and access the NCBI ftp server. Then the program would browse the records for each particular organism and download only the protein coding sequences (".faa") containing files. Each local genome file was also given a unique prefix based on taxonomic group.

The genomes from NCBI were supplemented with (60 files) genome records from the Integrated Microbial Genomes and Microbiome samples system (IMG/M) hosted by the Joint Genome Institute (JGI). These data can be accessed using the Globus data management platform. The JGI portal was accessed through Globus on January 23rd, 2017. This online directory system can be browsed by taxonomy. Only genomes from organisms of interest (ie. DPANN) were downloaded to supplement the NCBI dataset to reduce redundancy in well-established clades that already had substantial representation.

Recently published literature was browsed to increase genome representation for ultrasmall and deep-branching lineages (Lazar et al. 2017; Zaremba-Niedzwiedzka et al. 2017a).
All protein records from these sequencing projects were downloaded from Bioproject database entry PRJNA270657 and PRJNA319486, each to a separate single fasta file.

Eukaryotic sequences for the catalytic V-type ATPase were added on a per-genome basis. Sequences were identified by searching the non-redundant BLASTp database and filtering for a particular species or higher order classification (Camacho et al. 2009). The catalytic ATP synthase subunit from the Crenarchaeote *Sulfolobus islandicus* was used as a query for all searches (accession: WP_012711559.1). Two to three representatives from each major lineage were included when available.

**Compiling ATP synthase catalytic subunit sequence dataset:**

The BLASTP algorithm was used to search for catalytic ATP synthase subunit sequences in the 953 archaeal genomes downloaded from NCBI and JGI as “faa” file types (Altschul et al. 1990). The percent identity between catalytic ATP synthase subunit orthologues from different domains in the tree of life tends to be around 50%. There is a steep reduction in percent identity to roughly 20% when one compares the catalytic subunit to the paralogous non-catalytic subunit. For this reason, the catalytic ATP synthase subunit sequence from *Anopheles gambiae* was used as a query for the search (accession: XP_312843.1) A percent identity cutoff of 40% was imposed so as not to incorporate paralogous sequences into the analysis. An e-value cutoff was not imposed because these values were sufficiently low for all BLASTp results incorporated into the dataset (less than 1e -30). When the genome was available in a single file (891/953), only the top hit was added to the dataset.
The environmental non-redundant database of NCBI was searched to increase representation of deep-branching archaeal groups (Lokiarchaeota, Batharchaeota, Geoarchaeota, Diapherotrites, Parvarchaeota, Altiarchaeales, Korarchaeota, Nanoarchaeota, Nanohaloarchaeota, Theionarchaeota, Aigarchaeota, Aciduloprofundum). A separate search was done for each group using the online BLASTp portal (Camacho et al. 2009). In each case the query was an archaeal ATP synthase catalytic subunit sequence from a sequenced genome downloaded from either NCBI or JGI. A stringent cutoff was used (e-value < 10^-100 and percent ID > 60%) and only a handful of the total results were downloaded for each group. These sequences were added to the dataset with the “put” prefix in the sequence header to represent the fact that they were only putative representatives of that particular lineage.

Data Trimming and Sequence Alignment:

After data collection, the sequence file consisted of 604 entries. Sequences were aligned using MUSCLE v3.8.31 (Edgar 2004). 38 partial sequences were identified by eye and removed from the alignment. Gap characters (“-“) were removed and the 566 sequences were aligned again prior to clustering.

The dataset was clustered using Usearch under the “-cluster_fast” algorithm because it provides a fasta output of the centroid sequences when the “-uc” option is used (Edgar and Bateman 2010a). When clustered at 95% identity, this data set is simplified to 366 centroid sequences. As hoped for, only 1 out of 35 Eukaryotic sequences were removed by clustering. The only Lokiarchaeotal sequence that came from a sequenced genome in the dataset was removed by clustering, but was added back in to the centroid dataset because it was central to
the research question. Again, gap characters were removed and the sequences were aligned in MUSCLE.

Columns were removed from the alignment if they contained 10% gap characters or greater. This was done by passing the alignment to the trimAL program v1.4.rev15 with the option “-gt 0.9”. This reduced the number of columns in the multiple sequence alignment from 1,976 to 571. This cleaned alignment was then used for phylogenetic reconstruction (Capella-Gutiérrez et al. 2009). This action is justifiable due to the events such as intein invasion can occur convergently, and thus should not be considered relevant to history of the host protein in question, but rather separately. Additionally, the regions around gap characters often tend to be unreliably aligned.

**Compositional Heterogeneity in Sequences:**

The Tree-Puzzle program creates a maximum likelihood phylogeny for a multiple sequence alignment, but prior to doing so runs a 5% chi-square test for compositional heterogeneity in the dataset. This analysis checks whether the base composition of each sequence is identical to the average base composition across the multiple sequence alignment (Schmidt et al. 2002).

**Phylogenetic Reconstruction:**

The alignment was analyzed using ProtTest 3.4.2 to determine the evolutionary model for phylogenetic reconstruction in RAxML (Darriba et al. 2011; Stamatakis 2014). All possible models available in ProtTest were considered with and without invariable sites and a gamma parameter for among site rate variation. The starting tree was left at the default BioNJ option
and the search strategy was set to Maximum Likelihood. The LG model with invariable sites and
a gamma distribution for among site rate variation was returned as the most supported model
for all four testing criteria, AIC, BIC, AICC, and DT.

The multiple sequence alignment was passed to RAxML v.8.1.17 for maximum likelihood
phylogenetic reconstruction. The LG model was used with an estimation of invariable sites and
4 rate categories of among site rate variation modeled with a gamma distribution (\texttt{-m
PROTGAMMAILG}). Support values were estimated using RAxML’s rapid-bootstrapping
algorithm with 100 replicates (\texttt{-f a -# 100}). The trimmed multiple sequence alignment was also
recoded separately into 6 (ASTGP, DNEQ, RKH, MVILFYW, C) and 4 (ASTGP, DNEQ, RKH,
MVILFYW) Dayhoff classes. These multiple sequence alignments were each passed to RAxML
for phylogenetic reconstruction. To run recoded data in RAxML one has to name the first class
0, and number every class after that in ascending order (recode\_dayhoff4.py and
recode\_dayhoff6.py, GitHub repository). Both were run using the multi-state model under a
general time reversible framework with gamma parameter estimation for 4 categories of
among site rate variation. The multi-state model is the only way to handle character data in
RAxML that is neither the standard 4-class nucleotide classes nor the 20-state amino acid
classes. The GTR model gets the maximum likelihood estimation of the transition ratio between
dayhoff classes from the multiple sequence alignment input. Thus, substitutions between
classes do not occur with equal probability.
Results:

To ascertain the evolutionary position of several novel lineages, a 366 taxa catalytic A-ATPase phylogeny was inferred in RAxML v.8.1.17 (Stamatakis 2014). Dayhoff recoding was used on the alignments to reduce compositional bias in the data (Susko & Roger, 2008; RAxML Google Group). These datasets were also analyzed in RAxML and the resulting phylogenies were compared with the full original trimmed dataset. To see if Dayhoff recoding was adequately reducing the compositional bias in the alignment, the full alignment, trimmed alignment, 6-class recoding, and 4-class recoding were analyzed using a Chi-Square Test for compositional heterogeneity.

USEARCH Centroid Clustering

A MUSCLE alignment of the catalytic ATPase subunit for various archaeal and eukaryotic species was processed by the USEARCH program for clustering (Edgar, 2004; Edgar & Bateman, 2010). The ‘cluster_fast’ algorithm can return centroid sequences for an aligned dataset in fasta format. The dataset was clustered at the 95% identity threshold, which served to reduce the total number of sequences in the phylogeny from 566 to 366. Overall sequence diversity of the dataset was maintained because a representative sequence for each cluster, or centroid sequence, was returned by the analysis.

Gapped Column Alignment Trimming

After re-aligning the returned centroid sequences using MUSCLE, the centroid alignment consisted of 1976 alignment columns (figure 2.1A)(Edgar, 2004). This alignment was passed to the TrimAL program to remove any alignment columns that contained more than 10% gap
characters (Capella-Gutiérrez et al. 2009). This reduced the number of positions in the alignment to 571 (figure 2.1B). Since trimming has the tendency to remove insertions and inteins that only occur in a minority of the sequences, it also has the added benefit of reducing the compositional bias for sequences in the dataset (table 2.1). This is likely because genetic elements like inteins can have a different amino acid composition than their host proteins, which comes from the fact that mobile genetic elements can have a different evolutionary history than the sequences in which they reside.

Chi-Square Test for Compositional Heterogeneity

Four multiple sequence alignments were passed to the TreePuzzle program to determine whether individual sequences in the alignment have a composition that is significantly different from the average amino acid composition of the entire alignment (Schmidt et al. 2002). This is implemented as one of the first analyses of the TreePuzzle algorithm and is reported to the user in the main output file. Each sequence ID is printed on a separate line, also printed is whether they passed or failed a 5% Chi-Square Test for compositional heterogeneity.

The different alignments analyzed included the three that were passed to RAxML for phylogenetic analysis, as well as the version of the full alignment prior to gapped-column trimming implemented by the TrimAL program as described in the methods section (Capella-Gutiérrez et al. 2009). In the untrimmed alignment, some sequences had inteins in the ATP
Figure 2.1: Alignment snapshots displaying the removal of gapped alignment columns by the TrimAL program. The two images show how the TrimAL program effectively removes gapped columns from an alignment. (A) This is the alignment of the centroid dataset output by the UCLUST algorithm. Large regions comprised mostly of gap characters can be observed at positions where a handful of sequences contain inteins or insertions. (B) The trimmed alignment is much condensed compared to the centroid alignment and the sequences have less compositional heterogeneity (table 2.1).
synthase, and also insertions or deletions at various positions. Since only a minority of the sequences in the dataset had these genetic elements, this could lead to them being rejected by a 5% test for compositional heterogeneity. Indeed, in table 2.1 it can be seen that when the columns with greater than 10% gap characters are removed from the full alignment there is a stark reduction in the number of sequences that do not pass the Chi-Square Test.

Dayhoff recoding was used in order to reduce compositional bias that was identified in the 20 amino acid state full and trimmed alignments. By assigning the amino acid characters to classes based roughly on their biochemical characteristics one effectively reduces the distance between sequences. It can be seen in table 2.1 below that Dayhoff recoding successfully eliminates compositional bias in the dataset as not a single sequence fails the Chi-Square Test for compositional heterogeneity at the 5% threshold in either the 6-class or 4-class recoded alignments.

**Table 2.1**: Results for a 5% Chi-Square Test for compositional heterogeneity of sequences within a multiple sequence alignment dataset. For each sequence in the alignment, the average amino acid base composition is calculated to the average amino acid base composition for the entire alignment. If the sequence does not match the average composition at the 5% level in a Chi-Square Test, then it is reported as ‘failed’ to the user. Sequences that fail the test represent sequences that suffer from compositional bias. In the table below the results of this test are reported for 4 alignments as both raw counts and percentages. The lower three were used in subsequent phylogenetic analysis; they are the alignment produced by the TrimAL program with 20 amino acid states, the 6-class Dayhoff recoded alignment, and the 4-class Dayhoff recoded alignment.

<table>
<thead>
<tr>
<th>Alignment</th>
<th>Failed</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>71</td>
<td>19%</td>
</tr>
<tr>
<td>Trimmed</td>
<td>42</td>
<td>12%</td>
</tr>
<tr>
<td>6 - state</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>4 - state</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>
Maximum Likelihood Phylogenetic Analysis

After testing for compositional heterogeneity the full alignment was passed to ProtTest v3.3.2 to determine which evolutionary model should be used in the maximum likelihood analysis (Darriba, Taboada, Doallo, & Posada, 2011; Guindon & Gascuel, 2003). ProtTest takes a multiple sequence alignment as input in Phylip format and determines which model out of those selected fits the data best using information criteria metrics based on PhyML analysis. In this run all models that can be tested in ProtTest were considered as well as the choice of using a gamma parameter for among site rate variation and an estimate for proportion of invariant sites. The result of the analysis was that LG+I+G should be the model used to reconstruct the phylogeny of this dataset. RAxML was called with the following model ‘PROTGAMMAILG’ and 100 rapid bootstrapping replicates to determine support for branches in the resulting unrooted phylogeny (figure 2.2).

The 6-state and 4-state Dayhoff recoded alignments were not passed to ProtTest to determine the best fit evolutionary model. This was because in order to run Dayhoff recoding phylogenetic analysis in RAxML, the multi state model must be used. This model takes as input a multiple sequence alignment that has numbers for characters rather than amino acid or nucleotide code. This model can be run with or without a gamma parameter for among site rate variation and estimated proportion of invariant sites. Additionally, an option can be set to run this under a general time reversible (GTR) model of sequence evolution. The 6-state and 4-state alignments represented as numbered alignments (0-5 and 0-3 respectively) were thus passed to RAxML with the GTR ‘MULTIGAMMAI’ model and 100 rapid bootstrapping replicates for support (figure 2.3 and 2.4 respectively).
Characteristics of the Phylogenies

There are some interesting observations that can be drawn by looking at the phylogeny for the trimmed ATP-synthase catalytic subunit dataset in figure 2.2. This reconstruction did not support previous findings in the literature showing the Eukaryotes within either the ASGARD group or the Lokiarchaeota clade (Spang et al.; Saw et al. 2015; Zaremba-Niedzwiedzka et al. 2017b). This topology can be rejected with high confidence in the full sequence ATP synthase phylogeny because there is a high bootstrap support value (above 90) at the base of the Eukaryotic clade and at the base of the Lokiarchaeota clade. The phylogeny does not reconstruct the ASGARD archaea as a single group, but rather places the four groups in various places on the tree. The Lokiarchaeota group with the Korarchaeota in the TACK superphylum, but the other three members of ASGARD come out as sister clades to Euryarchaeal groups. There are no high confidence bootstrap support values on the internal branches separating the four members of the ASGARD archaea so we can’t say for certain that the ATP synthase phylogeny rejects this topology.

Another interesting result of this phylogenetic analysis was that the archaeal taxa comprising the DPANN superphylum did not form a single cohesive group in the tree. This contradicts previous phylogenetic analyses using various different genes for reconstruction (Rinke et al. 2013; Cindy J Castelle et al. 2015; Hug et al. 2016). Three groups, the Diapherotrites, the Micrarchaeota, and the Parvarchaeota do fall close to one another midway within a clade that is largely Euryarchaeota. However, there is not strong support for this proximity. The Nanoarchaeota group far away from these taxa, and instead are sister group to
Figure 2.2: Maximum likelihood phylogeny for 366 Eukaryotic and Archaeal A-ATPase catalytic subunit protein sequences. This phylogeny was inferred using the RAxML program v8.1.17 using the PROTGAMMAILG model. The phylogeny was inferred using the rapid bootstrapping algorithm option with 100 replicates. Green circles over a branch represent a bootstrap support value of 90 or greater. Clades were collapsed in FigTree for clarity and subsequently labeled based on resident taxa. The Eukaryotic sequence clade has been highlighted blue, the Haloarchaea in magenta, ASGARD archaea groups (Lokiarchaeota, Thorarchaeota, Odinarchaeota, and Heimdallarchaeota) in red, and DPANN group names in green. The phylogeny has been rendered in FigTree to separate the TACK superphylum, Lokiarchaeota, and Nanoarchaeota to the upper region of the tree and the Euryarchaeota, some DPANN groups, Eukaryotes, and some of the ASGARD groups in the lower region. This tree is unrooted and should be interpreted as so. Any clade that has been marked with a shovel icon has some sequences that are not officially assigned to a species because they were collected from the NCBI environmental non-redundant database.
the Crenarchaeotes within the TACK superphylum. The Woesearchaeota are sister group to a clade of Methanomicrobiales. The most convincing evidence for DPANN not being a cohesive group was the fact that the Haloarchaea and Nanohaloarchaeota were sister groups on the tree with a high bootstrap support value at the base of this clade. While several analyses place the Nanohaloarchaeota within DPANN, this is not the first reconstruction to place them as siter group to the Haloarchaea (Narasingarao et al. 2011; Lasek-Nesselquist and Gogarten 2013; Zhaxybayeva et al. 2013; Petitjean et al. 2015).

The phylogenetic reconstruction of the Dayhoff-recoded datasets show similarity to the full phylogeny in many regards. Neither of these two trees group all of the ASGARD taxa into a single cohesive group (see figures 2.3 and 2.4). However, there are still no strongly supported internal branches separating these groups, so the topology can’t be completely ruled out as a possibility. The Dayhoff-6 phylogeny places the Eukaryotes as a sister group to the Heimdallarchaeota, albeit with a very low bootstrap support value. The Dayhoff-4 phylogeny places the Eukaryotes as the sister group to the Lokiarchaeota, but again there is a low support value at the base of this bifurcation. We can again effectively rule out the scenario where the Eukaryotes arise from within the Lokiarchaeota in these two phylogenies because there are high bootstrap support values at the base of the Eukaryotic and Lokiarchaeotal clades in both recoded phylogenies. Both recoded phylogenies show a disruption of the DPANN superphylum similar to the phylogeny for the 20-state dataset. Nanoarchaeota again went to the TACK superphylum, and Woesearchaeota again group within the Methanomicrobiales, albeit with low bootstrap
Figure 2.3: the 6-state Dayhoff recoding maximum likelihood phylogenetic analysis for 366 A-ATPase protein sequences. This phylogeny was inferred from the 6-character state recoded alignment using the RAxML program v8.1.17 using the MULTIGAMMAI setting and general time reversible evolutionary model. Support values were generated using the rapid bootstrapping algorithm option with 100 replicates. Green circles over a branch represent a bootstrap support value of 90 or greater. Clades were collapsed in FigTree for clarity and subsequently labeled based on resident taxa. The Eukaryotic sequence clade has been highlighted blue, the Haloarchaea in magenta, ASGARD archaea groups (Lokiarchaeota, Thorarchaeota, Odinarchaeota, and Heimdallarchaeota) in red, and DPANN group names in green. The phylogeny has been rendered in FigTree to separate the TACK superphylum, Lokiarchaeota, and Nanoarchaeota to the upper region of the tree and the Euryarchaeota, other DPANN taxa, Eukaryotes, and other ASGARD groups in the lower region. This tree is unrooted and should be interpreted as so. Any clade that has been marked with a shovel icon has some sequences that are not officially assigned to a species because they were collected from the NCBI environmental non-redundant database.
Figure 2.4: the 4-state Dayhoff recoding maximum likelihood phylogenetic analysis for 366 A-ATPase protein sequences. This phylogeny was inferred from the 4-character state recoded alignment using the RAxML program v8.1.17 using the MULTIGAMMAI setting and general time reversible evolutionary model. Support values were generated using the rapid bootstrapping algorithm option with 100 replicates. Green circles over a branch represent a bootstrap support value of 90 or greater. Clades were collapsed in FigTree for clarity and subsequently labeled based on resident taxa. The Eukaryotic sequence clade has been highlighted blue, the Haloarchaeia in magenta, ASGARD archaea groups (Lokiarchaeota, Thorarchaeota, Odinarchaeota, and Heimdallarchaeota) in red, and DPANN group names in green. The phylogeny has been rendered in FigTree to separate the TACK superphylum, Lokiarchaeota, Eukaryotes, and Nanoarchaeota to the upper region of the tree and the Euryarchaeota, other DPANN taxa, and other ASGARD groups in the lower region. This tree is unrooted and should be interpreted as so. Any clade that has been marked with a shovel icon has some sequences that are not officially assigned to a species because they were collected from the NCBI environmental non-redundant database.
support. The Micrarchaeota, Parvarchaeota, and Diapherotrites again go together, but they fall within the Euryarchaeotal clade instead of at the base of the tree. Again, the only highly supported disruption of the DPANN superphylum was found at the base of the Haloarchaea and Nanohaloarchaeota group. Both Dayhoff recoding phylogenies also placed these groups as sister clades on the tree.

*Discussion and Future Directions*

The results of the phylogenetic reconstructions tell conflicting stories about the origin of Eukaryotes. The full 20-state and recoded 6-state phylogenies show them arising from within the Euryarchaeota clade, while the recoded 4-state phylogeny nests the Eukaryotes within the TACK superphylum. The phylogenies also don’t seem to agree on whether or not the Eukaryotes should be placed as a sister group to the ASGARD archaea. However, all three phylogenies show that there is no phylogenetic signal for the Eukaryota grouping within the Lokiarchaeota. Thus, the ATP synthase catalytic subunit phylogeny does not support the results found in other published phylogenetic analyses for these groups (Spang et al. 2015; Saw et al. 2015; Zaremba-Niedzwiedzka et al. 2017a).

The three phylogenies seem to be in agreement supporting the disruption of the DPANN superphylum. The most convincing evidence for this would be that the Haloarchaea and the Nanoahaloarchaeota group together with high bootstrap support in all three phylogenies. As these two groups hail from the same hypersaline environment, we would expect compositional bias to tend to group them together artificially in the 20-state full phylogeny. However, even in the dayhoff recoded phylogenies which reduce compositional bias the same relationship is
reconstructed with high support. Another line of support for a disrupted DPANN is the fact that
the Woesearchaeota and the Nanoarchaeota leave behind the Micrarchaeota, Parvarchaeota,
and Diapherotrites, but migrate to the same place on the tree in all three phylogenies.

In future work, the maximum likelihood analyses described here will be used to
compare the best tree to a tree that has been constrained to reflect different evolutionary
hypotheses. Three different constrained topologies will be constructed to test these
hypotheses. In one all ASGARD archaea will be grouped together into a single, distinct clan. A
second will have all DPANN archaea together in a single distinct clan. Finally, a topology will be
tested that places all Eukaryotes in a single distinct clan that falls within the Lokiarchaeota.

For each of these constraints the best maximum likelihood phylogeny will be
determined in RAxML, using the alignments and evolutionary models described above. These
trees will be compared to the overall best tree (calculated without constraint) using the tests
implemented in CONSEL (Shimodaira and Hasegawa 2001).
Appendix 1:

Supplementary Table 1:

### Abbreviations for Major groups

<table>
<thead>
<tr>
<th>Archaea</th>
<th>Bacteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acidobacteria</td>
<td>Acidobacteria</td>
</tr>
<tr>
<td>Archaeoglobi</td>
<td>Acidotrichales</td>
</tr>
<tr>
<td>Desulfurococcus</td>
<td>Desulfoflavobacteria</td>
</tr>
<tr>
<td>Rhodobacteroides</td>
<td>Rhodobacterales</td>
</tr>
<tr>
<td>Halobacteriaceae</td>
<td>Halobacteriales</td>
</tr>
<tr>
<td>Methanoperditales</td>
<td>Methanopenicillales</td>
</tr>
<tr>
<td>Methanococcales</td>
<td>Methanosarcinaceae</td>
</tr>
<tr>
<td>Methanoregulales</td>
<td>Methanopyrales</td>
</tr>
<tr>
<td>Methanomicrobales</td>
<td>Methanomicrobiota</td>
</tr>
<tr>
<td>Methanosarcinales</td>
<td>Methanospirillales</td>
</tr>
<tr>
<td>Nanarchaeota</td>
<td>Methylophilales</td>
</tr>
<tr>
<td>Nitrososphaerales</td>
<td>Methylococcus</td>
</tr>
<tr>
<td>Thermoacetococci</td>
<td>Methylophilales</td>
</tr>
<tr>
<td>Thermoproteococci</td>
<td>Methylophilales</td>
</tr>
<tr>
<td>Thermoplasmales</td>
<td>Methylophilales</td>
</tr>
</tbody>
</table>

The table above lists all of the primarily order level taxonomic classification used to rename the sequences obtained as phylip alignments from Bill Martin (University of Hanover, Germany). These were used in the PERL script that can be found in Appendix 2. After calling e-utils to obtain the taxonomy for a particular sequence, that sequence would be renamed incorporating the abbreviation for the group.
The histogram above includes branch length ratio values obtained from the 355 trees made from alignments obtained from Bill Martins, which were associated with the Weiss et al., 2016 publication. The process of isolating the ratios for each tree includes dividing the maximum within group distance by the interdomain branch length distance as outlined in equation 1. The pseudocode for the script used to measure these distances and calculate the ratios was explained in detail in the Results section. As can be seen in the plot above, most of the density of the data is centered between 0 and 100. There is also an outlier with a ratio of 724. In subsequent histograms, the range was reduced in order to better observe the data in the with a more precise scale. This was important to the analysis because if these proteins were present in LUCA, we would expect the majority of them to have a ratio that is less than 1. This conjecture was not upheld by the data, which lends support to the argument that the Weiss et al. 2016 analysis did not correctly identify gene families that were present in LUCA in the majority of cases. That or they neglected to capture the extent of sequence diversity contained in each of these gene families, leading to an artificially short interdomain branch length.
The histogram displayed above was made using the same branch length ratio dataset used to make supplementary figure 1 on the previous page, and figure 2 in the text. In this plot the range has been limited to 150 to exclude the outlier. This served to expand the graph in the region where the majority of the data lay. From this graph it can clearly be seen that the leftmost value corresponding to the lowest ratio range (0-1) has a much lower frequency of occurrence than the ratio values greater than 1. Figure 2 which is shown in the main body of the text limits the range to 50 in order to more clearly see the frequency of the values between 0 and 1 and those greater than 1.
Works Cited:


Becerra A, Delaye L, Isla S, Lazcano A. 2007. The Very Early Stages of Biological Evolution and


cNVlY7mnPGbx8LCb-yKV8Bk5exe5BOtPnP~Eb6atRq53IAg-

JjZvTd0MwRfvAVZslpNypR2K~v0szJlOrkO4AGbxw~4H2CChOfxecAu-

ERCTB30lnwKA7Jd0a~TJtIC6ptDlyWA8A-

jclQviQLvkUWnqgPKmvZSaFHI2HNUNAIlyQVOjuBAv~lQdwJoHwPZ5ZfZ7zBeRMsBFvdj7Wh

4m08dMoz1Z~oh6eQEYFk4QmyOKad0GNcBomKZIkS7vPFuhxEAb2xm3fw6l6pe4r9znGVt

WUOC7RZ1~h3iviLRO9w__&Key-Pair-Id=APKA1UCZBIA4LVPAVVW3Q


Herbeck JT, Degnan PH, Wernegreen JJ. Nonhomogeneous Model of Sequence Evolution Indicates Independent Origins of Primary Endosymbionts Within the Enterobacteriales (c-Proteobacteria). Available from: https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/mbi/22/3/10.1093/molbev/msi036/2/msi036.pdf?Expires=1492829604&Signature=Og99JTN0Fj8DhCtz6OC9Jlv4PXQdjodCYnk-EiZDS22UIAMNc-aPmLxR2NTtfZSbL-wpK6DYvJsIoX88TVrSCIA3e7~In9C3Xv3jRCFUdHrj3LQjj6S~BxdLTU3uhqcM5ItSHNxCPDNq7d11K9M7Vweftezr~ZrKNeEnqLFjDTZzC9h90WJRZ52PhT5LrVrodhK3cmvRLz6pYNdEjpl6JZWDuEuTUhZ5rchJmkeU3Be~EOLn8H2PDNyPyf9y-oUEvakJarztCn4~i9R3O~bUy4MMa7TEPy7gxFfJ8rk5THgPZxA2SO4d8SiR5Qz05aQKVVFmg64ajgtWpQYv__&Key-Pair-Id=APKAiUCZBIA4LVPAVW3Q


http://www.biomedcentral.com/1471-2148/3/2


http://www.pnas.org/content/86/12/4569.full.pdf


https://www.nature.com/nature/journal/v499/n7459/pdf/nature12352.pdf

Rota-Stabelli O, Lartillot N, Philippe H, Pisani D. 2013. Serine Codon-Usage Bias in Deep Phylogenomics: Pancrustacean Relationships as a Case Study. Syst. Biol [Internet] 121. Available from: https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/sysbio/62/1/10.1093/sysbio/sys077/2/sys077.pdf?Expires=1492829819&Signature=Y5gafKUbE7B6TK4Say8nOXZq9bLn275sXV-W1A82Uqm~SLCfa51z9h~hg3y0rBL1D59ndFe2N6p8hIIAewhPEhKwBzVSmXVLUV-cbrliu0EvASN9I4O96SQKrk1n89bJiWm75~tPp6E66zh-qmcye2lieKXs-FfOXXkjWy55za7BEsi5mQEEim4FbbN5bx5xq7qlseXGb77VRIJffdeKk6Zb2LN9EwBXbnGf1rfRx62Akkq8Byvv4yHo6Hs335ewQA062t6PHlDXsQ1Jkfk6JCw1CULrKyXZr8ZM8UGB7x25sAsDji z5imJZwiUnps~V4EVIWtFEV6U7VwKcH7BA__&Key-Pair-Id=APKAIUCZBIA4LVPAVW3Q

Saw JH, Spang A, Zaremba-Niedzwiedzka K, Juzokaite L, Dodsworth JA, Murugapiran SK, Colman

http://rstb.royalsocietypublishing.org/content/370/1678/20140328.long


https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/bioinformatics/17/12/10.1093/bioinformatics/17.12.1246/2/171246.pdf?Expires=1492828326&Signature=dqcc7EnG4Si7rSKmMZ8i1Oaj9gdanstbMR~HVxQmQphI9bcT~0D4kZrXpDSlcmCz5d6eHCB4KalZL9QsTRzD~c8cPBoqlfz6JRfQ4CQLahJhqoxGucV-cDtwnm8u3ufAWEWiibGQt49Daa1uKv6O3UA4FFE02ygqt1NDusE~tG5hnA2ALiRDMWg3g48F2w2x77MoSwIr42mKf1DfQk0Pqhjn9QBuddaQjkCLtxY8gG6KtfyUxlm36i8EKIdqmYyy7sUN4szGVMDEtoRxTPy1I8An5KvgFw7DOVmoWkZTbC6EK1cy4qnw3hz2hyRC1Fll6lqgECK40ZpAlql7yq7hwWA__&Key-Pair-Id=APKAIUCZBIA4LVPAVW3Q


Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW,

z&token2=exp=1493917973~acl=%2Fstatic%2Fpdf%2F116%2Fart%25253A10.1007%25252Fs00792-013-0514-z.pdf%3ForiginUrl%3Dhttp%253A%252F%252Flink.springer.com%2Farticle%2F10.1007%25252Fs00792-013-0514-z%25257%253Fsrc%25253Dlink.springer.com%2Farticle%2F10.1007%25252Fs00792-013-0514-
z%2526token2%3Dexp%253D1493917973~hmac%253D9cd21b50ab011d4738ce2a6ac5849dbce5aa43fe490ee72af18ec9999e02eccc