BEING *Aquifex aeolicus*: UNTANGLING A HYPERTHERMOPHILE'S CHECKERED PAST

by

Robert J.M. Eveleigh

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
December 2011

DALHOUSIE UNIVERSITY

DEPARTMENT OF COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "BEING *Aquifex aeolicus*: UNTANGLING A HYPERTHERMOPHILE'S CHECKERED PAST" by Robert J.M. Eveleigh in partial fulfillment of the requirements for the degree of Master of Science.

Dated: December 13, 2011

Co-Supervisors: _____

_____

Readers: _____

DALHOUSIE UNIVERSITY

DATE: December 13, 2011

AUTHOR: Robert J.M. Eveleigh

TITLE: BEING *Aquifex aeolicus*: UNTANGLING A HYPERTHERMOPHILE'S CHECKERED PAST

DEPARTMENT OR SCHOOL: Department of Computational Biology and Bioinformatics

DEGREE: MSc CONVOCATION: May YEAR: 2012

_____

Signature of Author

TABLE OF CONTENTS

List of Tables

List of Figures

Abstract

Lateral gene transfer (LGT) is an important factor contributing to the evolution of prokaryotic genomes. The Aquificae are a hyperthermophilic bacterial group whose genes show affiliations to many other lineages, including the hyperthermophilic Thermotogae, the Proteobacteria, and the Archaea. Previous phylogenomic analyses based on the concatenation of genes thought to be recalcitrant to LGT suggest that the Aquificae are sister to Thermotogae, but many phylogenies and certain cellular traits have suggested a stronger affiliation with the ε-Proteobacteria. Indeed, different scenarios for the evolution of the Aquificae yield different phylogenetic predictions. Here I outline these scenarios and consider the fit of the available data, including two recently sequenced genomes from members of the Aquificae, to different sets of predictions. Evidence from phylogenetic profiles and trees suggests that the ε-Proteobacteria have the strongest affinities with the three Aquificae analyzed. However, this phylogenetic signal is by no means the dominant one, with the Archaea, many lineages of thermophilic bacteria, and members of genus *Clostridium* and class δ-Proteobacteria also showing strong connections to the Aquificae. The phylogenetic affiliations of different functional subsystems showed strong biases: as observed previously, most but not all genes implicated in the core translational apparatus tended to group Aquificae with Thermotogae, while a wide range of metabolic systems strongly supported the Aquificae - ε-Proteobacteria link. Given the breadth of support for this latter relationship, a scenario of ε-proteobacterial ancestry coupled with frequent exchange among thermophilic lineages is a plausible explanation for the emergence of the Aquificae.

List of Abbreviations Used

BLAST ..... basic local alignment search tool
COG ..... clusters of orthologous groups
DP ..... dynamic programming
FSA ..... fast statistical alignment
GO ..... gene ontology
HMM ..... hidden Markov model
LGT ..... lateral gene transfer
MCMC ..... Markov chain Monte Carlo
ME ..... minimum evolution
ML ..... maximum likelihood
MP ..... maximum parsimony
MRP ..... matrix representation using parsimony
MSA ..... multiple sequence alignment
NJ ..... neighbor joining
OTU ..... operational taxonomical unit
RAxML ..... randomized axelerated maximum likelihood
SSU rRNA ..... small subunit ribosomal RNA
TOL ..... Tree of Life
UPGMA ..... unweighted pair group method with arithmetic mean
VPI ..... variable preference index

Acknowledgements

Firstly, I would like to thank my father, Eldon Eveleigh, for teaching me about the natural world very early on in life. It was these lessons that initiated my curiosity for the world around me and eventually lead to my love for science. Furthermore, the guidance and unwavering support provided by my family over the years especially the last few years has both been essential and an enormous help to me. I cannot thank them enough. I would also like to thank the newest member of my family, Amanda Morris, her passion for life and firey nature has greatly counter-balanced the stresses of this endeavor. Going through all of this has been much easier knowing she was in my corner at every step.

Secondly, I would like to thank all of my colleagues in the Archibald, Doolittle, Roger and Beiko labs. Without their imparted wisdom and experiences my journey to complete this project would have been an arduous one. So thank you. I would like to single out one person in particular, Dennis Wong, for constantly reminding me that there was more to life than just work and schooling. He showed me that time away from work is just as important as working.

Lastly, I want to thank my supervisors, Robert Beiko and John Archibald, for their brilliance and knowledge in guiding me subtly down this path of learning and discovery. But more importantly, I want to thank them for taking a chance on me; for recognizing and fostering my potential. I have learnt so much and I am grateful for the opportunity they have given me.

Chapter 1

Introduction

1.1 Introduction to Phylogenomics

Understanding the natural historical processes that gave rise to various organisms is central to almost any evolutionary study. The idea that evolution is a branching process, whereby populations are altered over time and may speciate into separate branches, follows from the evolutionary theory presented by Charles Darwin in *The Origin of Species* (Darwin 1859). Today, phylogenetics – the reconstruction of evolutionary history – relies on mathematical methods to infer the past from features present in extant organisms. These reconstructions, represented as a tree or phylogeny, involve the identification of homologous characters shared among different organisms and the inference of a tree based on the comparisons of these characters. The accuracy of the inference, therefore, is greatly impacted by the quality of the evolutionary models of such characters. Thus, some of the biggest challenges in the reconstruction of the evolutionary history of life on Earth are methodological - improving phylogenetic algorithms to correctly model the underlying evolutionary mechanisms.

Prior to the 1970s and the advent of molecular techniques for sequencing proteins and DNA (deoxyribonucleic acid; Zuckerkandl and Pauling 1965b), phylogenetic reconstructions focused on the analysis of phenotypic characteristics (e.g., morphological or ultrastructural characters) to infer phylogenetic relationships. Phenetic analyses, popular during the mid-twentieth century, played a dominant role in comparative studies involving fossil data and placement of extant species (Sneath and Sokal 1973). However,

the limited number of reliable homologous characters, almost non-existent among microorganisms, hampered these approaches. The introduction of molecular data into phylogenetics, however, increased the number of homologous characters for comparisons and greatly improved the resolving power of phylogenetic inference. Through their congruence (i.e., the agreement between single-gene phylogenies and classical morphological and ultrastructural studies, and subsequent multi-gene phylogenies), genes could be used to reconstruct what is often called the tree of life (TOL). Moreover, molecular phylogenetics has enabled the identification of genetic, ultrastructural and metabolic features of ancient life forms that lacked fossil data (Zuckerkandl and Pauling 1965a). Now the challenge has shifted from identifying morphological characters to identifying conserved and ubiquitous phylogenetic marker genes appropriate for microbial classification.

One gene in particular, the highly conserved small subunit ribosomal RNA (SSU rRNA) gene, became an invaluable reference marker. This gene revolutionized microbial classification, systematics and ecology, and its use helped identify the Archaea as a third distinct domain of the TOL (Woese and Fox 1977). Moreover, the rRNA molecule is one of only a few gene products present in all cells and its sequencing has helped classify organisms, calculate related groups and estimate rates of species divergence. However, the rRNA gene does have its drawbacks. It has been well documented that similar nucleotide composition (e.g., high G + C content among thermophilic lineages) among evolutionarily distant rRNA genes can incorrectly place organisms together in phylogenetic trees (Woese et al., 1991; Hasegawa and Hashimoto, 1993). Furthermore, inferring the phylogenies of organisms based on any single gene may often be misleading

(Jeffroy et al., 2006) and must be corroborated by the use of other phylogenetic makers, particularly protein-coding genes (e.g., *EF-Tu*, *rpoB*, *recA*, etc.; Ludwig and Klenk, 2005), which are less prone to compositional biases than SSU rRNA (Baldauf et al., 2000; Hasegawa and Hashimoto, 1993; Loomis and Smith, 1990). Consequently, as more genes were sequenced and analysed, topological conflicts between phylogenies identified numerous poorly resolved parts among all domains of life due in part to poor taxon sampling.

This situation improved with the advent of genomic sequencing. Each completed genome sequence brings with it the sequences of all protein-coding genes found in that organism. Currently thousands of large-scale genome sequencing projects have been completed such as those maintained by National Center for Biotechnology Information (NCBI). This wealth of data gave birth to a new field of research, termed phylogenomics, which utilizes phylogenetic principles to interpret genomic data (Eisen and Fraser, 2003). One branch of phylogenomics involves the use of these data to reconstruct the evolutionary history of organisms. Access to genomic data could potentially alleviate previous sampling biases by greatly expanding the number of characters for phylogenetic analysis. Again, with this increase, the emphasis of phylogenetic inference shifted from the search for informative characters to the development of more-sophisticated reconstruction methods for use with genomic data.

In this section, I will describe the current methods involved in phylogenomic inference and briefly discuss their merits and pitfalls. I will not cover all methods exhaustively, but will focus on approaches that have been used to infer the phylogenetic placement of the Aquificae group, the bacterial lineage examined in this thesis. I will then

describe the problems in resolving the position of the Aquificae, focusing on the mechanism of lateral gene transfer (LGT) also known as horizontal gene transfer (HGT) which results in two plausible scenarios describing the descent of this group.

1.2 Overview of Phylogenomics

Phylogenomic approaches provided the opportunity to use characters that are based not only on sequence data but on genomic structures which are only possible with (or greatly improved upon by) the analysis of complete genomes (Eisen and Fraser 2003). Two essential steps of classical phylogenetic inference, the identification of homologous characters and tree reconstruction, are utilized in phylogenomic studies. Therefore, prior to describing the various techniques employed by phylogenomics, I will briefly describe the steps involved in phylogenetic inference. I will begin by discussing the identification and alignment of homologous sequences, removal of ambiguously aligned columns, tree reconstruction methods and measures of phylogenetic confidence.

1.2.1 Introduction to Phylogenetic Inference

Modern phylogenetic methods are based on the comparison of sequence data (i.e., nucleotides or amino acids), and the reconstruction of phylogenetic trees inferred from multiple sequence alignments (MSA). Since sequence alignment techniques are based upon a model of divergent evolution or a substitution model (Thompson et al., 1994), the input of multiple sequence alignments are assumed, *a priori*, to be a set of homologous sequences. Programs such as basic local alignment search tool (BLAST; Altschul et al., 1990) or FASTA (Lipman and Pearson, 1985) employ heuristics to perform rapid searches of a query sequence to recover putative homologs. Despite the development of

statistical models of homology, non-homologous sequences may still inadvertently be added to the alignment set. Indeed, the selection of gene sets which arose unambiguously by vertical decent (i.e., orthologous) is difficult and subjective. Selection of such gene sets is usually achieved by targeting single-copy genes or through automated clustering methods. Many of these identification processes, however, rely heavily on BLAST similarity scores, which are known to be poor estimators of the true evolutionary distance (Koski and Golding 2001).

Sequence alignments organize a set of putative homologues to reveal conserved and variable sites. Initially this was achieved using dynamic programming (DP; Needleman and Wunsch, 1970) algorithms to determine the minimum number of mutations that may have occurred during the evolution of two or more sequences (Carrillo and Lipman 1988). For proteins, this method usually involves two sets of parameters: a gap penalty and a substitution "log-odds" matrix. For this purpose several amino acid substitution matrices, such as PAM (percent accepted mutation; Dayhoff et al., 1978) and BLOSUM (blocks of amino acid substitution matrix; Henikoff and Henikoff, 1992) series, have been developed, which estimate the evolutionary likelihoods of mutations and conservations of amino acids. Thus, given a specific scoring scheme that defines the scores for residue matches, mismatches and gaps, the DP algorithm optimally aligns two sequences to maximize their similarity. Although DP methods are an effective way to align sequences, applying these techniques to more than two sequences quickly becomes computationally intractable (i.e., the number of comparisons increases exponentially with the number of sequences).

An important innovation in multiple sequence alignment was the introduction of progressive alignment (Feng and Doolittle, 1990). This heuristic approach dramatically reduces the number of pairwise comparisons by progressively aligning sequences based on the order specified by an approximated phylogenetic tree – a guide tree. In this way, phylogenetic information is incorporated to guide the alignment process, such that pairs of sequences or pre-aligned blocks of sequences become aligned from the most similar pair to the most distantly related to produce a final MSA (Pei 2008). This greedy strategy, however, is highly dependent on the quality of the initial guide tree, as an error incorporated at any stage in the growing MSA will propagate through to the final MSA.

To help alleviate the greediness of this strategy, recent methods implement an iterative alignment procedure (Edgar 2004; Gotoh 1996; Hogeweg and Hesper 1984) that revisits previously calculated pairwise comparisons or sub-MSA to enhance the alignment quality. This refinement process corrects the MSA by gleaning increased information from previous alignments, starting with the initial guide tree, until a predetermined number of iterations or a predefined convergence criterion is reached. Such methods generally produce higher quality alignments when evaluated against reference alignment sets such as BAliBASE (Thompson et al., 2005). Alignment methods such as MUSCLE (multiple sequence alignment by log-expectation; Edgar, 2004) and FSA (Fast Statistical Alignment; Bradley et al., 2009) are among the top benchmarked approaches capable of dealing with challenging test cases covering most of the protein fold space (Thompson et al., 2005). MUSCLE improves upon the progressive methods using a more accurate distance measure (the Kimura correction for multiple substitutions at a single site; Kimura 1991) to assess the relatedness of two sequences. A recent

probabilistic procedure, FSA (Bradley et al. 2009), incorporates hidden Markov models (HMMs; Eddy 2011) to obtain pairwise estimates of homology for each character and uses a sequence annealing technique to find the MSA which is closest to the truth given the statistical model.

Prior to the inference of an evolutionary tree, the quality of the alignment should be assessed. The quality of the alignment may have an enormous impact on the resulting phylogenetic tree (Ogden and Rosenberg, 2006; Phillips et al., 2000). This is particularly true when the set of putative homologues in a set are very divergent and/or vary in length, requiring the introduction of gaps into the alignment (Rosenberg 2005). MSA procedures convert sequences of unequal length into sequences of equal length by determining the optimal placement of gaps (representations of insertion and deletions or *indels*), with the goal to infer homology among characters or positional homology (Talavera and Castresana, 2007). However, indels are treated in a variety of ways during MSA and phylogenetic reconstruction. Generally, phylogenetic approaches treat gaps as missing data and contribute no additional phylogenetic information, unless indels are explicitly modeled as discrete events such as modeling rare genomic changes (RGCs; Rokas and Holland 2000). Thus researchers can deal with aligned columns with gaps by either removing them entirely from the analysis or by treating them in an *ad hoc* fashion such as removing a subset of gaps based on flanking region conversation and gap frequency criteria (e.g., with GBlocks; Castresana 2000). Recent alignment confidence methods such as ALISCORE (Misof and Misof 2009) and GUIDANCE (GUIDe tree-based AligNment ConfidencE; Penn et al. 2010) assess the quality of each column (and residue in GUIDANCE) and have been shown to be superior to GBlocks (Penn et al., 2010).

Other programs that can output an MSA with confidence scores are Hmmer (version 3; Eddy, 2011) and FSA (Bradley et al., 2009), which use a statistical model that allows calculation of the uncertainty in the alignment.

The primary aim of molecular phylogenetic inference is to infer the progression of divergences that produced a group of observed sequences. Given a high-quality MSA and an evolutionary model to permit estimation of the genetic distance between two homologous sequences, phylogenetic inference finds the tree topology and branch lengths that best describe the phylogenetic relationships among the sequences. Several methods exist for inferring evolutionary relatedness, most inference algorithms can be differentiated by the scoring function used, which can be distance-, parsimony- or likelihood-based.

Distance-based methods use pairwise estimates of evolutionary divergence to infer a tree, either by an algorithmic clustering approach (e.g., neighbor joining and UPGMA; Saitou and Nei, 1987) or by assessing the fit of the distances to particular tree topologies (e.g., least-squares or minimum-evolution (ME) criterion; Fitch and Margoliash, 1967, Cavalli-Sforza and Edwards, 1967). Trees reconstructed in this manner employ these algorithms incorporating a model of evolution (i.e., amino acid substitution) to compute a distance matrix that relates the number of differences between each pair of sequences. Distance methods, particularly the clustering approaches, are extremely fast estimators of phylogenetic trees, and a popular approach to produce an initial tree for more sophisticated approaches. A potentially serious weakness for distance methods such as neighbour-joining (NJ) is that these methods do not use character data directly, and the information found in the distribution of character states is lost in the pairwise comparison

upon reduction to a distance measure. Furthermore, reliable estimates of pairwise distances can be hard to obtain for divergent sequences. For instance, multiple substitutions at the same site and in some cases an incorrect model selection (Susko et al., 2004) may confound the true distance making sequences appear closer to one another artificially (Holder and Lewis 2003).

Unlike distance-based methods, parsimony and likelihood-based methods map the history of gene sequences onto a tree by searching for the most optimal tree based on the characters at each position in the MSA (Holder and Lewis 2003). Maximum parsimony (MP) seeks to find the tree that is compatible with the minimum number of substitutions among sequences, i.e., the fewest evolutionary changes. Most parsimony approaches (Steel and Penny 2000) lack a formalized model of evolution and claim to be model-free, which is seen as an advantage by some researchers (Siddall and Kluge 1997). Indeed, MP can approximate likelihood-based methods when the branch lengths on the tree are short (Steel and Penny 2000). The lack of an explicit model, however, has been suggested to significantly limit these approaches, requiring the MP approach to have strict assumptions of consistency across sites and among lineages. Thus, MP performance may be significantly hampered when substitution rates differ between regions (Yang 1996) or if evolutionary rates are highly variable among evolutionary lineages (DeBry 1992). However, parsimony and likelihood can select exactly the same tree in the case of a ‘no common mechanism’ likelihood model i.e., using a Jukes+Cantor type substitution process and every site is assigned its own vector of branch length parameters optimized by ML (Steel and Penny 2000).

Likelihood-based methods are among the most popular approaches and use likelihood-based scoring functions. Likelihood is a measure proportional to the probability of observing the data given the parameters specifying an evolutionary model and branch lengths in the tree. These parameters describe how sequences change over time and how much change has occurred on particular lineages, respectively. Likelihood-based methods for tree reconstruction come in two varieties: Bayesian inference, which samples trees in proportion to the likelihood and prior probability, and maximum likelihood (ML), which searches for the tree that maximizes the likelihood function. The differences between these approaches are the way in which Bayesian and ML approaches treat parameters in the models of sequence evolution as well as the quantity of interest it estimates i.e. the posterior probability of the parameter (tree) rather than the probability of the data given the model/tree/parameters (as in ML; Huelsenbeck et al., 2002).

Bayesian inference includes a prior probability distribution over all parameters, including the tree and evolutionary model, describing how frequently one would expect values to be observed before evaluating evidence from the data (Holder and Lewis 2003). In Bayesian phylogenetic inference, the posterior probability of a tree (i.e., the probability that the tree is correct) is a function of the product of the tree's prior probability and its likelihood. The most commonly used Bayesian methods use Markov chain Monte Carlo (MCMC; Hastings, 1970; Metropolis et al., 1953), a stochastic sampling technique designed to visit different tree topologies with a frequency proportional to each tree's posterior probability. From the distribution of visited trees, the posterior probabilities for individual bipartitions (since trees are generally unrooted) can be obtained, which can then be used to assemble a tree in which each clan has some minimum posterior

probability (often >0.5). The use of a prior distribution on trees can be viewed as a strength of the method, particular when prior information is known about the problem, or a weakness, when prior knowledge is lacking. Furthermore, it is unclear for how long an MCMC chain should be run. Typically, the number of samples required for MCMC to successfully sample the posterior distribution is dependent on two factors: convergence and mixing. A chain has converged when it begins to accurately sample from the posterior distribution after the burn-in period. The mixing of a chain controls how quickly a chain converges as well as its ability to sample effectively from the posterior distribution. Despite diagnostic tools to assess these factors (e.g., running and comparing multiple chains), it remains difficult to confirm whether a chain has converged and has successfully mixed.

Maximum likelihood (ML) methods are based on a specific probabilistic model of evolution and search for the tree with maximum likelihood under these models (Edwards and Cavalli-Sforza 1964; Felsenstein 1973, 1981). ML methods optimize all parameters in the likelihood equation (i.e., branch lengths and parameters in the substitution model) to find the highest peak in the parameter landscape, which corresponds to the tree with the maximum likelihood (Holder and Lewis 2003). Therefore, complex and parameter-rich models necessitate heuristic approaches for searching tree space to speed up computation time. Indeed, some researchers believe that ML can lead to over-fitting artifacts when applied to models that are too rich in parameters, i.e., the "infinitely many parameter trap" (Felsenstein 2004). However, this artifact is not only present among ML methods but is also present in any statistical modeling paradigm including distance and Bayesian approaches (e.g., Rannala 2002). ML methods propose a tree, initially produced

using a distance-based method, and refine the tree using a heuristic traversal scheme until no further improvement to the likelihood is found. For each iteration, the traversal scheme moves around tree space by proposing alternative trees from the current best tree estimate. Popular traversal schemes, such as nearest neighbor interchange (NNI), subtree prune-and-regraft (SPR), and tree bisection and reconnection (TBR), propose candidate trees by making small rearrangements to the current tree and evaluating the change in likelihood.

SPR is the most common topological move method that is less susceptible to local optima than NNI and less expensive in terms of computational time than TBR. SPR generates candidate trees by pruning subtrees and regrafting the broken branch at different positions in the remaining tree. The number of candidate trees generated by an SPR operation rapidly increases as the number of sequences increases and results in larger steps in tree space and fewer local optima (Felsenstein 1989; Swofford 2003). However, SPR moves are more expensive in terms of computation time, as more trees need to be evaluated and the likelihood for each potential move has to be evaluated over the entire tree. Recently, however, innovative SPR-based searches have reduced computational time by limiting the number of candidate trees by bounding the number of steps away that a subtree can move from its original position and pre-evaluating potential moves by some fast likelihood approximations. For instance, the Lazy Subtree Rearrangement (LSR) search heuristic implemented in RAxML (Randomized Axelerated Maximum Likelihood) by Stamatakis et al. (2005) prunes a subtree from the current best tree and subsequently reinserts the subtree into all neighbouring branches up to a certain ‘rearrangement’ distance from the pruning point. For each possible subtree insertion

within the rearrangement distance, RAxML computes an approximate log-likelihood score by optimizing the lengths of only three branches adjacent to the insertion point. Upon completion of the fast pre-scoring of all alternative trees, only a small fraction of the best scoring topologies are retained for a complete optimization of the overall tree score.

Confidence measures play a vital role in phylogenetics, these measures help identify trees or parts of a tree (i.e., clades/clans) that are well supported by the data and thus adequate to serve as the basis for evolutionary inference of biological systems (Huelsenbeck et al., 2000; Lutzoni et al., 2001). Bootstrapping is the most commonly employed confidence method among distance-based, parsimony and ML approaches and involves the generation of pseudo-replicate data sets (generally 100 or 1,000 replicates) by re-sampling with replacement the characters in the original MSA data set (Efron and Tibshirani 1993; Felsenstein 1985). When optimality-criterion methods are used (i.e., parsimony and ML), a tree search is constructed from each replicate and the resulting phylogenies are assessed against the best tree estimate of the original alignment. For distance-based methods such as NJ, bootstrapping involves the application of a tree-building algorithm to each pseudo-replicate and the generation of a consensus tree (e.g. 50% majority rule consensus tree). Once the optimal tree is chosen, bootstrap support for a group of interest is calculated as the proportion of times that the group is obtained in the pseudo-replicates. Thus, groups that receive a high bootstrap proportion (BP; generally greater than 70%) would be expected to be recovered by repeated analyses of new data sets drawn from the same underlying process (Felsenstein 1985). For this reason bootstrapping is often described as a measure of repeatability (Andrieu et al., 1997;

Felsenstein 1985) and gives an indication of tree reliability that is conditional on the data and the method (Holder and Lewis 2003). Therefore, a method that inaccurately models the properties of the data (see section 1.2.2 for details), may not only infer an incorrect tree but also contain nodes receiving strong bootstrap support (Swofford et al. 2001).

Bayesian methods generate the best estimate of a phylogeny by exploring the posterior probability of all trees sampled by the MCMC chain and selects the tree that contains the highest posterior probability (i.e., the MAximum Posterior probability or MAP tree; Rannala and Yang 1996). However, the sample of trees produced by MCMC is highly auto-correlated. As a result, millions of cycles through MCMC are usually required. Furthermore, like bootstrap values, if the model is misspecified an incorrect tree can be supported with strong posterior probabilities.

1.2.2 Phylogenetic Incongruence and Assumptions

The comparison of molecular phylogenies based on single genes can often lead to conflicting results. This incongruence can be the result of: (a) violations of the orthology assumption (e.g., inclusion of non-homologous sequences in MSA, hidden paralogy, lateral gene transfer, etc.), (b) stochastic error due to insufficient phylogenetic signal (i.e., genes of short length) and/or (c) systematic error leading to tree reconstruction artifacts. In this section, I will briefly describe orthology predictions and the impact of lateral gene transfer on orthology assumptions, as well as discuss the impact of systematic error on tree reconstruction.

Homology designates a relationship of common origin between attributes of living entities, without further specification of the evolutionary scenario that gave rise to

them (Fitch 1970). Two fundamentally different types of homologous genes are key concepts of evolutionary genomics: orthologs and paralogs. Orthologs are homologous genes separated by a speciation event, whereas paralogs are homologous genes related via duplication. Accurate inference of orthologous genes is a key component in most comparative genomic studies. However, the identification of genome-wide orthologous and paralogous gene sets, particularly for distantly related organisms, is a difficult task due to the complex nature of gene evolution. The combination of speciation and duplication events, along with LGT, lineage-specific gene loss, gene rearrangements, and other evolutionary mechanisms further entangle orthologs and paralogs into complex webs of relationships. Disentangling these complex relationships to infer orthologous genes remains an active field of research (Kristensen et al., 2011). These approaches differentiate orthologs, in-paralogs (genes resulting from a lineage-specific duplication subsequent to a given speciation event) and out-paralogs (genes resulting from duplication preceding a given speciation event) to varying degrees of success, but remain susceptible to genes acquired by lateral gene transfer – genes which seemingly mimic orthology.

In general, organisms can inherit genes in one of two ways: by vertical decent, the passage of genes from parent to offspring, or lateral (or horizontal) gene transfer (LGT/HGT), the exchange of genes between species. LGT is an important adaptive force in evolution (Gogarten et al., 2002; Ragan and Beiko 2009), common among prokaryotes and some eukaryotes, which potentially confers some selective advantage to the recipient of the transfer. Indeed, a cell's acquisition of genes allows for immediate and effective exploitation of novel ecological niches such as the classic example of antibiotic resistance

gene acquisition, allowing organisms to survive in an otherwise lethal environment (Hall et al. 1999). Transfer events were also identified in numerous metabolic processes such as the cytochrome *c* biogenic and export pathway (Kranz and Goldman 1998), as well as informational genes (e.g., ribosomal protein S14 (RpsN); Brochier et al., 2000), despite the suggestion that informational genes are seldomly transferred (Jain et al., 1999).

Prokaryotes utilize a number of specialized processes, such as conjugation, transduction and transformation which confer the uptake of exogenous material (Thomas and Nielsen 2005). Genetic transfer may be mediated either by a) cell-to-cell contact or bridge-like connection or conjugation, b) viruses through the mechanism of transduction or c) through the uptake of exogenous genetic material from the surroundings or transformation. Numerous vectors (e.g., plasmids, (retero-)transposons, etc.) have been implicated to mediate genetic transfer and they have collectively been referred to as the mobilome (Frost et al., 2005). Upon entering the new host cell, the foreign material may or may not replace existing homologous material (i.e., the substitution of a resident gene by an exogenous version through homologous recombination). In cases where the exogenous material is not replaced, the resident gene may eventually be lost (Doolittle et al., 2003). Thus, gene trees inferred from putative homologous gene sets impacted by LGT will potentially be topologically discordant from other gene trees accepted as the 'true' organismal relationships (i.e., phylogeny based on the16S rRNA gene). However, new models for the emergence of bacterial lineages that emphasize LGT as a process contributing to the generation of phylogenetically coherent bacterial groups, as opposed to eroding them, have been proposed (e.g., Gogarten et al., 2002, Andam et al. 2010). To complicate matters, between-species phylogenetic analyses indicate that even widely

distributed and functionally conserved genes can be subject to LGT. The culmination of these effects suggests that LGT is indeed rampant and pervasive among prokaryotic lineages and that existence of the 'true' prokaryotic TOL remains an open question.

Systematic error can often be traced back to some violation of model assumptions by the data analyzed. In probabilistic frameworks, such as maximum likelihood or Bayesian inference, most expectations of character evolution are formalized as parameters in an explicit model of character change (Lewis 1998) that are either built empirically from through comparison of observed sequences or parametrically using the chemical/biological properties of DNA and amino acids. These conceptual models make several assumptions that are not given formal parameters. Some of these assumptions are: (a) mutations are independent and identically distributed (i.e., i.i.d.), (b) that lineages arise in a divergent manner without reticulation i.e. tree-like evolution, (c) stationarity: substitution processes are consistent from species to species through time, (d) Markov process: a stochastic process with Markov property, or memorylessness property, is a process conditional on the present state of the system and not the previous states, and (e) reversibility: a Markov chain is said to be reversible if there is a probability distribution over states, π, such that $\pi_i \Pr(X_{n+1} = j \mid X_n = i) = \pi_j \Pr(X_{n+1} = i \mid X_n = j)$. However, these general assumptions are often violated in reality. The best-understood causes for these violations are derived from models that do not properly account for: variable evolutionary rates across lineages; heterogeneous nucleotide/amino acid compositions, resulting in artificial grouping of species sharing the same bias (Lockhart et al., 1994); and heterotachy, the shift of position-specific evolutionary rates (Kolaczkowski and Thornton 2004; Philippe and Germot 2000). These types of systematic biases have led tree

reconstruction methods to support, even with high confidence, an incorrect topology. To deal with these violations, new methods and models have been developed to reduce the impact of erroneous signal, in order to obtain accurate and robust trees (e.g., RY coding; Phillips et al., 2004 and other recoding strategies; Susko and Roger 2007 or covarion models to accommodate for heterotachy; Fitch and Markowitz 1970; Penny et al., 2001).

1.2.3 Phylogenomic Inference

The appeal of phylogenomics lies in the possibility of using a variety of kinds of molecular information beyond the gene sequence level. Trees have been inferred from whole-genome features, including statistical properties of complete genomes, gene content, or gene order, and can be used to corroborate results based on gene sequences. For brevity considerations, I will focus on and discuss methods that use data at the genomic scale to reconstruct the phylogeny of organisms as well as methods for detecting LGT events. I will begin by discussing phylogenomic approaches that rely on gene sequence data, i.e., sequence-based approaches, then focus on approaches that utilize whole-genome features.

Sequence-based phylogenomic methods use multi-gene datasets to alleviate stochastic or sampling errors inherent to single gene phylogenies (Rokas et al., 2003). Thus, the rationale behind these approaches is to provide both an increased number of characters (e.g., genes) and species to improve the accuracy of the inference (Graybeal 1998; Rannala et al., 1998). However, to maintain a large depth of coverage (i.e., taxon sampling), the selection of genes is generally restricted to universal or semi-universal genes, as well as those that are recalcitrant to LGT, and present in a single copy in most

or all genomes. Upon alignment of the chosen orthologous gene sets and the determination of unambiguously aligned positions, two approaches can be used to infer the phylogenetic tree: either the supermatrix or supertree approach.

The supermatrix approach, also known as 'simultaneous-analysis', 'combined-analysis' or 'total-evidence' approach, is defined by the direct, simultaneous use of all character evidence from all included taxa in hopes of revealing emergent support hidden in separate analyses of each data partition. These methods demand that little or no LGT among genes used since the signals from all genes are combined. In reality, however, the rate of LGT and its implications for supermatrix approaches, especially among prokaryotes, is much debated (Bapteste et al. 2005; Doolittle 1999; Gogarten et al., 2002). Indeed, the phylogenetic signal from the bulk of the genes often overwhelms the signal of a small number of transferred genes, but in the presence of high incongruence among gene trees other reconstructions methods outperform supermatrix methods (Kupczok et al., 2010).

Supermatrix approaches concatenate individual genes, encode non-overlapping taxa as missing data and infer a tree under maximum likelihood (ML) or Bayesian approaches. These popular methods consider across-gene heterogeneity in evolutionary rates by using partitioned models, which allow each gene to evolve under a different model (Yang 1996). Despite the increased number of extra parameters introduced by using an independent model for each gene, these partitioned models have shown to be more accurate than a model inferred from the entire concatenated alignment (Yang, 1996). However, gene features like sequence lengths and taxon overlap influence the accuracy of these methods. Adding more genes may increase the amount of characters

but it may also increase the number of incongruent trees, while adding more taxa typically influences the amount and distribution of missing data.

The supertree approach considers each data partition containing overlapping taxa sets (i.e., any data represented by a tree) individually, and the resulting topologies derived from these independent analyses are used to produce a single, joint estimate of phylogeny (Bininda-Emonds, 2004; Delsuc et al., 2005). Current supertree methods can be differentiated into those identifying common or well-supported groupings among the set of source trees (i.e., agreement supertrees), or those generating supertree(s) which maximize the fit among the set of source trees according to some optimization criterion (Bininda-Emonds et al., 2002). For direct supertree methods (Wilkinson et al., 2001) such as consensus supertrees, this is generally accomplished by comparing taxon bipartition sets– taxa that appear on one side of an internal branch as opposed to the other – between different trees to generate a composite tree (Steel, 1992; Bininda-Emonds, 1998). Unfortunately, most consensus reconstructions only accept or reject components on the basis of agreement among all source trees. Therefore, many consensus methods cannot resolve incongruence among source trees. In contrast, most optimization-based or indirect supertree approaches (Wilkinson et al., 2001) produce a composite tree in which incongruence among the source trees can generate novel clades/clans (Bininda-Emonds, 1998; Bininda-Emonds, 2004).

Matrix representations form the basis of most optimization supertree techniques. Although many supertree techniques have been proposed (Bininda-Emonds et al., 2002; Steel and Rodrigo, 2008), the most popular optimization-based technique is the matrix representation using parsimony (MRP) described independently by Baum, 1992 and

Ragan 1992. Both papers arrived at the same method: representing the hierarchical structure of the source trees as a series of 'matrix elements' using additive binary (two-state character) code (Farris et al., 1970). In rooted cases, each informative node of the source tree is coded in turn, where all taxa descended from that node are scored as 1 while all other taxa are scored as 0 to generate the matrix representation of each tree. In practice, MPR trees can also be constructed using unrooted trees. Next, in combining multiple trees, each source tree matrix representation are concatenated into a single matrix, where the coding is slightly modified such that nodes unrepresented in a given sources tree are scored as missing data (Bininda-Emonds et al., 2002). Optimization of the combined matrix representations under parimony (MRP; Baum, 1992 and Ragan 1992) produces the supertree. Thus, the elements created by matrix representation are indirect statements of membership, and are only functionally equivalent to conventional characters (i.e., nucleotides and amino acids). Although some believe this is a disadvantageous property of this approach, there are still analytical advantages. Notably, matrix elements can be individually weighted to account for differential support or confidence (i.e., bootstrap or PP support) in a source tree or the nodes therein. Such weightings have been shown to improve the fit between source trees and its matrix representation. Indeed, in the presence of gene-tree conflicts, weighted MRP as well as other variants can outperform supermatrix methods (Kupczok et al., 2010).

With the continued increase in fully sequenced and annotated genomes, additional phylogenomic approaches were developed for the comparison of whole-genome features, also known as genome composition analyses. These methods move away from nucleotide or amino acids as characters for tree building, opting rather for the use of genes (or even

segments of protein-coding genes) as characters for analysis. Thus, these techniques utilize a large quantity of data (the use of entire genomes), and different types of data in hopes of targeting the cellular history of the microbes rather than the history of a subset of genes (Dagan and Martin 2006). Unfortunately, microbial systematics and genomics have yet to be reconciled, due in part to the intrinsic difficulties in inferring reasonable phylogenies from genomic sequences, particularly in light of the significant amount of LGT in prokaryotic genomes (Beiko et al., 2008), e.g., between Archaea and hyperthermophilic bacteria (Nelson et al., 1999). Nonetheless, studies have been performed using whole-genome based methods and they are believed to provide reasonable estimates of known species trees, i.e., 16S rRNA phylogenies (Snel et al., 2005). However, other whole-genome features methods such as non-tree based or 'surrogate' methods (Ragan 2001) have been used to detect atypical nucleotide composition or unexpected phylogenetic profiles which may have exogenous origins (Ragan et al., 2006). In this section, I will focus specifically on gene content methods including phylogenetic profiling as well as gene order-based approaches. I will forego discussing approaches based on statistical properties (i.e., nucleotide composition) and other rare genomic changes (i.e., gene fusion and fission, indels, etc.).

Comparing genomes on the basis of the fraction of genes they share was among the first comparative genomics analysis developed with the availability of completely sequenced genomes. This world-view of genomes as a 'bag of genes' was used in many functional and evolutionary analyses, such as differential genomics (Huynen and Bork 1998) or phylogenetic profiles (Gaasterland and Ragan, 1998; Huynen and Bork, 1998; Pellegrini et al., 1999; Date and Marcotte 2003), to predict protein function.

Phylogenetic profiling predicts protein function by correlating the phylogenetic distribution of different sets of genes. The original form of phylogenetic profiling used binary vectors (i.e., 0 or 1) to indicate which species a putative homolog (based on BLAST e-value threshold) is present or absent (Gaasterland and Ragan 1998; Pellegrini et al., 1999). The idea is that during evolution genes that are functionally related are gained and lost together, which results in a correlation of their presence and absence patterns. These phylogentic profiles or phyletic distributions can not only be used for functional prediction but can also be used to look for atypical presence and absence patterns of genes which may be a result of LGT. Atypical patterns are recognized when genes have highly restricted distributions, present in isolated taxa but not among closely related species, or genes with high similarity to unrelated taxa.

Genome trees have also been generated using these approaches under the assumption that the number of shared orthologs between two genomes decreases with their divergence times (Fitz-Gibbon and House 1999; Huynen and Bork 1998). A wide variety of implementations have been developed using either using simple or model-based approaches (Huson and Steel 2004), each varying by the metric used to calculate an evolutionary distance between two genomes. For example, the Snel et al. (1999) and Korbel et al. (2002) methods used the fraction of orthologs shared between two genomes divided by the smallest genome to initially define evolutionary similarity and then converts the similarity score to an evolutionary distance. More recently, distance defined by mean normalized BLAST score for orthologs shared between a pair of genomes has been implemented (Clarke et al., 2002; Gophna et al., 2005). By defining distance in this way, their method captures sequence divergence and is not adversely affected by

genome-size differences. More importantly, these methods acknowledge that some genes will show conflicting patterns of relationship (i.e., potential LGT candidates), identify these discordant sequences and eliminates them from the analyses. However, despite these improvements these methods are strongly affected by LGT (Beiko et al., 2008) and gene loss (big/small genome attraction artifact; Lake and Rivera 2004; Wolf et al., 2001). Another potential issue may be the interdependence of gene content characters due to selective pressures that may operate on whole sets of functionally linked genes or operons, thus violating the i.i.d. assumptions made in phylogenetic inference from multiple character states (Steel 2002).

Gene order or neighborhood methods use the physical co-localization of genes to identify regions of ordered conservation within an individual or species. Regions of conserved linkage between species can convey some notion of ancestry, either suggesting a functional link among adjacent genes, and/or shared regulatory mechanisms (Von Mering et al., 2003; Yanai and DeLisi 2002), although the transfer of adjacent gene clusters between species must also be considered (Iwasaki and Takagi 2009; Lawrence 1999). Indeed, clusters of co-transcribed genes or operons are believed to be the principal form of gene organization and regulation in prokaryotes (Jacob and Monod 1961). Disruption of the operon would disrupt this co-regulation and thus, operons should be maintained by purifying selection (Jacob and Monod 1961). However, comparative analysis of bacterial and archaeal genomes has shown that only a few operons are conserved across large evolutionary distances due to prokaryotes' propensity for rearrangements (Mushegian and Koonin 1996; Watanabe et al., 1997). Thus, rearrangements to a genome introduced by reversal, transposition and other operations

such as duplication, deletion and insertion can alter the arrangement of these genes and potentially disrupt adjacent genes or operons. Studies have shown that the rate at which gene order evolves varies substantially between taxa such that only 5-25% of the genes in bacterial and archaeal genomes belong to gene strings (probable operons) shared by at least two distantly related species (Wolf et al., 2001). Furthermore, unlike gene content, gene order evolves at much faster rate and, therefore, these approaches are more suited for comparative studies of closely related species (Kunisawa 2001, 2006).

Gene order trees estimate an evolutionary distance indirectly by defining an edit distance or breakpoint distance to calculate the minimum number of events required to transform one genome in the other. The breakpoint is defined as an ordered pair of genes such that the two genes are adjacent in one genome but not the other, and the distance is simply the number of breakpoints in one genome relative to another. Due to the mathematical complexity of this combinatorial problem, only inversions (i.e., inversion distance) are generally considered, however, better approximation algorthims (Larget et al., 2005; Wang et al., 2006) are continuing to be developed to include explicit models of gene-order evolution (Moret et al., 2005). Additionally, simplified implementations have been devised that use the presences and absence of gene pairs to construct trees (Wolf et al., 2001, Korbel et al., 2002), similar to the gene content tree building practices. Again, these approaches are hampered by the same limitations seen among gene content approaches; they are strongly impacted by LGT as well as big/small genome attraction artifacts (Lake and Rivera 2004; Wolf et al., 2001).

1.2.4 The Unresolved Tree of Life

Despite improvements made to phylogenomic methods through increased taxon sampling, and new methods/models to reduce the impact of erroneous signal, there still remain long-standing phylogenetic questions spanning all levels of the tree of life. Indeed, phylogenomic studies in prokaryotes have been hindered by the prominent role of LGT in shaping the evolution of these microorganisms, particularly in devising appropriate multi-gene data sets in which each gene is recalcitrant to LGT. Nevertheless, supertree and supermatrix analyses have yielded phylogenetic trees similar to the corresponding SSU rRNA tree. However, the relationships between numerous groups remain tentative, including the relationship between Archaea and hyperthermophilic bacteria (Aravind et al., 1998; Nelson et al., 1999). One group in particular, the Aquificae, exhibit affinities with Archaea, ε-Proteobacteria, Thermotogae, and other Bacteria but typically cluster with the Thermotogae at the base of the prokaryotic tree in 16S rRNA and phylogenomic phylogenies.

Thermophilic organisms appear to have shared a particularly large number of genes, suggesting that LGT may have played a key role in emerging adaptations to very hot environments (Aravind et al., 1998; Nelson et al., 1999; Zhaxybayeva et al., 2009). Members of the genus *Aquifex*, such as *Aquifex aeolicus* VF5, are among the most extreme thermophilic bacteria known, occupying a habitat originally thought to be exclusively occupied by members of the archaeal domain. *Aquifex* lends its name to the order Aquificales and phylum Aquificae, a group based on 16S ribosomal RNA (rRNA) phylogeny (Cole et al. 2009) with considerable phylogenetic, ecological, morphological and metabolic diversity, including the freshwater, filamentous *Thermocrinis ruber*

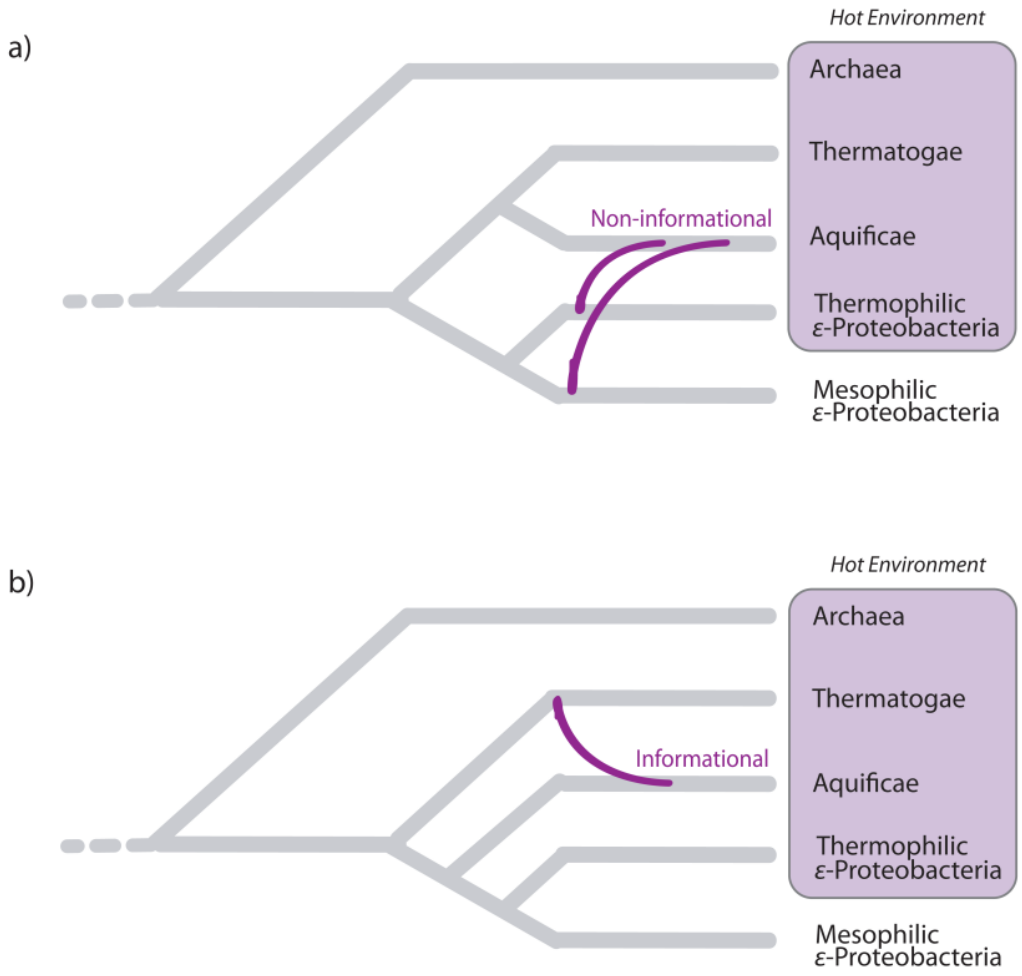


Figure 1.1: Two alternative hypotheses concerning the closest phylogenetic partners of the Aquificae. a) As suggested by 16S rRNA analysis and some concatenated protein phylogenies, the Aquificae are a deep branching phylum sister to the Thermotogae, with strong affinities for the ε-Proteobacteria due to large-scale gene sharing. b) The Aquificae are ε-Proteobacteria or a sister to this group, with extensive exchange of essential genes, either unidirectionally or reciprocally, with other thermophilic lineages such as Thermotogae and Archaea.

(Huber et al. 1998); the acidophile *Hydrogenobaculum acidophilum* (Stohr et al., 2001); and obligate anaerobes in the family Desulfurobacteraceae (L'Haridon et al. 2006).

Many studies have attempted to determine the evolutionary position of the enigmatic Aquificae phylum, with most results supporting one of two conflicting hypotheses (Figure 1.1). Either the Aquificae are most closely related to Thermotogae, a phylum containing many hyperthermophiles (Figure 1.1a); or the ε-Proteobacteria, a diverse class of Proteobacteria that includes environmental mesophiles, human-associated pathogens, and thermophilic and mesophilic species abundant in hydrothermal habitats (Figure 1.1b; Campbell et al., 2006; Nakagawa et al., 2007).

Analyses of widespread or universally distributed 'informational' genes involved in replication, transcription, and translation (Jain et al., 1999) that are believed to be relatively recalcitrant to LGT tend to place Aquificae near Thermotogae, in agreement with the 16S rRNA phylogeny. Datasets that supported this conclusion include the reciprocally rooted elongation factor Tu/G (Baldauf et al., 1996), RNA polymerase b/b_0 chain sequences (Bocchetta et al., 2000) and larger, concatenated alignment-based phylogenies of ribosomal proteins (Ciccarelli et al., 2006; Wolf et al., 2001). However, recent LGT studies have contradicted this claim. Beiko et al. (2005) reported a weakly supported *Aquifex+Thermotoga* affiliation ($\geq$ 0.95 posterior probability (PP) support among only 22 of the 110 constituent protein trees) with a larger number of protein trees supporting *A. aeolicus* as a basal member of the Proteobacteria, a sister to the ε-proteobacteria, or a lineage branching within this group. Moreover, phylogenetic profiling corrected for unequal taxon representation identified proteins with *Aquifex+Thermotogae* affinity (including the 22 constituent trees) to frequently co-occur

with the Archaea (Aravind et al., 1998; Zhaxybayeva et al. 2009), most notably the Euryarchaeota, suggesting that such proteins may have spread more recently via LGT. Additionally, genome trees based on gene content (Gophna et al. 2005; Korbel et al., 2002; Snel et al., 1999) and gene order (Korbel et al., 2002) have also identified *Aquifex* with the Archaea.

The alternative ε-proteobacterial affiliation was observed among other subsets of 'informational' genes such as the sigma transcription initiation factors (Gruber and D. A. Bryant 1998), the rpoBC operon (Klenk et al. 1999) and domain architecture studies of rpoC (Griffiths and Gupta 2004; Iyer et al., 2004). Biochemical studies of the cytochrome *bc* complex (Schütz et al. 2000) and cell wall characters (Cavalier-Smith 2002) also supported ε-proteobacterial affiliations. Furthermore, genome trees based on gene content (Korbel et al., 2002) and gene order (Wolf et al., 2001) have also placed *Aquifex* with ε-proteobacteria.

Based on the predictions of the complexity hypothesis, Boussau et al. (2008) focused on comparisons of counts between informational/non-informational gene families and their respective affiliations, and suggested that the *Aquifex* lineage was sister to the Thermotogae. However, biases in the selection of gene families by removal of potential LGT-derived gene families containing a non-monophyletic group neighboring *Aquifex* may have removed cases of LGT which have been informative for hypothesis testing. Wu and Eisen (2008) performed concatenated analyses of many widely distributed proteins and recovered a grouping of Aquificae with Thermotogae. An analysis using the same method, but including the genome of *Sulfurihydrogenibium* sp. YO3AOP1 as well, moved the phylum into a sister position to the ε-Proteobacteria (Wu

et al., 2009). This repositioning suggests that different members of phylum Aquificae show different degrees of affinity to other lineages, calling into question the existence of Aquificae as a monophyletic group.

In light of these apparently mosaic genomic affinities of the Aquificae, it is unclear which of the competing hypotheses regarding the positioning of phylum Aquificae (Figure 1.1) is correct; indeed it is unclear whether phylogenomic data can distinguish between these two (and potentially other) alternatives. If the scenario implied by aggregated analysis of 'informational' genes is correct, then the Aquificae are a deep-branching phylum, sister to the Thermotogae, while the ε-proteobacterial, archaeal and other affinities reflect large-scale "highways" of gene sharing (Beiko et al., 2005). Alternatively, the Aquificae may be unique ε-Proteobacteria, either descendents of a thermophilic or mesophilic ε-proteobacterial ancestor that exchanged essential genes either reciprocally or non-reciprocally with other thermophilic lineages (i.e., Thermotogae and Archaea) due to their common residence in very hot habitats. Although thermophilic ε-Proteobacteria have been identified from hydrothermal habitats using 16S rRNA analysis (Campbell et al., 2006; Nakagawa et al. 2005), only recently have the genomes of such organisms been sequenced. For example, the genomes of the thermal vent ε-Proteobacteria *Nitratiruptor* sp. SB155-2 (a thermophile) and *Sulfurovum* sp. NBC37-1 (a mesophile) were determined in Nakagawa et al. (2007). If, as suggested by their lifestyle, such organisms are the closest relatives of Aquificae, then their inclusion in genome-level studies should provide vital data in support of this relationship.

Given the apparent sensitivity of aggregation-based approaches (e.g., trees from concatenated alignments, and supertrees to a lesser extent) to taxonomic sampling and the

availability of additional genome sequences, consideration of phylogenetic trees both individually and at the level of molecular systems may provide new insights into the affinities of phylum Aquificae. Here I perform a comprehensive phylogenomic investigation of the genomes of three members of this group (comprising *A. aeolicus* and two recently sequenced isolates, *Hydrogeobaculum* sp. Y04AAS1 and *Sulfurihydrogenibium* sp. YO3AOP1) in light of previously published hypotheses and using a reference set of 774 completely sequenced prokaryotic genomes available as of December 2008.

Chapter 2

Materials and Methods

2.1 Genome Retrieval

In order to assess the phylogenetic position of the Aquificae group, I designed a dataset containing all sequenced genomes available in December 2008 from the National Center for Biotechnology Information (NCBI) FTP site (ftp://ftp.ncbi.nih.gov/genomes/) using the rsync (http://rsync.samba.org/) program. The set comprised 774 genomes, with 721 genomes from 20 bacterial phyla and 53 genomes from four archaeal phyla. The genome set included 633 mesophiles, 47 thermopiles, 28 hyperthermophiles, 14 psychrophiles and 52 with no identified temperature preference. Three Aquificae included in the dataset were *Aquifex aeolicus* VF5 (*Aquifex*), *Hydrogenobaculum* sp. Y04AAS1 (*Hydrogenobaculum*) and *Sulfurihydrogenibium* sp. YO3AOP1 (*Sulfurihydrogenibium*), all annotated as hyperthermopiles. The other phyla containing thermophiles and hyperthermophiles are shown in Table A1.

2.2 Putative Cluster Determination

In order to generate putative Aquificae clusters, BLASTP version 2.2.19 (using an expectation (e)-value of 1×10^{-3}, alignment view in tabular format (-m 8), upper limit of one line description of database sequences set (-v) to 100000, and upper limit of database sequence alignments set (-b) 100000; Altschul et al., 1990) was used to compare the encoded proteins for each of the three Aquificae genomes against the full set of genomes. Putative homologous gene sets or clusters were defined by first identifying all BLAST matches in which the one Aquificae species had a significant BLAST match (e-values ≤ 1

$\times 10^{-10}$) with another Aquificae. Next, a graph was constructed with each Aquificae protein sequence as a node and edges connecting pairs of proteins with bidirectional BLASTP matches. Using Kosaraju's linear time algorithm (Kosaraju 1978; Tarjan 1972) from the Graph perl module, clusters were selected by determining all strongly connected components (SCC; or subgraphs) within the initial graph such that each SCC node contains a path to every other node in the component and all SCC nodes were not a subset of any larger SCC(s). Aquificae clusters were then generated by merging all BLASTP matches (e-values $\leq 1 \times 10^{-5}$) reported for each connected Aquificae node by using one *Aquifex* node, when applicable, as the seed node and removing duplicated BLASTP matches.

2.3 Alignment Preparation and Phylogenetic Inference

Clusters identified with specific affiliations were aligned using fast statistical alignment (FSA) version 1.15.3 (using fast and maxsn commands; Bradley et al., 2009), and then Hmmer (version 3 using -trim command; Eddy 2011). hmmalign was used on existing FSA alignments (FSA+hmmalign) to generate profile hidden Markov model (HMM) alignments. Ambiguously aligned regions were masked when the consensus posterior (CP) probability for a column was less than 0.80. To reduce the size of large sequence sets, neighbor-joining phylogenies were inferred using the PHYLIP v3.68 package (Felsenstein 1989) : any genera with > 2 represented genomes that constituted a homogenous clan (Lapointe et al., 2010) in the neighbor-joining tree were reduced to two representative sequences, with one representative sampled from each descendant of the earliest implied split in that genus. The sets of retained congener sequences were re-aligned with FSA+hmmalign as above. Due to the high rates of LGT among the

Aquificae, I purposefully refrained from the use of supermatrix or supertree phylogenies opting to evaluate each gene individually. Maximum-likelihood phylogenies were inferred using RAxML 7.04 (Stamatakis 2006), using the WAG + Γ (four discrete rate categories) substitution model with 100 rapid bootstrap replicates.

2.4 Determination of the Cohesion Within the Aquificae Phylum

In order to determine whether the Aquificae are a distinct lineage, I implemented two separate strategies to assess the strength of support for the phylogenetic cohesion of the three Aquificae genomes. For specific subsets of clusters a ranked phylogenetic profiling and tree-based approach were utilized and are described in further detail below.

2.4.1 Ranked Phylogenetic Profiling Approach

Each cluster of proteins was first interpreted as a phylogenetic profile. For each profile, a protein *P* from *A. aeolicus* was assigned a rank of 1, and all other proteins in the profile were ranked in ascending order of BLASTP e-value (i.e., in decreasing order of statistical significance) obtained from a comparison using *P* as query and each other protein as subject. If *k* other proteins from phylum Aquificae were present in the profile, then their expected ranking would be 2, 3, …, $k + 1$ if the group were unaffected by LGT involving other phyla: clusters exhibiting this pattern were termed 'clean'. If, however, one or more proteins from members of other phyla were ranked higher than some Aquificae proteins, we termed the putatively orthologous set a 'dirty' cluster. Such patterns generally arise due to a) gene acquisitions by at least one of the Aquificae or non-Aquificae species, b) the inclusion of paralogs (i.e., ancient duplication and differential loss) introduced by the clustering approach and/or c) statistical artifacts (e.g., Koski and Golding 2001).

2.4.2 Tree-based Cohesion Approach

The ranked phylogenetic profiles approach was complemented by a tree-based assessment of the cohesion of the Aquificae clan in trees inferred with RAxML 7.0.4 (Stamatakis 2006). In cases where multiple Aquificae genomes were present along with genomes from non-Aquificae lineages, the resulting tree would either contain a single homogeneous clan (Lapointe et al., 2010) with all represented Aquificae genomes and no other genome (i.e., 'cohesive') or would contain a heterogeneous clan in which the members of the clan belong to more than one group of OTUs (i.e., 'non-cohesive').

2.5 Assessment of Relationship between Aquificae and Other Lineages

2.5.1 Phylogenetic Profile Construction

The orthologous clusters were analyzed in terms of their presence / absence distribution across all sequenced genomes (i.e., phylogenetic profiles: Gaasterland and Ragan 1998; Pellegrini et al., 1999). Given my focus on the putative origins of the Aquificae, I considered profiles in which at least one such genome (*Aquifex* = A; *Hydrogenobaculum* = H; *Sulfurihydrogenibium* = S) was represented, and then considered the presence or absence of Archaea (R), ε-Proteobacteria (E), Thermotogae (T), and other bacteria (O). Phylogenetic profiles could also be *exclusive* (designated with ø) to the lineages identified or potentially *inclusive* (designated with *) of other groups not explicitly named: for example, profiles designated ET-ø have at least one represented protein from Aquificae, Thermotogae, and the ε-Proteobacteria, and no other lineage, while profiles designated ET-* could potentially include representatives from other groups (e.g., other proteobacterial classes or Cyanobacteria) as well.

2.5.2 Phylogenetic and Bipartition Analysis

To access the conflicting phylogenetic signals within gene trees and identify sets of gene trees supporting particular hypotheses (i.e., a complete Aquificae clan adjacent to E, T and/or R groups). Each of the 100 bootstrap trees generated by RAxML (Stamatakis 2006) was represented as a set of splits to assess the relative positions of each operational taxonomical unit (OTU), with respect to a homogeneous clan of Aquificae. For any pair of taxonomic groups X and Y, the relative support for each of these two groups in association with phylum Aquificae was determined by enumerating the number of bootstrap trees in which group X was closer to the Aquificae (i.e., separated by fewer internal edges) than was group Y. The number of trees supporting X closer to the Aquificae are then subtracted from the number of trees supporting Y. The balance of support for X-Y ranged between 100 (all trees support a closer affinity of Aquificae to group X) to -100 if the reverse was true. Therefore, replicates in which both X and Y were equidistant to the Aquificae contributed 0 to the total score. The balance of support threshold ≥ 50 or ≤ -50 was applied to select subsets of gene trees to identify strong preferences for one affinity versus the other.

2.6 Functional Classification of Clusters

2.6.1 Clusters of Orthologous Groups (COGs)

All Aquificae gene sets were assigned functions based on the clusters of orthologous groups database (COGs; Tatusov et al., 1997), which contains 25 specific functional categories grouped into four parent categories. Here I used the following approaches: (1) clusters that contained *Aquifex aeolicus* VF5 were annotated by directly mapping the NCBI locus ID to the associated COG locus ID using the NCBI COG database available

at ftp://ftp.ncbi.nih.gov/pub/COG/COG/; (2) Clusters not assigned in the first step were annotated by determining the most frequent COG annotation among all BLAST matches with an e-value threshold of 1×10^{-15}; (3) Clusters that were restricted to the Aquificae or lacked a defined COG function were assigned Gene Ontology (GO) terms if the evidence codes were experimentally (IMP, IGI, IPI, IDA, or IEP) or computationally (ISS, IGC, or ICA) derived. GO terms were assigned COG functions by using the COG2GO database provided by Gene Ontology (Ashburner et al. 2000); (4) Clusters still lacking a COG or GO annotation were designated as unknown and assigned a functional role of 'poorly characterized'.

Characterizations of biological subsystems/pathways were first performed by identifying general phyletic patterns using the COG designations. For each COG category, a 'variable preference' index (VPI) was computed to contrast the affinities between R, E and T. This metric expresses the proportion of non-ubiquitous profiles that contained inclusive R, E, or T relative to the total number of profiles, excluding lineage-restricted profiles. For example, the VPI for ε-proteobacterial signal was calculated as all inclusive E profiles excluding RET (i.e., E+RE+ET) divided by the total number of profiles (i.e., RET+RE+RT+ET+R+E+T+Other). Comparisons of the VPI values across the 21 COG categories identified specific functional groups in light of the competing Aquificae hypotheses.

2.6.2 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways

To identify metabolic pathways and complexes within the COG classification scheme, each Aquificae NCBI RefSeq GI number was mapped to a KEGG orthology (KO)

number which consists of a manually defined, similarity- and positional-based orthologous gene set that corresponds to a node (enzyme or protein) in a specific KEGG pathway (or network; Kanehisa and Goto 2000). For each pathway, a manually drawn reference (denoted by a KO number) was constructed to identify the presence / absence of genes within the network of nodes. Metabolic pathways, which were generally widely conserved, were represented with one manually drawn reference pathway from which many organism-specific pathways were computationally generated. Conversely, regulatory pathways were found to be far more divergent and required the construction of separate organism-specific pathways by identifying reference pathways common among groups of organisms (e.g., three ribosomal assembly diagrams for Bacteria, Archaea and Eukaryota). Each enzyme or protein present in an Aquificae metabolic or regulatory pathway was coupled with manual curation of their associated putatively orthologous cluster and subjected to phylogenetic and bipartition analysis.

2.7 Recombination Analysis

To determine if homologous recombination occurred among specific functional groups, I implemented a recombination detection procedure using a two-phase strategy based on Chan et al. (2007). This strategy first detects the occurrences of recombination events in the sequence dataset and then identifies breakpoints of such events. The first phase is a quick screening of the dataset including only members of the Aquificae, ε-Proteobacteria, Thermotogae and Archaea for possible recombination events using PhiPack (Bruen et al., 2006). PhiPack utilizes the neighbour similarity score (NSS; Jakobsen and Easteal 1996), maximal chi-squared (MaxChi; Smith 1992) and pairwise homoplasy index (PHI; Bruen et al., 2006) to evaluate phylogenetic discrepancy across the set of sequences. In the

second phase, datasets detected by PhiPack to contain a recombination event are then analyzed to determine the corresponding breakpoints using another approach called DualBrothers (Minin et al., 2005). This Bayesian phylogenetic approach uses reversible jump MCMC and dual multiple change-point model (MCP) to infer changes in tree topology and evolutionary rates across site within the dataset.

Chapter 3

Results

3.1 Affinities within the Aquificae

Using the protein-coding genes of the three Aquificae genomes as seeds for the clustering algorithm, 2295 clusters (2019 (88% of 2019 clusters) putatively orthologous, single-copy and 276 (12% of 2019 clusters) multiple-copy clusters containing either Aquificae in- or out-paralogs) were generated using the method described in the previous chapter. Among the 2019 single-copy Aquificae clusters (Figure 3.1), 1204 (60% of 2019 clusters) were exclusive to one of the three Aquificae (A, H, or S) and 288 (14% of 2019 clusters) were represented in two Aquificae (AH, AS, HS). These subsets correspond to 'variable' genes with a patchy distribution due to shared ancestry and subsequent gene loss in one or more Aquificae lineages, genes invented in specific lineages, and/or LGT. The Aquificae 'core' comprising 527 (26% of 2019 clusters) shared clusters (designated as AHS) could potentially represent genes that were present in a common ancestor and retained in all sampled descendent lineages.

3.2 Cohesion of the Aquificae Phylum

In order to assess the cohesion of the Aquificae phylum, I focused on 527 phylogenetic profiles containing single-copy representatives of all three sampled Aquificae genomes (AHS). Evaluation of the phyletic distribution (Figure 3.2) as described in the previous chapter, revealed that the two most frequent ‘core’ AHS clusters were the 350 (66% of the 527 ‘core’ clusters) ET-* profiles (i.e., RET and ET subsets) that presumably contain universal or semi-universal gene sets. Moreover, the 350 ET-* profiles were

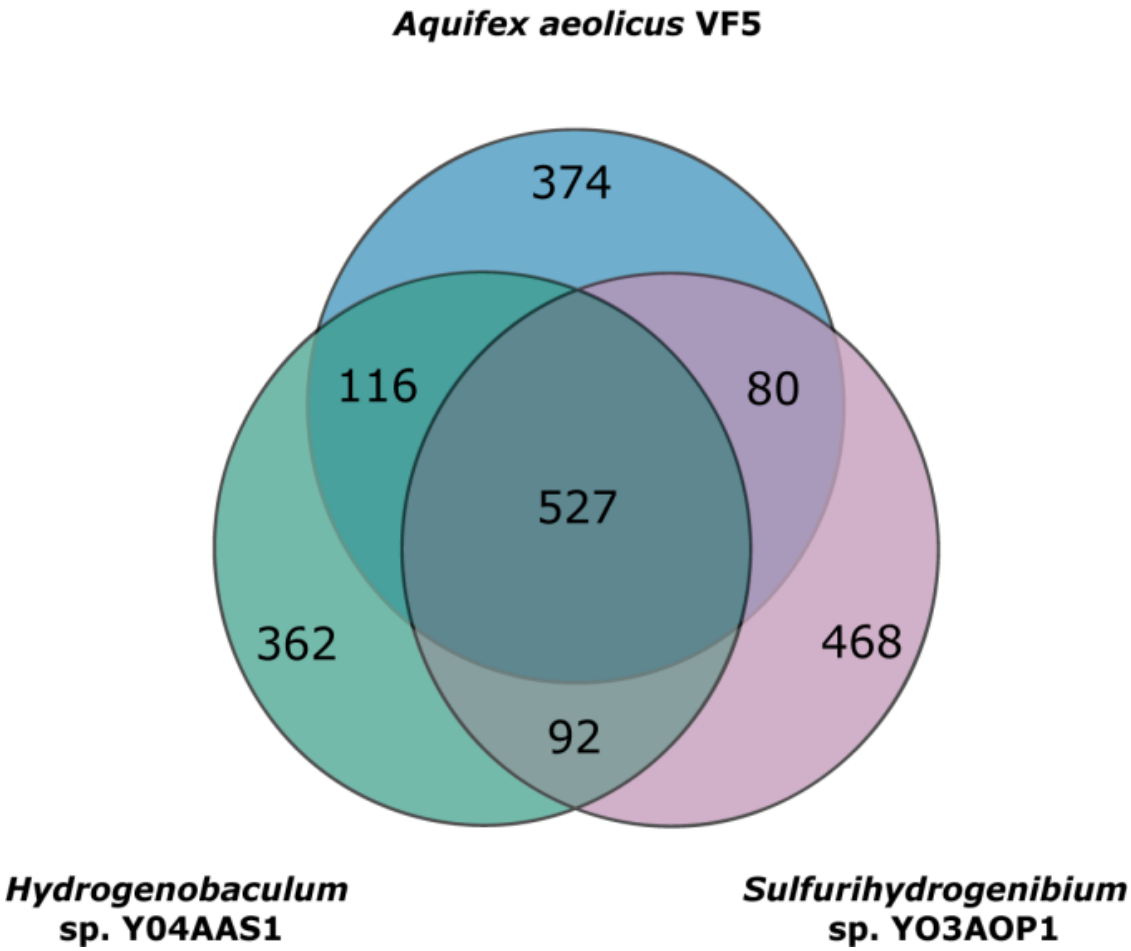


Figure 3.1: Overlap in homologous gene content among the three sequenced members of phylum Aquificae.

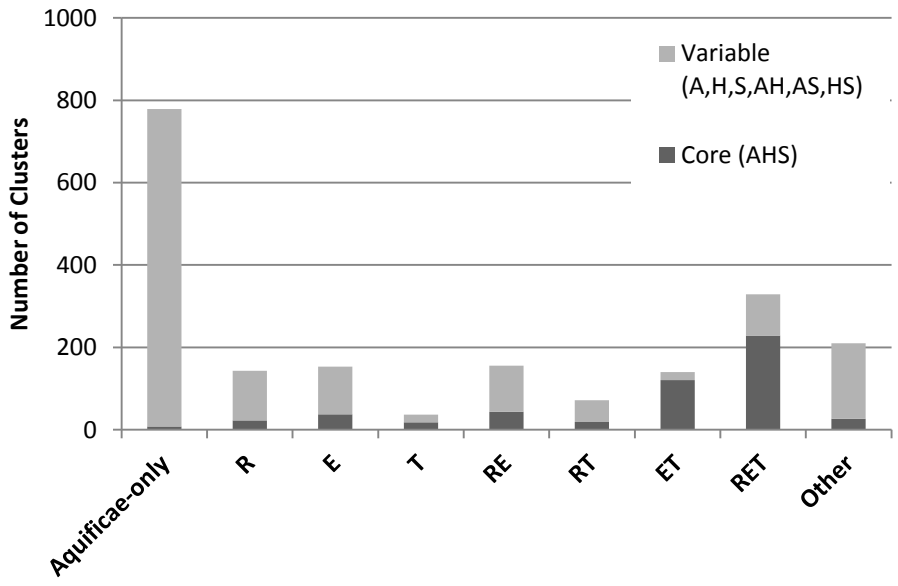
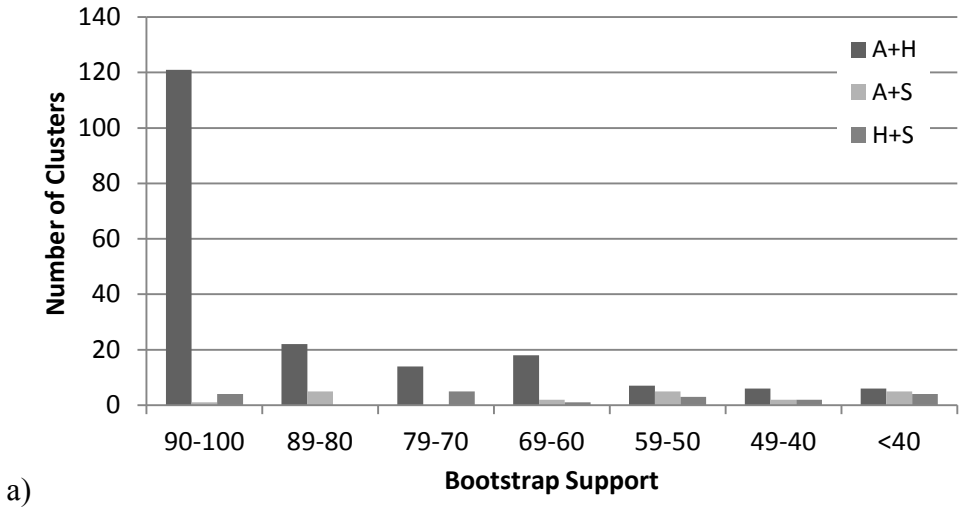


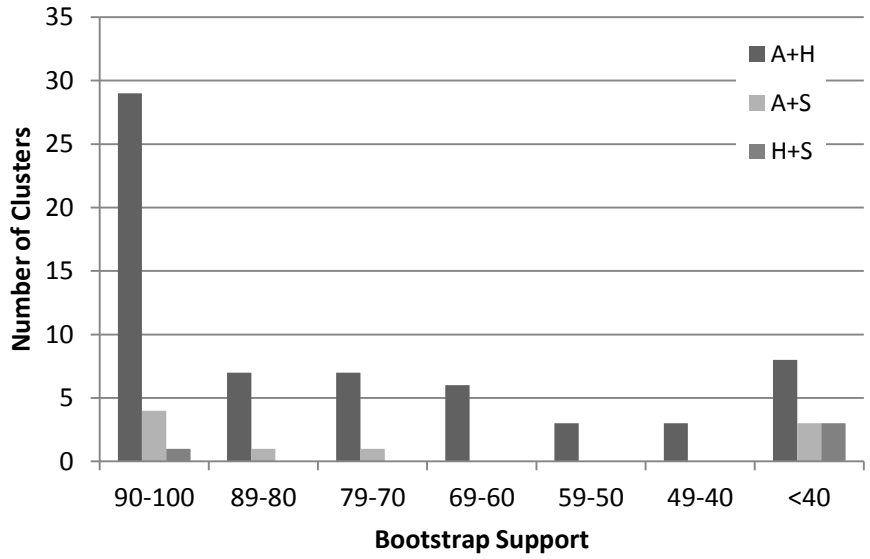
Figure 3.2: Summary of all phylogenetic profiles of all Aquificae subsets, subdivided into the core AHS subset and the variable subset. Single letter abbreviations are as follows: *Aquifex* = A; *Hydrogenobaculum* = H; *Sulfurihydrogenibium* = S) was represented, and then considered the presence or absence of Archaea (R), ε-Proteobacteria (E), Thermotogae (T), and other bacteria (O).

distinguished by the presence (229 RET) or absence (121 ET) of Archaea, i.e., gene sets found in both archaeal and bacterial domains or those specific to the bacterial domain. However, other 'core' clusters contain highly restricted distributions present in isolated taxa, e.g., the Aquificae restricted (AHS-ø; 8; 2% of the 527 'core' clusters), R-only (R-ø; 23;4% of the 527 'core' clusters), E-only (E-ø; 38; 7% of the 527 'core' clusters), T-only (T-ø; 18; 3% of the 527 'core' clusters) and Other-only (Other-ø; 27;5% of the 527 'core' clusters), suggesting that LGT may be present even among the 'core' clusters which are presumed to be vertically inherited.

To assess whether the three sampled Aquificae constitute a distinct lineage through ancestry rather than LGT generating the appearance of a cohesive grouping, I applied both the ranked BLAST approach and tree-based approaches to the 350 ET-* profiles as described in the previous section 2.4. The ranked approach identified 230 (64% of the ET-* profiles) 'clean' profiles consisting of 154 (67% of the 229 RET profiles) RET and 76 (62% of the 121 ET profiles) ET sets, and 190 (36% of the 350 ET-* profiles) 'dirty' profiles consisting of 75 (33% of the 229 RET profiles) RET and 45 (38% of the 121 ET profiles) ET sets. Similarly, the tree-based cohesion analysis identified 160 (71% of the 225 RET trees) RET and 77 (65% of the 118 ET trees) ET trees with a homogeneous (cohesive) Aquificae clan (Figure 3.3a); trees of the remaining 65 (29% of the 225 RET trees) RET and 42 (35% of the 118 ET trees) ET profiles yielded a heterogeneous (non-cohesive) Aquificae clan (Figure 3.3b). The expected pattern that clean profiles would be correlated to cohesive clans or dirty profiles as non-cohesive clans were generally observed. However, in some cases, 'clean' profiles were identified as non-cohesive clans and 'dirty' profiles as cohesive clans. These case were



a)



b)

Figure 3.3: Bootstrap distributions for the pairings of different Aquificae among the a) 237 cohesive and b) 107 non-cohesive maximum likelihood trees of the AHS core subset where A=*Aquifex*, H=*Hydrogenobaculum* and S=*Sulfurihydrogenibium.*

observed more frequently among the ET (30% of the 118 ET profiles/trees) than in RET (18% of 225 RET profiles/trees) sets suggesting that the RET subset are more recalcitrant to LGT than the ET subset. Furthermore, evidence from phylogenetic profiles and trees suggests that the Aquificae do constitute a distinct lineage, albeit one that is frequently affected by LGT.

I performed further analyses on the 350 ET-* profiles to assess the branching order among the three Aquificae to assess the potential rate of within-Aquificae LGT. Upon removal of six clusters containing large sequence sets (>1500 sequences) generated by multiple non-Aquificae copies, I assessed the bootstrap support among members of the homogeneous Aquificae clan. Two hundred (58%) of the remaining 344 trees had an associated bootstrap support value of 70% or greater for the pairing of *Aquifex* + *Hydrogenobaculum*, as compared with only 4% (12/344) supporting *Aquifex* + *Sulfurihydrogenibium* and 3% (10/344) supporting *Hydrogenobaculum* + *Sulfurihydrogenibium*. A further 10% (35/344) contained no grouping of the Aquificae (Figure3.3a, 3.3b). The dominant branching pattern of *Aquifex* + *Hydrogenobaculum* as an adjacent group together with *Sulfurihydrogenibium* in a cohesive clan, is consistent with 16S-based taxonomy (Cole et al. 2009) which places *Aquifex* and *Hydrogenobaculum* together in family Aquificaceae, with *Sulfurihydrogenibium* joining the other two only at the class level (Aquificales). However, these results suggest that within-Aquificae transfer may occur or systematic bias such as compositional biases or heterotachy may artificially cause these alternative branching patterns.

3.3 Affinities of Aquificae with Key Groups

In order to assess the overall genomic affiliations of the three Aquificae, I analyzed the phylogenetic profiles (Figure 3.2) by first focusing on the distribution of profile counts across both the variable (A, H, S, AH, AS, HS) and core (AHS) sets, then the core set specifically. Among the 2019 single-copy profiles identified in Figure 3.1, 779 (39% of the 2019 clusters) were found only in the Aquificae. In the remaining 1240 clusters, both exclusive and inclusive profiles of the combined variable and core clusters revealed more profiles with an ε-Proteobacteria affinity (E-*;778;63% of the 1240 profiles and E-ø;153; 13% of the 1240 profiles, respectively) than to either Archaea (R-*;700;56% of the 1240 profiles) and R-ø;143;12% of the 1240 profiles, respectively) or Thermotogae (T-* ;578;47% of the 1240 profiles and T-ø; 37;7% of the 1240 profiles, respectively). Within the E-* subset, more matches were associated with the thermal vent ε-Proteobacteria *Nitratiruptor* sp. SB155-2 (588; 76% of the 778 E-* profiles) and *Sulfurovum* sp. NBC37-1 (550; 71% of the 778 E-* profiles) than to any other single ε-proteobacterium. Additionally, among the subset of 527 profiles that covered all Aquificae (AHS), the E-* count (432; 83% of the 527 profiles) was greater than T-* (387; 75% of the 527 profiles) and R-* (315; 61% of the 527 profiles) suggesting that the core AHS clusters are less likely to be influenced by LGT from the Archaea than the variable clusters (Figure 3.2).

3.4 Affinities of Aquificae with Other Groups

Other taxonomic groups also showed genetic affinities to the three Aquificae (AHS subset). The most notable affiliations were to organisms with large genomes such as the delta (δ)-Proteobacteria from the genus *Geobacter* (416-424 depending on *Geobacter* species and strain; 80-82% of the 527 'core' profiles) and/or the

thermophilic/hyperthermophilic organisms including a member of the Nitrospirae *Thermodesulfovibrio yellowstonii* DSM 11347 (420; 81% of the 527 'core' profiles), and *Carboxydothermus hydrogenoformans* Z-2901 (381; 73% of the 527 'core' profiles), *Thermoanaerobacter tengcongensis* MB4 (352; 68% of the 527 'core' profiles) and *Caldicellulosiruptor saccharolyticus* DSM 8903 (357; 69% of the 527 'core' profiles) in class Clostridia. Indeed, the *δ*-Proteobacteria and Clostridia contain a vast repertoire of genes, particularly metabolic genes, believed to have been acquired by LGT (Beiko 2011; Dagan et al., 2010; Gophna et al., 2006). The majority of these relationships cannot reflect vertical signal (especially those involving the Gram-positive Clostridia), and the substantial number of affinities with many different groups raises the question of whether Aquificae can be phylogenetically placed at all without giving special status to a small subset of genes such as 16S or the ribosomal apparatus.

3.5 Affinities of Each Individual Aquificae Genome

As described in the introductory chapter, the analysis of Wu et al. (2009) demonstrates that different members of phylum Aquificae show varying degrees of affinity for other lineages. To address the affiliations of each Aquificae genome individually, I contrasted phyletic patterns between the lineage-restricted subsets (e.g., *Aquifex*-only: A) and the inclusive (*) Aquificae subsets (e.g., *Aquifex* and possibly others: A, AH, AS and AHS; Figure 3.4). Each inclusive Aquificae subset (colored bars in Figure 3.4) showed a similar breakdown of affinities to other major lineages: affinities with *ε*-Proteobacteria were observed more frequently than Archaea (R-*), which were in turn more frequent than Thermotogae. Among the Aquificae-only sets (gray bars in Figure 3.4), 704 (58% of the 1204 profiles) were represented in only a single member of the Aquificae (i.e., 'orphan'

proteins with respect to the set of 774 genomes considered), with H highest (37% of the 704 Aquificae-only profiles), followed by S (36% of the 704 Aquificae-only profiles) and A (28% of the 704 Aquificae-only profiles). Comparisons of the remaining 500 lineage-specific profiles against the three inclusive Aquificae subsets (A-*, H-* and S-*) identified the *Aquifex* genome to be strongly influenced by both the Archaea (R-*: 409; 57% of 723 AS, AH and AHS profiles) and Thermotogae (T-*: 439; 61% of the 723 AS, AH and AHS profiles), *Hydrogenobaculum* by only the Archaea (R-*: 421; 63% of AH, HS and AHS profiles) and *Sulfurihydrogenibium* by only ε-Proteobacteria (E-*: 502; 75% of AS, HS, and AHS profiles).

3.6 System-level Analysis

Following the bulk characterization of taxonomic affinities of Aquificae proteins, I analyzed the contributions of Archaea, ε-Proteobacteria, and Thermotogae-associated genes in terms of their functional role in the cell. If most or all constituents of a molecular system show similar patterns of inheritance, subsystem analyses may help to identify the global Aquificae affinities and the potential origins of the Aquificae rather than aggregated counts of individual gene phylogenies (Doolittle and Zhaxybayeva 2009). To assess the placement of the Aquificae among gene trees, a balance of support threshold of ≥ 50 or ≤ -50 was applied to each molecular system to identify strong preferences for one hypothesis versus another.

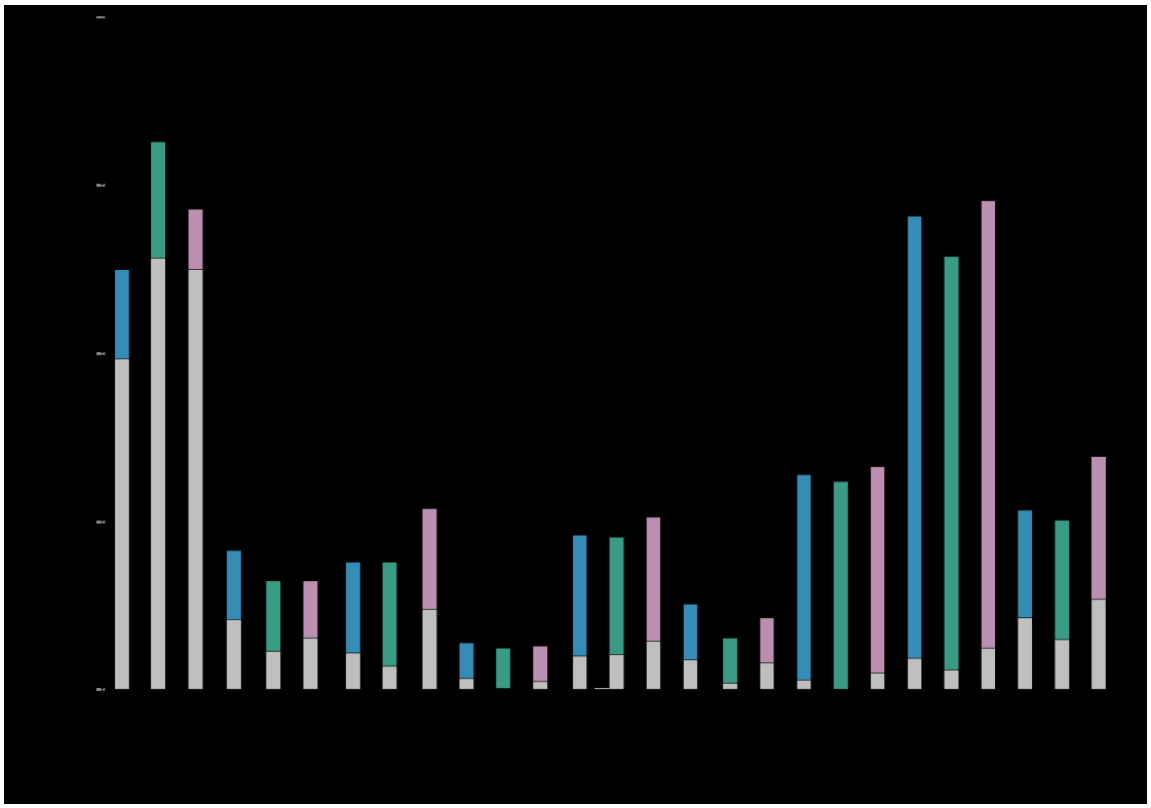


Figure 3.4: Summary of all phylogenetic profiles for each Aquificae genome (*Aquifex* = A; *Hydrogenobaculum* = H; *Sulfurihydrogenibium* = S), subdivided into lineage-restricted subsets (A-only, H-only, S-only in grey) and inclusive Aquificae subsets (in color) whereby the inclusive *Aquifex* subsets (blue) includes AH, AS, and AHS clusters, the inclusive *Hydrogenobaculum* subsets (green) includes AH, HS, and AHS clusters and the inclusive *Sulfurihydrogenibium* subset (purple) include AS, HS, and AHS clusters.

3.6.1 Phylogenetic Profile Analysis of Functional Groups

In a manner described in section 2.6, all 2295 single and multi-copy clusters were classified according to the clusters of orthologous groups (COG; Table 3.1) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways databases in order to identify broad functional groups and metabolic/regulatory pathway with interesting phyletic patterns (Table 3.1and Table 3.2). 63% (1541/2433) of all clusters were labeled with one or more COG designations and mapped to one of four broad categories with 36% (552/1541) labeled as 'metabolism', 24% (368/1541) as 'cellular processing and signaling', 19% (296/1541) as 'information storage and processing' and the remaining 21% (325/1541) were 'poorly characterized'.

I assessed the relative degree of support for affinities of Aquificae with R, E and T by computing a 'variable preference' index (VPI) that expresses the number of non-ubiquitous profiles that contained R, E, or T relative to the total number of profiles, excluding those found only in the Aquificae. Comparison of the VPI values for each of the three distinct lineages across the 21 COG categories (see Table A.2) indicates the degree to which different lineages were found in association with the phylum Aquificae. As partner to Aquificae, the ε-Proteobacterial group was dominant in a handful of categories (Figure 3.5), notably cell wall biogenesis (M), intracellular trafficking and secretion (U), and lipid biosynthesis and transport (I); and to a lesser extent transcription (K) and secondary metabolites (Q). The Archaea were frequent partners in many metabolic categories, with the notable exception of lipids, which differ substantially in their composition between Bacteria and Archaea, posttranslational modification and proteins of unknown function. Translation (J) was the only one of 21 COG categories in

which Thermotogae have the largest VPI score; even in this case their score (0.43) was not considerably greater than that of the ε-Proteobacteria (0.38). The ε-Proteobacteria were well represented in all 21 categories, with a VPI score that was always greater than half of the best VPI score for a given category. Conversely, in many cases the VPI for either Thermotogae or Archaea was much lower than that of the two other lineages. The Archaea had VPI scores < 0.15 for several cellular process categories including cell cycle, cell wall biosynthesis, motility (VPI = 0), and trafficking; outside this group, translation and unassigned functions also had VPI scores < 0.15. The Thermotogae had low VPI scores in the signal transduction category, along with energy production, and the metabolism of amino acids, coenzymes, inorganic ions and secondary metabolites. Cases in which VPI for ε-Proteobacteria was lowest (< 0.2) correspond to functions that are most widespread: over half of all protein sets in the amino acid and nucleotide transport categories have RET distributions, which uniformly decrease all VPI scores since ubiquitous profiles can only decrease the VPI.

Indeed, the primary focus and debate regarding the descent of the Aquificae revolves around the identification of genetic markers that appropriately describe the relationships of the Aquificae to the ε-Proteobacteria, Thermotogae and Archaea. Analysis of phylogenetic profiles coupled with functional classification identified translation (J), cell wall biosynthesis (M), and cell motility (N) to each contain distinct affiliations which may help differentiate between the two opposing Aquificae hypotheses (see Figure 1.1). Moreover, supertree analyses revealed that phylogenies of genes involved in protein synthesis and cell wall biosynthesis were less discordant when compared against the 'species' tree than the protein-coding genes of cell motility

Table 3.1: Phyletic pattern breakdown of all 2433 Aquificae COGs at the parent level classification

Parental Category	RET	ET	RE	RT	R	E	T	Aq-only	Other	Total	R-*	E-*	T-*
Cellular processes and Signaling	94	69	33	11	18	42	8	48	45	368	156	238	182
Information storage and Processing	104	66	16	18	18	10	12	34	18	296	156	196	200
Metabolism	253	25	103	31	49	39	3	23	26	552	436	420	312
Poorly Characterized	62	20	59	24	77	88	21	708	158	1217	222	229	127

Table 3.2: Phyletic pattern breakdown of four biological subsystems of interest identifying the number of ubiquitous, inclusive and exclusive R, E and T profiles.

KEGG Pathways	RET	R-*	E-*	T-*	E not T	T not E	R not E	R not T
Ribosome	17	18	40	43	2	4	1	0
Flagellar assembly	0	0	16	17	1	2	0	0
Lipopolysaccharide biosynthesis	1	3	10	1	9	0	0	2
Oxidative phosphorylation	13	29	29	16	14	0	8	15

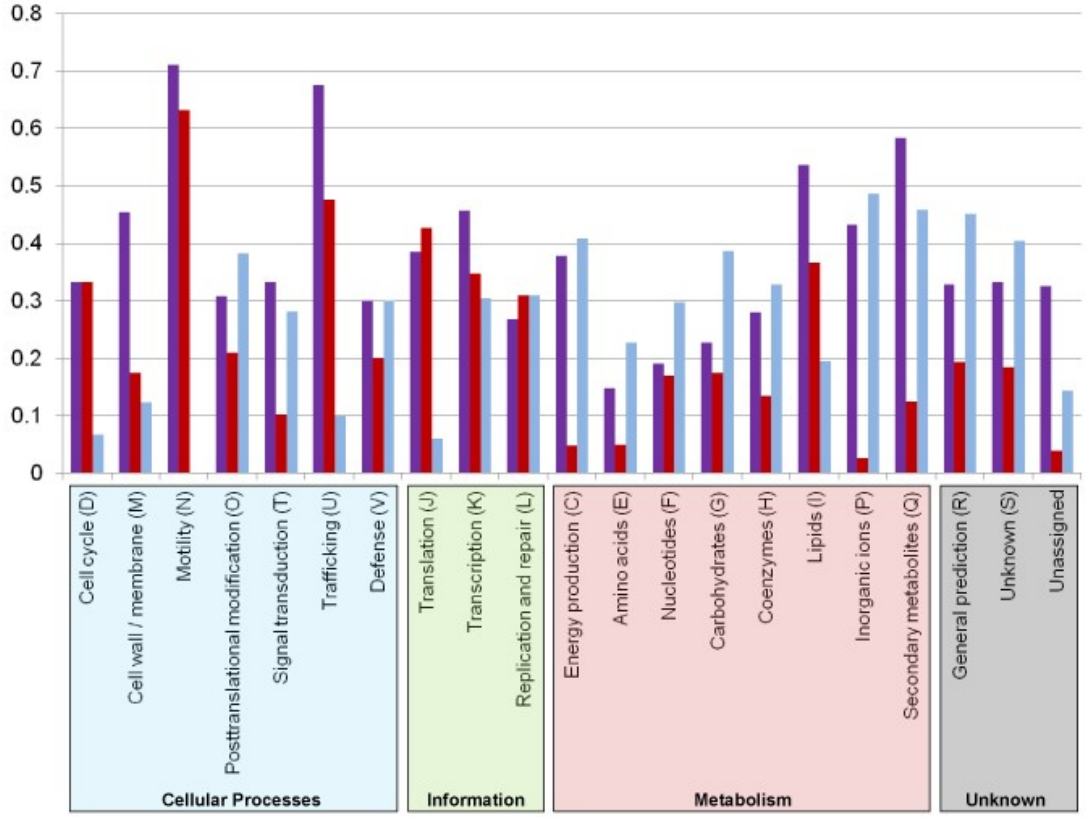


Figure 3.5: Relative support for the affinities of Aquificae with Archaea (blue), ε-Proteobacteria (purple), Thermotogae (red) evaluated with the variable preference index (VPI) which expresses the number of non-ubiquitous profiles that contain inclusive Archaea, ε-Proteobacteria, or Thermotogae counts relative to the total number of profiles, excluding those found only in the Aquificae.

(Beiko et al., 2005). These results conflict with the *a priori* assumptions of the complexity hypothesis (Jain et al., 1999) whereby cell wall biosynthesis was thought to be more susceptible to LGT. Other studies (Cavalier-Smith 2002; Plötz et al., 2000; Slonim et al., 2006; Wang and Quinn 2010) suggested that particular components of this subsystem are

highly conserved across different Gram-negative bacteria and inherited vertically. Conversely, the archaeal affiliation (Aravind et al.,1998) of energy metabolism (oxidative phosphorylation) suggests that particular genes or complexes involved in respiration were acquired from the Archaea to confer some niche-specific adaptation (Boucher et al., 2003).

3.6.2 Ribosomal Structure and Biogenesis

Ribosomal proteins, which are complexed with rRNA to form ribosomes during the cellular process of translation, are widely distributed informational genes containing many protein-protein and protein-rRNA interactions thought to be refractory to transfer between divergent species according to the complexity hypothesis of Jain et al. (1999). Thus, the ribosomal proteins are frequently used in microbial classification to infer organismal phylogeny through the use of concatenated phylogenies (Ciccarelli et al., 2006; Matte-Tailliez et al., 2002). However, barriers to transfer of ribosomal components are not absolute: for instance, Asai et al. (1999) showed that in *Escherichia coli* the rRNA operon can be successfully replaced by that of a distantly related species; furthermore, cases of LGT have been observed in the rps14 ribosomal protein in Bacteria (Brochier et al., 2000).

To understand the evolutionary processes and LGT susceptibility of the Aquificae translational machinery, I evaluated 34 proteins of the ribosomal complex. This consisted of 20 large subunit (LSU) and 14 small subunit (SSU) proteins from a reduced set of 47 KEGG-annotated (ko03010) clusters through the removal of six Aquificae-restricted clusters (L29, L30, L32, L35, S20 and S21) and an additional seven clusters determined by Brochier et al. (2005) to contain evidence of LGT within the archaeal domain. Despite suggestions that the ribosomal complex is recalcitrant to LGT (Jain et al.,1999), the reduced dataset contained a diverse distribution of phyletic signatures (RET, RT, ET, T, E and Other). Moreover, analyses of the exclusive profiles and VPI identified similar Thermotogae and ε-Proteobacteria distributions (T-*; 31; 91% vs E-*; 30; 88% and VPI: T: 47% vs E: 41%). Bipartition analysis of the individual gene phylogenies identified 23 of 34 trees (67%) where the Thermotogae were adjacent to a cohesive Aquificae clan; however, in all cases the sister to Aquificae consisted of Thermotogae coupled with at least one other major lineage such as Clostridia, δ-proteobacteria, or ε-Proteobacteria.

Studies of the *E. coli* ribosome identified the organization of the ribosomal assembly into five large operons where approximately half of the protein-encoded genes map to the five operons: *str*, *spc*, *S10*, *α*, and *L11+rif*, and the remaining genes are scattered around the genome in clusters of size one to four (Lindahl and Zengel 1986). Among the three Aquificae, the organization of the five operons in *Hydrogenobaculum* and *Sulfurihydrogenibium* genomes were sequentially arranged as *L11+rif+str+S10+spc*+X+*α* unlike *Aquifex*, which contains the operon arrangement (*S10*+X+*α*+*str*+*spc*+*str*+*L11+rif*) such that all but the *L11+rif* operons are present in a different order and location. The ordering of the co-regulated genes within each operon

(Figure 3.6) including X, which contains four ubiquitous genes - adenylate kinase (*kad*), methionine aminopeptidase (*map*), translation initiation factor 1 (IF-1) and ribosomal L36 - located between the SecY gene and *α* operon, were conserved among all Aquificae. Interestingly, similar ordering of the four genes was observed in the Thermotogae and Dictyoglomi lineages but not among the *ε*-Proteobacteria, which lacked the *kad* gene. Bipartition analysis of the 25 co-regulated genes present in the six operons identified the Thermotogae as frequently (72% - 18/25) adjacent to the Aquificae clan, although the proximity of these two groups to the exclusion of *ε*-Proteobacteria and Archaea was not always strongly supported by the bootstrap analysis. Among the remaining co-regulated genes, the sequential S19 and L22 genes found in the *S10* operon, the L6 and S5 genes of the *spc* operon and S13 of the *α* operon were adjacent to Euryarchaeota, and L14 and S4 of the *spc* and *α* operons, respectively, were adjacent to the *ε*-Proteobacteria. The consistent Thermotogae affiliation among the single-gene phylogenies suggests either shared ancestry with the Thermotogae (Boussau et al., 2008) where the archaeal (L22; Coenye and Vandamme 2005) and *ε*-proteobacterial affiliations were laterally acquisitions or the superoperon (i.e., the five operons) was acquired laterally from the Thermotogae followed by independent gene displacement of the 7 genes with alternative affiliations (Omelchenko et al., 2003). Furthermore, the PhiPack program identified no instances of recombination among the genes of the *S10* operon. The remaining nine ribosomal proteins (L13, L19, L20, L21, L31, S2, S6, S9 and S15), which were scattered around the Aquificae genomes in small clusters, lacked matches to the Archaea and contained identical inclusive Thermotogae and *ε*-Proteobacteria profile counts (E-* and T-*: 7/9; 78%). Bipartition analysis of the six ribosomal proteins containing ET profiles

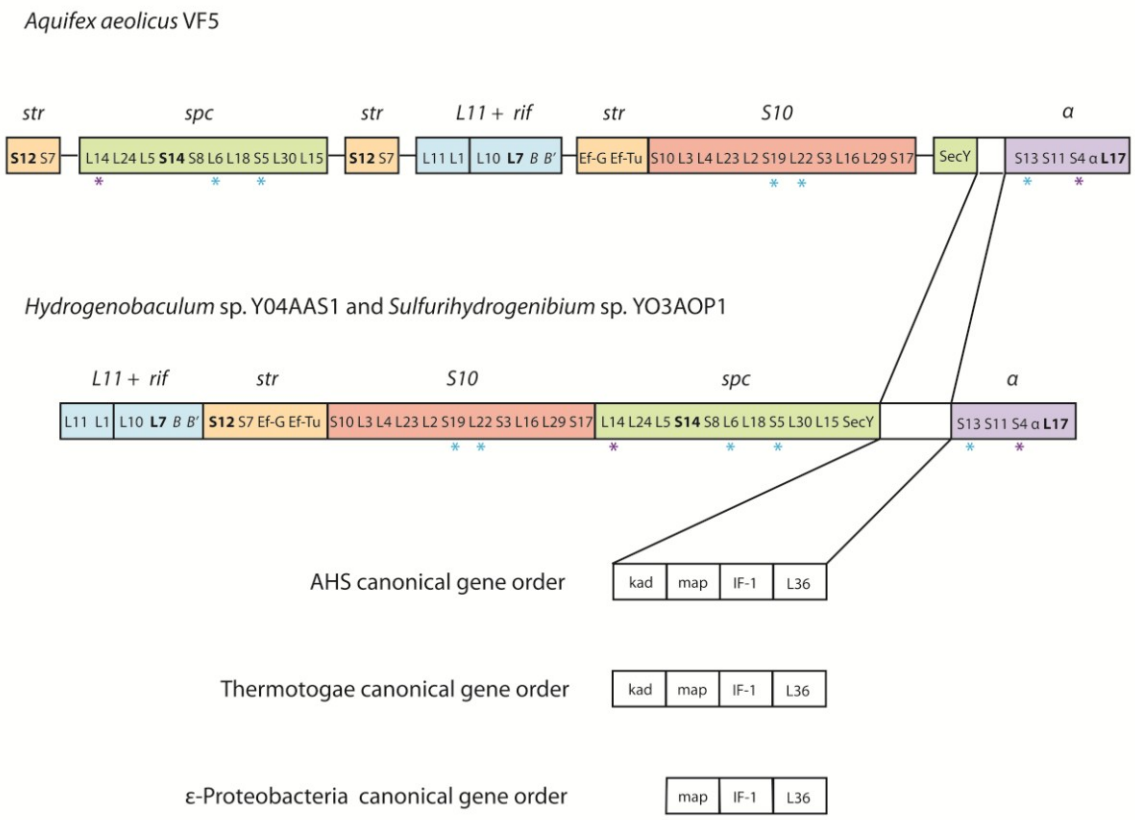


Figure 3.6: Linear arrangement of five ribosomal operons: L11+ rif (blue), str (orange), S10 (red), spc (green) and α (purple) and their respective gene order in *Aquifex*, *Hydrogenobaculum* and *Sulfurihydrogenibium*. Purple asterisks represent Aquificae genes with ε-proteobacterial affinity and blue asterisks indicate archaeal affinities. *Aquifex* contains a novel operon arrangement where the majority of the individual genes are located in different locations (denoted by a line separating the operons), whereas *Hydrogenobaculum* and *Sulfurihydrogenibium* have a compact operon arrangement with four genes (adenylate kinase (kad), methionine aminopepitase (map), translation initiation factor 1 (IF-1) and ribosomal L36) separating the spc and α operons. The gene order was conserved amongst all Aquificae and Thermotogae, however, among the ε-Proteobacteria the kad gene is absent.

revealed the same trend, identifying three trees where either Thermotogae or ε-Proteobacteria were more often adjacent to the Aquificae than the other. Additionally, gene order conservation among the three Aquificae differed; *Sulfurihydrogenibium* and frequently *Hydrogenobaculum* contained similar gene arrangements as were found in ε-Proteobacteria.

3.6.3 Cell Motility: Flagellar Assembly

The bacterial flagellar system is both a motor organelle and a protein export/assembly apparatus extending from the cytoplasm to the cell exterior which plays a central role in cell motility, adhesion, biofilm formation and host invasion (Harshey 2003). Recent evolutionary analysis of the flagellar complex (Liu and Ochman 2007) suggests that the 'core' flagellar genes were derived from a single ancestor through successive duplications and diversifications where LGT played only a minor role. However, Doolittle and Zhaxybayeva (2007) refuted this claim arguing that faulty BLAST settings, the disregard of seven discordant gene trees (potential LGT-driven events) involved in 14-gene concatenated phylogeny, and biased comparisons between the concatenated tree to the 'species' tree (reconstructed from mostly ribosomal proteins) downplay the role of LGT of this complex.

To investigate evolutionary processes and LGT susceptibility of the Aquificae flagellar assembly, I evaluated twenty-three Aquificae clusters identified as components of the flagellar assembly pathway (ko02040). Removal of seven Aquificae-restricted (FlgA, FlgB, FlgD, FlgL, FlgM, FliE, and FliN/FliY) and three multi-copy clusters (FlhB, MotB/MotB-like, and FlgG+FlgE) reduced the dataset to 13 putatively orthologous clusters. The observed phyletic patterns were consistent with the independent

origins of the bacterial and archaeal flagellar machinery (Ng et al., 2006) with all profiles lacking archaeal (R) signal. Phylogenetic analysis identified nine trees in which the thermophilic ε-proteobacterium *Nitratiruptor* sp. SB155-2 was sister to *Sulfurihydrogenibium* within the Aquificae clan. Bipartition analysis revealed that the seven trees (54%; 7/13) contained the mesophilic ε-Proteobacteria (Liu and Ochman 2007) adjacent to the Aquificae + *Nitratiruptor* group while in the remainder of cases (46%; 6/13) this group was adjacent to a Thermotogae clan.

Gene ordering in the flagellar operons of the Aquificae was similar to the well-studied regulon of *Salmonella enterica* serovar *typhimurium* (Chilcott and Hughes 2000) and revealed that the ordering of 30 flagellar genes was well conserved between the thermophilic *Nitratiruptor* sp. SB155-2 and *Sulfurihydrogenibium* suggesting recent LGT of the large flagellar regulon from *Sulfurihydrogenibium* to *Nitratiruptor*. This is consistent with the identification of a large genomic region among *Nitratiruptor* sp. SB155-2 exhibiting an atypical G+C content (Nakagawa et al., 2007). The flagellar genes were also shown to exhibit the highest degree of similarity to Aquificae (i.e., *Aquifex aeolicus*; Nakagawa et al., 2007).

3.6.4 Lipopolysaccharide (LPS) Biosynthesis

The distinctive property of cell wall architecture of most Gram-negative bacteria, including the Aquificae, Thermotogae and most ε-Proteobacteria, is the presence of the outer membrane. A prominent constituent of the outer leaflet of the outer membrane is lipopolysaccharide (LPS), which is composed of O-antigen repeats, core oligosaccharide region and the membrane-anchoring lipid A molecule. Previous ultrastructure (Cavalier-Smith 2002; Plötz et al.,2000) and phylogenetic studies (Beiko et al., 2005) suggested

that cell wall proteins tend to be inherited vertically and are informative for classification. Conversely, Boussau et al., (2008) claimed that these operational genes were likely to be of significant adaptive value and suggested that the resemblance of the outer membrane between *Aquifex* and other Proteobacteria was a result of LGT. Indeed, certain components of the LPS molecule are highly variable such as the core oligosaccharide or O-antigen repeats (Wang and Quinn 2010). However, extensive studies of the lipopolysaccharide biosynthesis (ko00540) pathways in *E. coli* and other bacteria has revealed that the structure of lipid A as well as the enzymes involved in biosynthesis of the molecule are conserved across Gram-negative bacteria (Slonim et al., 2006) and required for cell growth (Raetz et al., 2007). Among all three Aquificae, only lipid A and 3-deoxy-D-*manno*-oct-2-ulosonic acid (Kdo) synthesis were inferred to be present, whereas the presence of the O-antigen enzymes were variable.

The first stage in the biosynthesis of the LPS is the synthesis of the Kdo_2-lipid A. Nine clusters implicated in Kdo_2-lipid A synthesis among the Aquificae, eight (89%) profiles were shared exclusively with the ε-Proteobacteria while the Thermotogae and Archaea were represented only in one profile (kdsA, which is present in many Gram-positive organisms: Slonim et al., 2006). These observations are consistent with absence of LPS pathways among the Thermotogae (Nelson et al. 1999; Plötz et al., 2000) and the structural difference between the bacterial and archaeal domains (Koga and Morii 2007). Investigations of the lipid A component revealed a reduced set of six enzymes. The first three reactions, mediated by LpxA, LpxC and LpxD enzymes, synthesize UDP-diacyl-GlcN from UDP-N-acetylglucosamine (UDP-GlcNAc). The hydrolysis of UDP-diacyl-GlcN to Lipid X by the LpxH enzyme, however, was absent from the three Aquificae.

The remaining three enzymes, LpxB, LpxK and KdtA/WaaA were predicted to convert UDP-diacyl-GlcN to the minimal LPS structure Kdo_2-lipid IV_A where Kdo_2/Kdo (Mamat et al., 2009), synthesized by Kdo biosynthesis enzymes (KdsA, KdsC and KpsU), was transferred to Lipid IV_A by KdtA/WaaA (Wang and Quinn 2010). The last two enzymes in the lipid A pathway, LpxL and LpxM, were not detected in *Aquifex* and *Hydrogenobaculum* and LpxL was present in *Sulfurihydrogenibium* and other mesophilic and thermophilic ε-Proteobacteria.

Phylogenetic and bipartition analyses of the six lipid A enzymes (LpxA, LpxB, LpxC, LpxD, LpxK and KdtA) produced trees in which the mesophilic ε-Proteobacteria were adjacent to a cohesive Aquificae clan. Similarly, the KdsA, KdsC and KpsU enzymes involved in Kdo synthesis revealed that the majority of trees contained mesophilic and thermophilic ε-Proteobacteria branching with a cohesive Aquificae clan.

3.6.5 Energy Metabolism: Oxidative Phosphorylation

The mechanistically complex oxidative phosphorylation (ko00190) forms adenosine triphosphate (ATP) as a result of the transfer of electrons from NAHD or $FADH_2$ to a final electron acceptor (usually molecular oxygen; O_2) through a series of electron carriers. The flow of electrons through a sequential set of large proton-pump supercomplexes, NADH (nicotinamide adenine dinucleotide) dehydrogenase (complex I), cytochrome *c* reductase (complex III) and cytochrome *c* oxidase (complex IV) generates an electrochemical potential gradient that drives the production of ATP by F_OF_1 ATP synthase (complex V). Evidence has been accumulating over the past few decades that respiratory chains are dynamic systems that display great variability in their components. Studies have revealed that the genes involved in these essential complexes have

experienced frequent exchange across vast phylogenetic distances (Hilario and Gogarten 1993).

To investigate the respiratory chain of the Aquificae, I evaluated thirty-seven clusters: 13 in complex I, 2 in complex II (succinate dehydrogenase/fumarate reductase), 3 in complex III, 10 in complex IV and 9 in complex V. Phylogenetic profiling of the five complexes identified similar affinities of the Aquificae to Archaea (24 R-*) and ε-proteobacteria (22 E-*), and fewer genes in common with the Thermotogae (10 T-*). Indeed, the anaerobic Thermotogae lack the majority of aerobic complexes with the exception of nine genes from complex I and all of the proteins in complex V (Slonim et al., 2006).

Complex I or NADH dehydrogenase is the first entry point for electrons into the respiratory electron transport chain. Electrons from NADH are transferred through a series of electron carriers or prosthetic groups - flavin mononucleotide (FMN) and numerous iron-sulfur clusters (Fe-S) - to ultimately translocate four protons (H+) across the inner membrane. The conserved L-shape structure (Clason et al., 2010) found among the three Aquificae is comprised of 13 'core' subunits (Hirst 2010) where all gene sets excluding the prosthetic groups (NuoB, E, F and G) contained multiple Aquificae copies (NuoA, CD, H, I, J, K, L, M, N). Gene order of complex I (NuoA to NuoN) among the Aquificae deviated from the observed order in *E. coli* (Weidner et al., 1993) whereby the electron carriers NuoE, F and G subunits were located in separate regions of each Aquificae genome.

Phylogenetic analysis of each subunit revealed that most of the duplicated Aquificae copies were in-paralogs with the exception of the out-paralog NuoL2 in

Aquifex. Bipartition analysis of the contiguous subunits NuoA, B, CD, H, I, J, K, L, M, and N revealed the ε-Proteobacteria to be adjacent to a cohesive Aquificae clan among all subunits excluding NuoB. The prosthetic-containing NuoB, E, G (Fe-S) and F (FMN) subunits identified the Euryarchaeota to be adjacent to the Aquificae more often than were the Thermotogae (NuoE, F, and G) or the ε-Proteobacteria (NuoB). These observations suggest that the genes of complex I, despite the duplication events, contained a consistent ε-proteobacterial signature while the prosthetic-group containing subunits were acquired from the Archaea (e.g., the *in situ* displacement of the NuoB gene).

Complex II or succinate dehydogenase/fumarate reducatase is the second entry point into the respiration chain where electrons from $FADH_2$ reducing ubiquinone (Q) to ubiquinol (QH_2). The expression of either succinate dehydrogenase (Sdh) or fumarate reductase (Frd) enzymes depends on growing conditions; under aerobic conditions Sdh is expressed and Frd is repressed and the reverse occurs under anaerobic conditions (Cecchini 2003). Four genes of complex II were often found in a compact operon encoding two hydrophilic subunits, the catalytic A and Fe-S cluster-containing B subunits, and two small membrane-bound C and D subunits in *E. coli* (Cecchini 2003). Among the Aquificae, no gene order conservation was observed and the presence of subunit SdhB was variable: both *Aquifex* and *Hydrogenobaculum* contained the SdhA and FrdB genes whereas *Sulfurihydrogenibium* contained SdhA, B and FrdB. The hydrophilic C and D subunits were not identified in the three Aquificae. However, these subunits can be hard to detect using homology-based methods.

Phylogenetic and bipartition analysis placed the mesophilic ε-Proteobacteria adjacent to a cohesive Aquificae in the SdhA phylogeny. The SdhB/FrdB phylogeny, however, placed the five Aquificae FrdB copies (two *Aquifex*, two *Hydrogenobaculum* and one *Sulfurihydrogenibium*) adjacent to the euryarchaeotes and the *Sulfurihydrogenibium* SdhB copy branched with mesophilic ε-Proteobacteria. Assuming the annotations for Sdh/Frd genes are correct, these observations suggest that SdhB was functionally replaced in *Aquifex* and *Hydrogenobaculum* by an archaeal FrdB homolog and the ancestral SdhB homolog was simultaneously or subsequently lost.

Complex III or the cytochrome bc_1 complex catalyzes the transfer of electrons from QH_2 to the mobile electron carrier cytochrome *c* leading to the effective net transport of four protons (Berry et al., 2000). In all three Aquificae, genes encoding the highly conserved complex III subunits (Crofts 2004): the Rieske iron sulfur protein (ISP), cytochrome *b* (Cytb), and cytochrome c_1 (Cyt1) – were partitioned into three distinct clusters and genome locations. Phylogenetic and bipartition analyses identified the ε-Proteobacteria to be adjacent to the Aquificae among the IPS and Cyt1 genes, whereas in the Cytb tree the Aquificae were not cohesive, with *Sulfurihydrogenibium* adjacent to the ε-Proteobacteria while the other two Aquificae branched with a group that includes Archaea, δ-Proteobacteria and Actinobacteria.

Complex IV or cytochrome *c* oxidase is the last of the three proton-pumping assemblies in the respiratory chain, catalyzing the transfer of electrons from reduced cytochrome *c* (cytochrome *c* oxidase) or quinol (quinol oxidase) to the final acceptor (usually molecular oxygen), translocating four protons. The majority of aerobic bacteria utilize multiple oxidases enabling the organisms to customize their respiratory systems to

meet the demands of a variety of oxygen concentrations (García-Horsman et al., 1994). Indeed, the expression of different cytochrome *c* or quinol oxidases are linked with the type of heme the complex harbors. Complexes containing heme A-type which are typically found among mitochrondial aa_3-type and other A-type oxidases (García-Horsman et al., 1994) predominate when cells grow aerobically with high oxygen tension. Alternative complexes containing variants of the typical heme A such as heme B (e.g., cbb_3-type), heme O (e.g. bo_3-type quinol oxidase) or heme D (e.g., *bd*-type quinol oxidase) are expressed under microaerophilic (low O_2) conditions (García-Horsman et al., 1994; Michel et al., 1998;Heinemann et al., 2008).

The microaerophilic Aquificae contain three different cytochrome *c* oxidases: the aa_3-type present only in *Aquifex,* the minimal cbb_3-type cytochrome *c* oxidase found in both *Hydrogenobaculum* and *Sulfurihydrogenibium*, and *bd*-type quinol oxidase present in all three Aquificae. The *Aquifex* aa_3-type oxidase is composed of the three core subunits CoxI, CoxII and CoxIII (Michel et al., 1998) and the gene order was inconsistent with any of the hypothesized relatives. Two sets of adjacent gene sets were identified in *Aquifex* separated by a 550 base pair intergenic spacer: the first set with CoxIII-x-CoxII-CoxI ordering and the second set with CoxII2-CoxI2-x, where x represents the SCO1/senC gene believed to promote the assemblage of CoxI and CoxII post-translationally (Swem et al., 2005). Phylogenies of the subunits CoxI2 and CoxII2 copies identified a cohesive Aquificae clan adjacent to the Archaea whereas CoxI, II and III of the first set were adjacent to a Proteobacteria clan to the exclusion of ε-Proteobacteria. Among the heme biosynthesis genes, CyoE/CtaB preferentially branched with the Proteobacteria (α, β, and γ), whereas the COX15/CtaA gene branched with the Archaea.

The *cbb*$_3$-type found in *Hydrogenobaculum* and *Sulfurihydrogenibium* contained only subunit I which was confirmed to have a novel function among certain α-Proteobacteria (García-Horsman et al.,1994; Gray et al., 1994; Saraste 1994). The phylogenetic placements of the two Aquificae subunits CoxI were adjacent to the ε-Proteobacteria and other Proteobacteria, which tend to utilize this microaerobic oxidase (Nakagawa et al., 2007). The three Aquificae utilize another high oxygen affinity complex IV – *bd*-type quinol oxidase and has been shown to be widely distributed (e.g., Blattner et al., 1997; Heidelberg et al., 2002; Kunst et al., 1997), particularly in Gram-negative heterotrophs and are present among some Archaea (Lubben 1995). This enzyme complex is a membrane-bound heterodimer encoded by two subunits CydA and CydB, which were both found to branch with the Archaea.

Complex V or Adenosine Triphosphate (ATP) synthase is the final enzyme supercomplex in the oxidative phosphorylation pathway that synthesizes ATP generated by the downhill flow of protons, produced by complexes I, III and IV, across the inner membrane. The highly conserved and ubiquitous F_OF_1 ATP synthase is composed of five hydrophilic components of the F_1 complex (α, β, δ, ε and γ subunits) which catalyzes ATP hydrolysis/synthesis and the three transmembrane-containing a, b and c subunits of the F_O complex which acts as the proton channel (Yoshida et al., 2001). Phylogenies and bipartition analysis of the F_O complex placed the Aquificae with the mesophilic ε-Proteobacteria, while the second copy of subunit b was adjacent to a heterogeneous ε-Proteobacteria and Thermotogae clan. The F_1 components showed weak bootstrap support for the grouping of the Aquificae with the ε-Proteobacteria with the exception of the non-catalytic α subunit, which was ε-Proteobacteria-derived.

Chapter 4

Discussion

In this study I examined three Aquificae genomes: *Aquifex aeolicus* VF5, *Hydrogenobaculum* sp. Y04AAS1 and *Sulfurihydrogenibium* sp. YO3AOP1 (Reysenbach et al., 2009) and identified 527 gene sets predicted to be present in the common ancestor of the three Aquificae. Phylogenetic analyses of the broadly distributed constituents of this "core" set of proteins revealed that the majority of these genes exhibit identical branching patterns within the Aquificae to those seen in 16S rRNA phylogenies (Cole et al., 2009). The cohesion of this group was frequently observed in phylogenetic trees built from profiles distributed across both archaeal and bacterial domains (i.e., RET profiles) suggesting that the Aquificae are indeed a distinct lineage. However, genes specific to the bacterial domain (i.e., ET profiles) tended to produce less-cohesive trees, containing other lineages interleaved amongst the Aquificae, particularly thermophilic bacteria of the ε-Proteobacteria (*Nitratiruptor* sp. SB155-2), and *Nitrospira*. Furthermore, many proteins in this core set also showed numerous exclusive affiliations with the Archaea, ε-Proteobacteria, Thermotogae and other Bacteria. These results imply that LGT is, indeed, rampant even among the core gene set that contains universal and semi-universal genes including the 'informational' genes (Jain et al., 1999).

Functional categorization of this core Aquificae set identified a subset of genes involved in translation (category J; see Figure 3.4) to be the sole functional group that is primarily affiliated with the Thermotogae (Boussau et al., 2008; Nelson et al., 1999). In-depth examination of an essential component involved in translation - the ribosomal protein complex - revealed a mosaic of affiliations among the genes present among the

six major ribosomal operons. Additionally, gene organization studies revealed the consistent presence of four genes (*kad*, *map*, *IF-1* and *L36*, located between the *SecY* gene and *α* operon) in all Aquificae and Thermotogae (and other thermophilic lineages) while the adenylate kinase (*kad*) gene was absent from this region in all *ε*-proteobacterial genomes (Figure 3.6). The affiliation of the Aquificae with *ε*-Proteobacteria, however, dominated numerous functional categories (see Figure 3.4), with cell envelope/outer membrane biogenesis (M) as the most striking example. Further investigations identified the lipid A biosynthesis pathway known to be present in most Gram-negative bacteria and absent in Gram-positive bacteria (Slonim et al., 2006). Phylogenetic analyses of the constituent enzymes from this widely conserved pathway (Mamat et al. 2009; Wang and Quinn 2010) were consistent with exclusive *ε*-proteobacterial affiliations identified in previous studies (Beiko et al., 2005; Cavalier-Smith 2002; Plötz et al., 2000). Thus, the lipid A biosynthetic pathway may have been preferentially inherited among the three Aquificae for the expression of the Gram-negative trait.

If one considers these results under the assumptions of the complexity hypothesis (Jain et al., 1999) and in the context of the two scenarios depicted in Figure 1.1, my findings suggest that the Aquificae are sister to the Thermotogae (Boussau et al., 2008). This supposition is largely based on the analyses of a particular subset of the informational genes - the ribosomal protein complex (Figure 1.1a). However, recent reevaluations of the complexity hypothesis (Hao and Golding 2008) and the numerous studies identifying LGT among informational genes (Gogarten et al., 2002; Kanhere and Vingron 2009) suggest that the functional distinction between informational and operational genes is of limited utility as a predictive tool for identifying transferred genes.

Indeed, individual gene phylogenies within the six major ribosomal operons show alternative affiliations to Archaea or ε-Proteobacteria (see Figure 3.5), and transcriptional genes (COG class K; see Figure 3.5) preferentially show ε-proteobacterial affiliations (Gruber and Bryant 1998; Klenk et al., 1999). Thus, another plausible interpretation of the data presented herein is that the evolution of different gene sets reflects the lifestyle of the organisms in which they reside—in this case thermophily or mesophily—rather than their functional category.

The Aquificae contain a significant fraction of genes that were potentially acquired from or donated to other thermophilic lineages, establishing a plausible connection between the similarity in lifestyle of evolutionarily distant organisms and the apparent rate of LGT (Aravind et al., 1998; Beiko et al., 2008; Nelson et al., 1999). Indeed, many Aquificae genes, particularly those with metabolic functions, are related to the Archaea and were likely acquisitions enabling differing strategies for ecological adaptation, such as the *bd*-type quinol oxidase, which may confer adaptation to microaerophilic conditions. The affiliations to other thermophilic lineages, particularly the Thermotogae, *Nitratiruptor* sp. SB155-2 and *Nitrospira*, however, may have been acquisitions among these bacterial lineages simply due to their proximity with other thermophiles in the environment. Moreover, bacterial lineages that were initially mesophilic and later colonized a hot environment were shown to have widespread amino acid biases (i.e., a significant increase in charged residues) in their proteome (Berezovsky and Shakhnovich 2005; Boucher et al., 2003; Singleton and Amelunxen 1973). Thus, the acquisitions of specific gene sets from other thermophilic bacteria (e.g., the ribosomal complex and flagellar assembly genes) may be of selective advantage to these organisms,

conferring thermal stabilization of these important protein complexes. This is in contrast to the cell membrane in which structural differences, such as the increase in saturated and branch-chained fatty acids, i.e., branched glycerol dialkyl diethers, has been proposed to contribute to the thermal stability of the membrane (Boucher et al., 2003).

Given the patterns of phylogenetic relatedness seen in the subsystems that were investigated, it appears that the majority of genes in the Aquificae that appear related to thermophiles were likely lateral acquisitions, whereas those with ε-proteobacterial affiliations may be remnants of a mesophilic past that predated its colonization of a thermophilic environment (Figure 1.1b). Environmental studies of hydrothermal adaptations have revealed that the ε-Proteobacteria have developed diverse strategies to colonize many deep-sea substrates, due in part, to their high growth rates, rapid adaptations to changing geochemical conditions and metabolic versatility (Lopez-Garcia et al., 2003; Nakagawa et al., 2007). Thus, the ε-Proteobacteria may be major contributors in the colonization processes where they play a vital role in the cycling of carbon, nitrogen and sulfur (Alain et al., 2004; Campbell et al., 2006). Under this scenario, one could speculate that the Aquificae, as a derived ε-Proteobacterium, could acquire essential genes to adapt to a thermophilic environment, explaining the observed mosaic patterns present among the Aquificae. Indeed, there is convincing phylogenetic and physiological (e.g., cell wall biosynthesis; Cavalier-Smith 2002) evidence to suggest that the Aquificae are highly modified ε-Proteobacteria and, therefore, the frequent positioning with the Thermotogae at the base of the tree a consequence of preferential LGT and potentially a pull in the direction of the Archaea as well. However, given the rate of LGT there may not be enough consistent information to draw any definitive

conclusions. Therefore, further insights will likely be gained with deeper genomic sampling of the diversity within phylum Aquificae.

Chapter 5

Conclusion

Despite the ever-increasing number of complete prokaryotic genomes available, the placement of the Aquificae remains surprisingly unchanged. Indeed, in the face of LGT, an undeniable source of genome evolution and innovation, evolutionary biologists search for genes apparently unhindered by LGT, which drastically reduces the set of potential genes used to infer the prokaryotic TOL (Dagan and Martin 2006). The narrowed dataset includes such universal genes as SSU rRNA and other essential genes predicted to be recalcitrant to LGT (i.e., 'informational' genes like transcriptional and translational genes; Jain et al., 1999). Phylogenies based on the SSU rRNA and multi-locus datasets of these pre-selected genes provide the current picture of the prokaryotic TOL. However, numerous studies have shown that even informational genes are not exempt from transfers (e.g., Brochier et al., 2000), particular among microorganisms adapted to a hot lifestyle (Aravind et al., 1998; Nelson et al., 1999). Evidently, the Aquificae are among such organisms adapted to thermophilic environments with complex evolutionary histories impacted by a large influx of genes. Nonetheless, the work presented herein provides insight into the evolution of the Aquificae phylum and compares these findings against conclusions presented in the literature (e.g., Boussau et al., 2008, Beiko et al., 2005; Cavalier-Smith 2002).

I have developed and implemented a phylogenomic pipeline using up-to-date and sophisticated algorithms (e.g., FSA, Hmmer3 and RAxML). This pipeline first identifies all the evolutionary patterns present in the three Aquificae and frames the phylogenetic affiliations of the functional groups and pathways in light of previously published

hypotheses in hopes to identify the true cellular history of this phylum. In-depth analyses using phylogenetic and phylogenomic methods (e.g., phylogentic profiling and gene order investigations) determined the branching patterns of the Aquificae and help delineate patterns of vertical inheritance versus lateral acquisition.

Interestingly, the most prevalent phylogenetic patterns among the variable and core Aquificae clusters was to the ε-Proteobacteria, where the Thermotogae ranked second among the core set due in part to the large number of translational genes. Indeed, clusters involved in translation, ribosome structure and biosynthesis (J) were the sole functional group with significant Thermotogae signal, whereas numerous functional groups were affiliated with ε-Proteobacteria (e.g., LPS biosynthesis) and Archaea (mostly the metabolic pathways including complexes of the oxidative phosphorylation). Based on the complexity hypothesis (Jain et al., 1999) the Aquificae were proposed to be sister to the Thermotogae (Boussau et al., 2008). Analysis of the gene order and gene phylogenies of the ribosomal operons suggests that the majority of the ribosomal complex may have been acquired from Thermotogae or that systematic biases (e.g., compositional or heterotachy) may have resulted in artificial attraction between the Aquificae and the Thermotogae. Thus, suggesting that the hyperthermophilic and thermophilic affiliations of the Aquificae were acquisitions among this group to confer thermophilic adaptation and the ancestral signal, weak among *Aquifex* and *Hydrogenobaculum*, is to the ε-Proteobacteria.

In summary, my work has provided insight into the evolution of the Aquificae phylum. I have shed light on the multiple affiliations of this phylum, identified functional groups/pathways containing each of these affiliations and identified which relationship is

likely ancestral. Although many important questions surrounding the evolution of the Aquificae and other organisms adapted to hot environments remain, I have added to the body of knowledge surrounding the enigmatic ancestry of the Aquificae and developed methods to help delineate ancestry even when the rates of LGT are high.

References

Alain, K., M. Zbinden, N. L. Bris, F. Lesongeur, J. Quérellou, and F. Gaill. 2004. Early steps in microbial colonization processes at deep-sea hydrothermal vents. Environmental Microbiology 6: 227-241.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. Journal of Molecular Biology 215(3): 403–410.

Andrieu, G., G. Caraux, and O. Gascuel. 1997. Confidence intervals of evolutionary distances between sequences and comparison with usual approaches including the bootstrap method. Molecular Biology and Evolution 14(8): 875-82.

Aravind, L., R. L. Tatusov, Y. I. Wolf, D. R. Walker, and E. V. Koonin. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends in Genetics 14(11): 442-4.

Asai, T., D. Zaporojets, C. Squires, and C. L. Squires. 1999. An Escherichia coli strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. Proceedings of the National Academy of Sciences of the United States of America 96(5): 1971-1976.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, et al. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25(1): 25-29.

Baldauf, S. L., J. D. Palmer, and W. F. Doolittle. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proceedings of the National Academy of Sciences of the United States of America 93(15): 7749-7754.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A Kingdom-Level Phylogeny of Eukaryotes Based on Combined Protein Data. Science 290(5493): 972-977.

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

Baum, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41(1): 3–10.

Beiko, R. G. 2011. Telling the Whole Story in a 10,000-Genome World. Biology Direct 6(1): 34.

Beiko, R. G., W. F. Doolittle, and R. L. Charlebois. 2008. The impact of reticulate evolution on genome phylogeny. Systematic Biology 57(6): 844-56.

Beiko, R. G., T. J. Harlow, and M. A. Ragan. 2005. Highways of gene sharing in prokaryotes. Proceedings of the National Academy of Sciences of the United States of America 102(40): 14332-7.

Berezovsky, I. N., and E. I. Shakhnovich. 2005. Physics and evolution of thermophilic adaptation. Proceedings of the National Academy of Sciences of the United States of America 102(36): 12742-7.

Berry, E. A., M. Guergova-Kuras, L. Huang, and A. R. Crofts. 2000. Structure and function of cytochrome bc complexes. Annual Review of Biochemistry 69(1): 1005–1075.

Bininda-Emonds, O. R. P. 1998. Properties of matrix representation with parsimony analyses. Systematic Biology 47(3): 497-509.

Bininda-Emonds, O. R. P., J. L. Gittleman, and M. A. Steel. 2002. The (Super)Tree of Life: Procedures, Problems, and Prospects. Annual Review of Ecology and Systematics 33(1): 265-289.

Bininda-Emonds, O.R.P. 2004. The evolution of supertrees. Trends in Ecology & Evolution 19(6): 315-22.

Blattner, F. R., G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, et al. 1997. The Complete Genome Sequence of Escherichia coli K-12. Science 277(5331): 1453-1462.

Bocchetta, M., S. Gribaldo, A. Sanangelantoni, and P. Cammarano. 2000. Phylogenetic depth of the bacterial genera Aquifex and Thermotoga inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. Journal of Molecular Evolution 50(4): 366–380.

Boucher, Y., C. J. Douady, R. T. Papke, D. A. Walsh, M. R. Boudreau, C. L. Nesbø, et al. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annual Review of Genetics 37: 283-328.

Boussau, B., L. Guéguen, and M. Gouy. 2008. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. BMC Evolutionary Biology 8: 272.

Bradley, R. K., A. Roberts, M. Smoot, S. Juvekar, J. Do, C. Dewey, et al. 2009. Fast statistical alignment. PLoS Computational Biology 5(5): e1000392.

Brochier, C., S. Gribaldo, Y. Zivanovic, F. Confalonieri, and P. Forterre. 2005. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? Genome Biology 6(5): R42.

Brochier, C., H. Philippe, and D. Moreira. 2000. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends in Genetics 16(12): 529-33.

Bruen, T. C., H. Philippe, and D. Bryant. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics 172(4): 2665-2681.

Campbell, B. J., A. S. Engel, M. L. Porter, and K. Takai. 2006. The versatile epsilon-proteobacteria: key players in sulphidic habitats. Nature Reviews. Microbiology 4(6): 458-68.

Carrillo, H., and D. Lipman. 1988. The multiple sequence alignment problem in biology. SIAM Journal on Applied Mathematics: 1073–1082.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17(4): 540-52.

Cavalier-Smith, T. 2002. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. International Journal of Systematic and Evolutionary Microbiology 52(Pt 1): 7-76.

Cavalli-Sforza, L. L., and A. W. Edwards. 1967. Phylogenetic analysis. Models and estimation procedures. American Journal of Human Genetics 19(3 Pt 1): 233-57.

Cecchini, G. 2003. Function and structure of complex II of the respiratory chain. Annual Review of Biochemistry 72: 77-109.

Chan, C. X., R. G. Beiko, and M. A. Ragan. 2007. A two-phase strategy for detecting recombination in nucleotide sequences. Molecular Biology: 1-7.

Chilcott, G. S., and K. T. Hughes. 2000. Coupling of flagellar gene expression to flagellar assembly in Salmonella enterica serovar typhimurium and Escherichia coli. Microbiology and Molecular Biology Reviews 64(4): 694-708.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science (New York, N.Y.) 311(5765): 1283-7.

Clarke, G. D., R. G. Beiko, M. A. Ragan, and R. L. Charlebois. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. Journal of Bacteriology 184(8): 2072.

Clason, T., T. Ruiz, H. Schägger, G. Peng, V. Zickermann, U. Brandt, et al. 2010. The structure of eukaryotic and prokaryotic complex I. Journal of Structural Biology 169(1): 81-8.

Coenye, T., and P. Vandamme. 2005. Organisation of the S10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes. FEMS Microbiology Letters 242(1): 117-26.

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, et al. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research 37(Database issue): D141-D145.

Crofts, A. R. 2004. The cytochrome bc1 complex: function in the context of structure. Annual Review of Physiology 66: 689-733.

Dagan, T., M. Roettger, D. Bryant, and W. Martin. 2010. Genome networks root the tree of life between prokaryotic domains. Genome Biology and Evolution 2: 379-92.

Dagan, T., and W. Martin. 2006. The tree of one percent. Genome Biology 7(10): 118.

Darwin, C. 1859. 6 Screen 22-79 The Origin of Species. ed. John Murray. John Murray.

Date, S. V., and E. M. Marcotte. 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nature Biotechnology 21(9): 1055-62.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. In In Atlas of protein sequence and structure, Citeseer.

DeBry, R. W. 1992. The consistency of several phylogeny-inference methods under varying evolutionary rates. Molecular Biology and Evolution 9(3): 537-51.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews: Genetics 6(5): 361-75.

Doolittle, W. F. 1999. Phylogenetic Classification and the Universal Tree. Science 284(5423): 2124-2128.

Doolittle, W. F., Y. Boucher, C. L. Nesbø, C. J. Douady, J. O. Andersson, and A. J. Roger. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philosophical transactions of the Royal Society of London. Series B, Biological sciences 358(1429): 39-57; discussion 57-8.

Doolittle, W. F., and O. Zhaxybayeva. 2007. Evolution: reducible complexity -- the case for bacterial flagella. Current Biology 17(13): R510-R512.

Doolittle, W. F., and O. Zhaxybayeva. 2009. On the origin of prokaryotic species. Genome Research 19(5): 744-56.

Eddy, S. R. 2011. Accelerated Profile HMM Searches. ed. William R. Pearson. PLoS Computational Biology 7(10): e1002195.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32(5): 1792-7.

Edwards, A. W. F., and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. eds. Vernon Hilton Heywood and J McNeill. Phenetic and Phylogenetic Classification 6(6): 67-76.

Efron, B., and R. J. Tibshirani. 1993. 57 Chapman and Hall 436 An Introduction to the Bootstrap. ed. Chapman HallCRC. Chapman & Hall.

Eisen, J., and C. M. Fraser. 2003. Phylogenomics: intersection of evolution and genomics. Science (New York, N.Y.) 300(5626): 1706-7.

Farris, J. S., A. G. Kluge, and M. J. Eckardt. 1970. A numerical approach to phylogenetic systematics. Systematic Biology 19(2): 172.

Felsenstein, J. 1989. PHYLIP -- Phylogeny Inference Package (Version 3.2). Cladistics 5(164-166): 164-166.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39(4): 783-791.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17(6): 368-376.

Felsenstein, J. 2004. 266 Methods in Enzymology 418-27 Inferring Phylogenies. ed. Joseph Felsenstein. Sinauer Associates.

Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Zoology 22(3): 240-249.

Feng, D F, and R F Doolittle. 1990. Progressive alignment and phylogenetic tree construction of protein sequences. Methods in Enzymology 183: 375-87.

Fitch, W. M. 1970. Distinguishing Homologous from Analogous Proteins. Systematic Zoology 19(2): 99-113.

Fitch, W. M., and E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochemical Genetics 4(5): 579-593.

Fitz-Gibbon, S. T., and C. H. House. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Research 27(21): 4218-22.

Frost, L. S., R. Leplae, A. O. Summers, and A. Toussaint. 2005. Mobile genetic elements: the agents of open source evolution. Nature Reviews: Microbiology 3(9): 722-32.

Gaasterland, T., and M. A. Ragan. 1998. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. Microbial and Comparative Genomics 3(4): 199–217.

García-Horsman, J. A., B. Barquera, J. Rumbley, J. Ma, and R. B. Gennis. 1994. MINIREVIEW The Superfamily of Heme-Copper Respiratory Oxidases. Journal of Bacteriology 176(18): 5587-5600.

Gogarten, J . P., W. F. Doolittle, and J. G. Lawrence. 2002. Prokaryotic evolution in light of gene transfer. Molecular Biology and Evolution 19(12): 2226-38.

Gophna, U., R. L. Charlebois, and W. F. Doolittle. 2006. Ancient lateral gene transfer in the evolution of Bdellovibrio bacteriovorus. Trends in Microbiology 14(2): 64-69.

Gophna, U., W. F. Doolittle, and R. L. Charlebois. 2005. Weighted genome trees: refinements and applications. Journal of Bacteriology 187(4): 1305.

Gotoh, O. 1996. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. Journal of Molecular Biology 264(4): 823-838.

Gray, K. A., M. Grooms, H. Myllykallio, C. Moomaw, C. Slaughter, and F. Daldal. 1994. Rhodobacter capsulatus contains a novel cb-type cytochrome c oxidase without a CuA center. Biochemistry 33(10): 3120-3127.

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Systematic Biology 47(1): 9.

Griffiths, E., and R. S. Gupta. 2004. Signature sequences in diverse proteins provide evidence for the late divergence of the Order Aquificales. International Microbiology 7(1): 41-52.

Gruber, T. M., and D. A. Bryant. 1998. Characterization of the group 1 and group 2 sigma factors of the green sulfur bacterium Chlorobium tepidum and the green non-sulfur bacterium Chloroflexus aurantiacus. Archives of Microbiology 170(4): 285-96.

Hall, R. M., C. M. Collis, M. J. Kim, S. R. Partridge, G. D. Recchia, and H. W. Stokes. 1999. Mobile gene cassettes and integrons in evolution. Annals of the New York Academy of Sciences 870: 68-80.

Hao, W., and G. B. Golding. 2008. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. BMC Genomics 9: 235.

Harshey, R. M. 2003. Bacterial motility on a surface: many ways to a common goal. Annual review of microbiology 57: 249-73.

Hasegawa, M., and T. Hashimoto. 1993. Ribosomal RNA trees misleading? Nature 361(6407): 23.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1): 97.

Heidelberg, J. F., I. T. Paulsen, K. E. Nelson, E. J. Gaidos, W. C. Nelson, T. D. Read, et al. 2002. Genome sequence of the dissimilatory metal ion-reducing bacterium Shewanella oneidensis. Nature Biotechnology 20(11): 1118-23.

Heinemann, I. U., M. Jahn, and D. Jahn. 2008. The biochemistry of heme biosynthesis. Archives of Biochemistry and Biophysics 474(2): 238-51.

Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America 89(22): 10915-9.

Hilario, E., and J. P. Gogarten. 1993. Horizontal transfer of ATPase genes--the tree of life becomes a net of life. Bio Systems 31(2-3): 111-119.

Hirst, J. 2010. Towards the molecular mechanism of respiratory complex I. The Biochemical Journal 425(2): 327-39.

Hogeweg, P., and B. Hesper. 1984. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. Journal of Molecular Evolution 20(2): 175-186.

Holder, M., and P. O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nature Reviews: Genetics 4(4): 275-84.

Huber, R., W. Eder, S. Heldwein, G. Wanner, H. Huber, R. Rachel, and K. O. Stetter. 1998. Thermocrinis ruber gen. nov., sp. nov., a Pink-Filament-Forming Hyperthermophilic Bacterium Isolated from Yellowstone National Park. Applied and Environmental Microbiology 64(10): 3576-3583.

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Systematic biology 51(5): 673-88.

Huelsenbeck, J. P., B. Rannala, and Masly J. P. 2000. Accommodating Phylogenetic Uncertainty in Evolutionary Studies. Science 288(5475): 2349-2350.

Huson, D. H., and M. Steel. 2004. Phylogenetic trees based on gene content. Bioinformatics 20(13): 2044-2049.

Huynen, M. A., and P. Bork. 1998. Measuring genome evolution. Proceedings of the National Academy of Sciences of the United States of America 95(11): 5849-56.

Iwasaki, W., and T. Takagi. 2009. Rapid Pathway Evolution Facilitated by Horizontal Gene Transfers across Prokaryotic Lineages. ed. Ivan Matic. PLoS Genetics 5(3): 8.

Iyer, L. M., E. V. Koonin, and L. Aravind. 2004. Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. Gene 335: 73-88.

Jacob, F., and J. Monod. 1961. Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology 3(3): 318-356.

Jain, R., M. C. Rivera, and J . A. Lake. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proceedings of the National Academy of Sciences of the United States of America 96(7): 3801-6.

Jakobsen, I. B., and S. Easteal. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Computer Applications in the Biosciences: CABIOS 12(4): 291-295.

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends in Genetics 22(4): 225-31.

Kanehisa, M., and S. Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28(1): 27-30.

Kanhere, A., and M. Vingron. 2009. Horizontal Gene Transfers in prokaryotes show differential preferences for metabolic and translational genes. BMC Evolutionary Biology 9: 9.

Kimura, M. 1991. The neutral theory of molecular evolution: a review of recent evidence. Japan Journal of Genetics 66(4): 367-386.

Klenk, H. P., T. D. Meier, P. Durovic, V. Schwass, F. Lottspeich, P. P. Dennis, and W. Zillig. 1999. RNA polymerase of Aquifex pyrophilus: implications for the evolution of the bacterial rpoBC operon and extremely thermophilic bacteria. Journal of molecular evolution 48(5): 528-41.

Koga, Y., and H. Morii. 2007. Biosynthesis of ether-type polar lipids in archaea and evolutionary considerations. Microbiology and Molecular Biology Reviews 71(1): 97-120.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431(7011): 980-984.

Korbel, J. O., B. Snel, M. A. Huynen, and P. Bork. 2002. SHOT: a web server for the construction of genome phylogenies. Trends in Genetics 18(3): 158-62.

Kosaraju, R. S. 1978. Unpublished.

Koski, L. B., and G. B. Golding. 2001. The closest BLAST hit is often not the nearest neighbor. Journal of Molecular Evolution 52(6): 540-2.

Kranz, R. G., and B. S. Goldman. 1998. Evolution and horizontal transfer of an entire biosynthetic pathway for cytochrome c biogeneis: Helicobacter, Deinococcus, Archae and more. Molecular Microbiology 25(6): 1177-1184.

Kristensen, D. M., Y. I. Wolf, A. R. Mushegian, and E. V. Koonin. 2011. Computational methods for Gene Orthology inference. Briefings in Bioinformatics 12(5): 379-391.

Kunisawa, T. 2006. Dichotomy of major bacterial phyla inferred from gene arrangement comparisons. Journal of Theoretical Biology 239(3): 367-75.

Kunisawa, T. 2001. Gene arrangements and phylogeny in the class Proteobacteria. Journal of theoretical biology 213(1): 9-19.

Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, et al. 1997. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 390(6657): 249-256.

Kupczok, A., H. A. Schmidt, and A. von Haeseler. 2010. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Algorithms for Molecular Biology 5(1): 37.

Lake, J. A., and M. C. Rivera. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. Molecular Biology and Evolution 21(4): 681-690.

Lapointe, F., P. Lopez, Y. Boucher, J. Koenig, and E. Bapteste. 2010. Clanistics: a multi-level perspective for harvesting unrooted gene trees. Trends in Microbiology 18(8): 341-7.

Larget, B., D. L. Simon, J. B. Kadane, and D. Sweet. 2005. A bayesian analysis of metazoan mitochondrial genome arrangements. Molecular Biology and Evolution 22(3): 486-95.

Lawrence, J. G. 1999. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Current Opinion in Genetics and Development 9(6): 642-8.

Lewis, P. O. 1998. Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. In Molecular Systematics of Plants, eds. D E Soltis, P S Soltis, and J J Doyle. Kluwer Academic Publishsers, p. 132-163.

Lindahl, L., and J. M. Zengel. 1986. Ribosomal genes in Escherichia coli. Annual Review of Genetics 20: 297-326.

Lipman, D. J., and W. R. Pearson. 1985. Rapid and sensitive protein similarity searches. Science 227(4693): 1435.

Liu, R., and H. Ochman. 2007. Stepwise formation of the bacterial flagellar system. Proceedings of the National Academy of Sciences of the United States of America 104(17): 7116-21.

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

Loomis, W. F., and D. W. Smith. 1990. Molecular phylogeny of Dictyostelium discoideum by protein sequence comparison. Proceedings of the National Academy of Sciences of the United States of America 87(23): 9093-7.

Lopez-Garcia, P., S. Duperron, P. Philippot, J. Foriel, J. Susini, and D. Moreira. 2003. Bacterial diversity in hydrothermal sediment and epsilonproteobacterial dominance in experimental microcolonizers at the Mid-Atlantic Ridge. Environmental Microbiology 5(10): 961-976.

Lubben, M. 1995. Cytochromes of archaeal electron transfer chains. Biochimica et Biophysica Acta. Bioenergetics 1229(1): 1–22.

Ludwig, W., and H. P. Klenk. 2005. Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics. Museum: 49–66.

Lutzoni, F., M. Pagel, and V. Reeb. 2001. Major fungal lineages are derived from lichen symbiotic ancestors. Nature 411(6840): 937-940.

L'Haridon, S., A. Reysenbach, B. J. Tindall, P. Schonheit, A. Banta, U. Johnsen, et al. 2006. Desulfurobacterium atlanticum sp. nov., Desulfurobacterium pacificum sp. nov. and Thermovibrio guaymasensis sp. nov., three thermophilic members of the Desulfurobacteriaceae fam. nov., a deep branching lineage within the Bacteria. International Journal of Systematic and Evolutionary Microbiology 56(Pt 12): 2843-52.

Mamat, U., H. Schmidt, E. Munoz, B. Lindner, K. Fukase, A. Hanuszkiewicz, et al. 2009. WaaA of the hyperthermophilic bacterium Aquifex aeolicus is a monofunctional 3-

deoxy-D-manno-oct-2-ulosonic acid transferase involved in lipopolysaccharide biosynthesis. The Journal of Biological Chemistry 284(33): 22248-62.

Matte-Tailliez, O., C. Brochier, P. Forterre, and H. Philippe. 2002. Archaeal phylogeny based on ribosomal proteins. Molecular Biology and Evolution 19(5): 631-9.

Von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. 2003. STRING: a database of predicted functional associations between proteins. Nucleic Acids Research 31(1): 258-261.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller. 1953. Equations of state calculations by fast computing machines. Journal of Chemical Physics.

Michel, H., J. Behr, A. Harrenga, and A. Kannt. 1998. Cytochrome c oxidase: structure and spectroscopy. Annual Review of Biophysics and Biomolecular Structure 27: 329-56.

Minin, V. N., K. S. Dorman, F. Fang, and M. A. Suchard. 2005. Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics 21(13): 3034-3042.

Misof, B., and K. Misof. 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Systematic Biology 58(1): 21-34.

Moret, B. M. E., J. Tang, and T. Warnow. 2005. Reconstructing phylogenies from gene-content and gene-order data. In Mathematics of Evolution and Phylogeny, ed. Olivier Gascuel. Oxford University Press, p. 321-352.

Mushegian, A. R., and E. V. Koonin. 1996. Gene order is not conserved in bacterial evolution. Trends in Genetics 12(8): 289-290.

Nakagawa, S., K. Takai, F. Inagaki, H. Hirayama, T. Nunoura, K. Horikoshi, and Y. Sako. 2005. Distribution, phylogenetic diversity and physiological characteristics of epsilon-Proteobacteria in a deep-sea hydrothermal field. Environmental Microbiology 7(10): 1619-32.

Nakagawa, S., Y. Takaki, S. Shimamura, A. Reysenbach, and K. Takai. 2007. Deep-sea vent e-proteobacterial genomes provide insight into emergence of pathogens. Proceedings of the National Academy of Sciences 104(29): 12146-12150.

Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3): 443-53.

Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, et al. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399(6734): 323-9.

Ng, S. Y. M., B. Chaban, and K. F. Jarrell. 2006. Archaeal flagella, bacterial flagella and type IV pili: a comparison of genes and posttranslational modifications. Journal of Molecular Microbiology and Biotechnology 11(3-5): 167-91.

Ogden, T. H., and M. S. Rosenberg. 2006. Multiple sequence alignment accuracy and phylogenetic inference. Systematic Biology 55(2): 314-28.

Omelchenko, M. V., K. S. Makarova, Y. I. Wolf, I. B. Rogozin, and E. V. Koonin. 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biology 4(9): R55.

Pei, J. 2008. Multiple protein sequence alignment. Current Opinion in Structural Biology 18(3): 382-6.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences of the United States of America 96(8): 4285-8.

Penn, O., E. Privman, G. Landan, D. Graur, and T. Pupko. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. Molecular BIology and Evolution 27(8): 1759-67.

Penny, D., B. J. McComish, M. A. Charleston, and M. D. Hendy. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. Journal of Molecular Evolution 53(6): 711-723.

Philippe, H, and A Germot. 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. Molecular Biology and Evolution 17(5): 830-834.

Phillips, A., D. Janies, and W. Wheeler. 2000. Multiple sequence alignment in phylogenetic analysis. Molecular Phylogenetics and Evolution 16(3): 317-30.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Molecular Biology and Evolution 21(7): 1455-8.

Plötz, B. M., B. Lindner, K. O. Stetter, and O. Holst. 2000. Characterization of a novel lipid A containing D-galacturonic acid that replaces phosphate residues. The structure of the lipid a of the lipopolysaccharide from the hyperthermophilic bacterium Aquifex pyrophilus. The Journal of Biological Chemistry 275(15): 11222-11228.

Raetz, C. R. H., C. M. Reynolds, M. S. Trent, and R. E. Bishop. 2007. Lipid A modification systems in gram-negative bacteria. Annual Review of Biochemistry 76: 295-329.

Ragan, M. A. 2001. On surrogate methods for detecting lateral gene transfer. FEMS Microbiology Letters 201(2): 187-191.

Ragan, M. A. 1992. Phylogenetic inference based on matrix representation of trees. Molecular Phylogenetics and Evolution 1(1): 53-58.

Ragan, M. A., T. J. Harlow, and R. G. Beiko. 2006. Do different surrogate methods detect lateral genetic transfer events of different relative ages? Trends in Microbiology 14(1): 4-8.

Ragan, M. A., and R. G. Beiko. 2009. Lateral genetic transfer: open issues. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 364(1527): 2241-51.

Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Systematic biology 51(5): 754-60.

Rannala, B., J. P. Huelsenbeck, Z. Yang, and R. Nielsen. 1998. Taxon sampling and the accuracy of large phylogenies. Systematic Biology 47(4): 702-10.

Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution 43(3): 304-11.

Reysenbach, A., N. Hamamura, M. Podar, E. Griffiths, S. Ferreira, R. Hochstein, et al. 2009. Complete and draft genome sequences of six members of the Aquificales. Journal of Bacteriology 191(6): 1992-3.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425(6960): 798-804.

Rokas, A., and P. W. Pw Holland. 2000. Rare genomic changes as a tool for phylogenetics. Trends in Ecology and Evolution 15(11): 454-459.

Rosenberg, M. S. 2005. Multiple sequence alignment accuracy and evolutionary distance estimation. BMC Bioinformatics 6: 278.

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

Saraste, M. 1994. Structure and evolution of cytochrome oxidase. Antonie van Leeuwenhoek 65(4): 285-7.

Schütz, M., M. Brugna, E. Lebrun, F. Baymann, R. Huber, K. O. Stetter, et al. 2000. Early evolution of cytochrome bc complexes. Journal of Molecular Biology 300(4): 663-75.

Siddall, M. E., and A. G. Kluge. 1997. Probabilism and Phylogenetic Inference. Cladistics 13(4): 313-336.

Singleton, R., and R. E. Amelunxen. 1973. Proteins from thermophilic microorganisms. Bacteriological Reviews 37(3): 320-42.

Slonim, N., O. Elemento, and S. Tavazoie. 2006. Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. Molecular Systems Biology 2: 2006.0005.

Smith, J. M. 1992. Analyzing the mosaic structure of genes. Journal of Molecular Evolution 34(2): 126-129.

Sneath, P. H. A., and R. R. Sokal. 1973. San Francisco 573 Numerical Taxonomy: The Principles and Practice of Numerical Classification. eds. D Kennedy and R B Park. W.H. Freeman.

Snel, B., P. Bork, and M. A. Huynen. 1999. Genome phylogeny based on gene content. Nature Genetics 21(1): 108-110.

Snel, B., M. A. Huynen, and B. E. Dutilh. 2005. Genome trees and the nature of genome evolution. Annual Review of Microbiology 59: 191-209.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21): 2688-2690.

Stamatakis, A., T. Ludwig, and H. Meier. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21(4): 456-63.

Steel, M. 2002. Some statistical apects of the maximum parsimony method. In Molecular Systematics and Evolution Theory and Practice, Birkhaüser Verlag, p. 125-139.

Steel, M. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. Journal of Classification 9(1): 91–116.

Steel, M., and D. Penny. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Molecular Biology and Evolution 17(6): 839-50.

Steel, M., and A. Rodrigo. 2008. Maximum likelihood supertrees. Systematic Biology 57(2): 243-50.

Stohr, R., A. Waberski, H. Völker, B. J. Tindall, and M. Thomm. 2001. Hydrogenothermus marinus gen. nov., sp. nov., a novel thermophilic hydrogen-oxidizing bacterium, recognition of Calderobacterium hydrogenophilum as a member of the genus Hydrogenobacter and proposal of the reclassification of Hydrogenobacter acidophilus a. International Journal of Systematic and Evolutionary Microbiology 51(Pt 5): 1853-1862.

Susko, E., Y. Inagaki, and A. J. Roger. 2004. On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. Molecular biology and evolution 21(9): 1629-42.

Susko, E., and A. J. Roger. 2007. On reduced amino acid alphabets for phylogenetic inference. Molecular Biology and Evolution 24(9): 2139-50.

Swem, D. L., L. R. Swem, A. Setterdahl, and Bauer C. E. 2005. Involvement of SenC in assembly of cytochrome c oxidase in Rhodobacter capsulatus. Journal of Bacteriology 187(23): 8081-8087.

Swofford, D. L. 2003. PAUP*: phylogenetic analysis using parsimony, version 4.0b10. 21 Libro: 11pp.

Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Systematic Biology 50(4): 525-39.

Talavera, G., and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 56(4): 564-77.

Tarjan, R. 1972. Depth-First Search and Linear Graph Algorithms. SIAM Journal on Computing 1(2): 146-160.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A Genomic Perspective on Protein Families. Science 278(5338): 631-637.

Thomas, C. M., and K. M. Nielsen. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature reviews: Microbiology 3(9): 711-21.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22(22): 4673-80.

Thompson, J. D., P. Koehl, R. Ripp, and O. Poch. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins 61(1): 127-36.

Wang, L., T. Warnow, B. M. E. Moret, R. K. Jansen, and L. A. Raubeson. 2006. Distance-based genome rearrangement phylogeny. Journal of Molecular Evolution 63(4): 473-83.

Wang, X., and P. J. Quinn. 2010. Lipopolysaccharide: Biosynthetic pathway and structure modification. Progress in Lipid Research 49(2): 97-107.

Watanabe, H., H. Mori, T. Itoh, and T. Gojobori. 1997. Genome plasticity as a paradigm of eubacteria evolution. Journal of Molecular Evolution 44 Suppl 1(suppl. 1): S57-64.

Weidner, U., S. Geier, A. Ptock, T. Friedrich, H. Leif, and H. Weiss. 1993. The gene locus of the proton-translocating NADH: ubiquinone oxidoreductase in Escherichia coli. Organization of the 14 genes and relationship between the derived proteins and subunits of mitochondrial complex I. Journal of Molecular Biology 233(1): 109-122.

Woese, C. R., L. Achenbach, P. Rouviere, and L. Mandelco. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts. Systematic and Applied Microbiology 14(4): 364-371.

Woese, C. R., and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America 74(11): 5088-90.

Wolf, Y. I., I. B. Rogozin, N. V. Grishin, R. L. Tatusov, and E. V. Koonin. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evolutionary Biology 1: 8.

Wolf, Y. I., I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin. 2001. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Research 11(3): 356-372.

Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462(7276): 1056-60.

Wu, M., and J. A. Eisen. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biology 9(10): R151.

Yanai, I., and C. DeLisi. 2002. The society of genes: networks of functional links between genes from comparative genomics. Genome biology 3(11): research0064.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. Journal of Molecular Evolution 42(5): 587–596.

Yoshida, M., E. Muneyuki, and T. Hisabori. 2001. ATP synthase--a marvellous rotary engine of the cell. Nature reviews: Molecular Cell Biology 2(9): 669-77.

Zhaxybayeva, O., K. S. Swithers, P. Lapierre, G. P. Fournier, D. M. Bickhart, R. T. DeBoy, et al. 2009. On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proceedings of the National Academy of Sciences of the United States of America 106(14): 5865-70.

Zuckerkandl, E., and L. Pauling. 1965a. Evolutionary divergence and convergence in proteins. Evolving genes and proteins 97: 166.

Zuckerkandl, E., and L. Pauling. 1965b. Molecules as documents of evolutionary history. Journal of Theoretical Biology 8(2): 357–366.

Appendix A

Supplementary Materials

Table A.1: The species distribution of 774 genomes, 53 Archaea and 721 Bacteria, categorized by domain, phylum, class and number of thermophiles used for phylogenomic analysis.

			Species/Strains	Thermophiles
Archaea			53	31(58%)
	Crenarchaeota	Thermoprotei	16	15
	Euryarchaeota	Archaeoglobi	1	1
		Halobacteria	5	1
		Methanobacteria	3	1
		Methanococci	7	1
		Methanomicrobia	9	1
		Methanopyri	1	1
		Thermococci	5	5
		Thermoplasmata	3	3
		Unclassified Euryarchaeota	1	0
	Korarchaeota	Unclassified Korarchaeota	1	1
	Nanoarchaeota	Unclassified Nanoarchaeota	1	1
Bacteria			721	44(6%)
	Acidobacteria	Acidobacteria	1	0
		Solibacteres	1	0
	Actinobacteria	Actinobacteria	55	3
	Aquificae	Aquificae	3	3
	Bacteroidetes	Bacteroidia	8	0
		Flavobacteria	4	0
		Sphingobacteria	2	0
		Unclassified Bacteroidetes	1	0
	Candidate division TG1	Unclassified Candidate division	1	0
	Chlamydiae	Chlamydiae	13	0
	Chlorobi	Chlorobia	11	1
	Chloroflexi	Chloroflexi	4	3
		Dehalococcoidetes	3	0
	Cyanobacteria	Gloeobacteria	1	0
		Unclassified Cyanobacteria	32	3
	Deinococcus-Thermus	Deinococci	4	2
	Dictyoglomi	Dictyoglomia	1	1
	Firmicutes	Bacilli	99	5
		Clostridia	37	12
	Fusobacteria	Fusobacteria	1	0
	Nitrospirae	Nitrospira	1	1
	Planctomycetes	Planctomycetacia	1	0
	Proteobacteria	Alphaproteobacteria	93	0
		Betaproteobacteria	61	0
		Deltaproteobacteria	21	0
		Epsilonproteobacteria	22	1
		Gammaproteobacteria	190	1
		Unclassified Proteobacteria	1	0
	Spirochaetes	Spirochaetes	16	0
	Tenericutes	Mollicutes	23	0
	Thermotogae	Thermotogae	7	7
	Verrucomicrobia	Opitutae	1	0
		Unclassified Verrucomicrobia	1	1
		Verrucomicrobiae	1	0

Table A.2: Functional breakdown of all child categories of a) Cellular processes and signaling, b) Information storage and processing, c) Metabolism and d) poorly characterized.

Functional Category	RET	ET	RE	RT	R	E	T	Aq-only	Other	Total
Cell cycle control, cell division, chromosome partitioning (D)	7	3	0	1	0	2	1	1	1	16
Cell wall/membrane/envelope biogenesis (M)	36	15	8	1	3	21	1	7	12	104
Cell motility (N)	6	22	0	0	0	5	2	18	3	56
Posttranslational modification, protein turnover, chaperones (O)	23	8	13	7	11	4	2	7	13	88
Signal transduction mechanisms (T)	13	3	7	0	4	3	1	2	8	41
Intracellular trafficking, secretion, and vesicular transport (U)	4	17	3	1	0	7	1	13	7	53
Defense mechanisms (V)	5	1	2	1	0	0	0	0	1	10
Extracellular structures (W)	0	0	0	0	0	0	0	0	0	0
Nuclear structure (Y)	0	0	0	0	0	0	0	0	0	0
Cytoskeleton (Z)	0	0	0	0	0	0	0	0	0	0

a)

Functional Category	RET	ET	RE	RT	R	E	T	Aq-only	Other	Total
RNA processing and modification (A)	0	0	0	0	1	0	0	0	0	1
Chromatin structure and dynamics (B)	1	0	0	0	0	0	0	0	0	1
Translation, ribosomal structure and biogenesis (J)	57	40	2	3	2	3	7	19	3	136
Transcription (K)	11	11	8	3	3	2	2	5	6	51
Replication, recombination and repair (L)	35	15	6	12	12	5	3	10	9	107

b)

Functional Category	RET	ET	RE	RT	R	E	T	Aq-only	Other	Total
Energy production and conversion (C)	39	0	25	5	12	14	0	10	8	113
Amino acid transport and metabolism (E)	72	1	12	4	7	2	0	2	3	103
Nucleotide transport and metabolism (F)	30	1	8	5	1	0	2	1	0	48
Carbohydrate transport and metabolism (G)	27	2	9	7	6	2	1	2	3	59
Coenzyme transport and metabolism (H)	43	6	15	5	7	2	0	2	4	84
Lipid transport and metabolism (I)	9	13	2	2	4	7	0	1	4	42
Inorganic ion transport and metabolism (P)	28	0	24	2	10	8	0	2	2	76
Secondary metabolites biosynthesis, transport and catabolism (Q)	5	2	8	1	2	4	0	3	2	27

c)

Functional Category	RET	ET	RE	RT	R	E	T	Aq-only	Other	Total
General function prediction only (R)	47	11	39	17	28	11	8	13	25	199
Function unknown (S)	13	7	14	5	27	17	9	12	22	126
Unassigned COG function (Unassigned)	2	2	6	2	22	60	4	683	111	892

d)